Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges

Tongtong Yuan[1], Xuange Zhang[1], Kun Liu[2], Bo Liu[1]*, Chen Chen[3], Jian Jin[4], Zhenzhen Jiao[5]
[1]Beijing University of Technology, CN [2]Beijing University of Posts and Telecommunications, CN
[3]Center for Research in Computer Vision, University of Central Florida, USA
[4]Institute of Industrial Internet of Things, CAICT, CN
[5]Beijing Teleinfo Technology Co., Ltd., CAICT, CN
{yuantt, liubo}@bjut.edu.cn, szxg923@emails.bjut.edu.cn, liu_kun@bupt.cn,
chen.chen@crcv.ucf.edu, jin.jian@caict.ac.cn, jiaozhenzhen@teleinfo.cn

Abstract

*Surveillance videos are important for public security. However, current surveillance video tasks mainly focus on classifying and localizing anomalous events. Existing methods are limited to detecting and classifying the predefined events with unsatisfactory semantic understanding, although they have obtained considerable performance. To address this issue, we propose a new research direction of surveillance video-and-language understanding (VALU), and construct the first multimodal surveillance video dataset. We manually annotate the real-world surveillance dataset UCF-Crime with fine-grained event content and timing. Our newly annotated dataset, UCA (**U**CF-**C**rime **A**nnotation)[1], contains 23,542 sentences, with an average length of 20 words, and its annotated videos are as long as 110.7 hours. Furthermore, we benchmark SOTA models for four multimodal tasks on this newly created dataset, which serve as new baselines for surveillance VALU. Through experiments, we find that mainstream models used in previously public datasets perform poorly on surveillance video, demonstrating new challenges in surveillance VALU. We also conducted experiments on multimodal anomaly detection. These results demonstrate that our multimodal surveillance learning can improve the performance of anomaly detection. All the experiments highlight the necessity of constructing this dataset to advance surveillance AI.*

1. Introduction

Surveillance videos are crucial and indispensable for public security. In recent years, various surveillance-video-oriented tasks have been widely studied, *e.g.*, anomaly detection, anomalous/human action recognition, *etc*. However, the existing surveillance video datasets [20, 25, 26, 38] just provide the category labels and timing of anomalous events, and require all categories to be predefined. Thus, the related methods are still limited to detecting and classifying predefined events merely, lacking the semantic understanding capacity of video content. However, automatic understanding of surveillance video content is crucial to enhance the existing investigative measures. Some surveillance applications often need to search for specific event queries rather than board categories, *i.e.*, using queries to retrieve events in surveillance videos. Meanwhile, intelligent surveillance exhibits a trend toward multimodal directions, especially in video-and-text interaction.

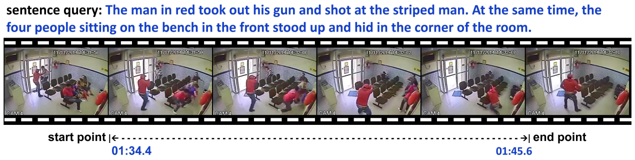

sentence query: The man in red took out his gun and shot at the striped man. At the same time, the four people sitting on the bench in the front stood up and hid in the corner of the room.
start point
01:34.4
end point
01:45.6

Figure 1. Annotation examples in our UCA dataset, including fine-grained sentence queries and the corresponding timing.

A large number of multimodal video datasets [31, 32, 36] have been released, on which various multimodal learning tasks [1, 16, 17] are being explored to semantically understand the video content. However, to our knowledge, surveillance-video-oriented multimodal learning is still understudied. For example, some tasks for investigating detailed cases in public security such as temporal sentence grounding in surveillance video, and surveillance video captioning, have been rarely proposed and studied. One main reason is that the current surveillance datasets lack sentence-level language annotations, which hinders the learning of

*corresponding author

[1] The dataset is provided at https://xuange923.github.io/Surveillance-Video-Understanding.

multimodal tasks. Another implicit reason is the increased difficulty in learning multimodal patterns from surveillance videos due to the unique characteristics they possess, which differentiate them from conventional video datasets. Therefore, this calls for a timely multimodal surveillance dataset to validate the challenges of multimodal surveillance video learning and facilitate the development of surveillance AI.

To address these above issues, we propose extending the existing surveillance video datasets for anomaly detection to multimodal scenarios. Specifically, the multimodal surveillance video dataset should be composed of real-world videos, detailed event and timing descriptions, and as many labeled events as possible. Therefore, we investigated several surveillance datasets [2, 5, 20, 25, 26, 38] and selected UCF-Crime [38] as the foundation of our new dataset. Because UCF-Crime is the largest real-world surveillance dataset and contains a variety of realistic anomalies. To obtain a multimodal surveillance dataset, we contribute new annotations for UCF-Crime [38]. Our dataset is termed UCA (**U**CF-**C**rime **A**nnotation), and it is collected by making manually fine-grained annotations of event content and event timing on UCF-Crime [38]. Some examples are shown in Figure 1.

Compared with UCF-Crime, the main features of our UCA include: (1) The annotation information is relatively fine-grained, and includes as many event descriptions as possible. (2) Considering the realistic demand for temporal localization in the security field, our UCA also includes event timing along with the activity descriptions. (3) UCA can support multiple multimodal understanding tasks, such as Temporal Sentence Grounding in Videos (TSGV) [17], Video Captioning (VC) [1], Dense Video Captioning (DVC) [16], Multimodal Anomaly Detection (MAD) [9]. Our main contributions can be summarized as follows:

- We establish the first real-world and multimodal surveillance dataset with both language descriptions and event timing, which is a comprehensive dataset for the surveillance field to develop the multimodal understandable capacity of machine intelligence.
- We are the first to establish the comprehensive baselines for multimodal tasks (*i.e.*, TSGV, VC, and DVC) on surveillance videos. This research provides a foundational basis for understanding surveillance videos through the integration of video and language.
- We also demonstrate the effectiveness of our UCA in improving the existing surveillance applications. In experiments on multimodal anomaly detection (*i.e.*, MAD), our methods can improve the performance of anomaly detection by providing a basic surveillance video captioning model as a plug-and-play module.
- We thoroughly analyze these experiments and have discovered that mainstream models, commonly used in publicly conventional VALU datasets, exhibit poor performance in multimodal surveillance video scenarios. This suggests that the basic learning model architectures must be modified and designed based on the unique characteristics of surveillance video datasets. This finding also underscores the necessity of constructing our dataset and highlights the challenges associated with multimodal surveillance video learning.

2. Related Work

Table 1. Comparison of the statistics of our UCA and other multimodal video datasets. **Avg word** means the average number of words per sentence. **Temp.anno.** means temporal annotation. The statistics of previous datasets have been recorded in [8].

Dataset	Domain	#Videos	Duration(h)	#Queries	Avg word	Temp. anno.
MSVD [7]	Open	1,970	5.3	70,028	8.7	✓
TACos [31]	Cooking	127	10.1	18,818	9.0	✓
YouCook [10]	Cooking	88	2.3	3,502	12.6	✗
MPII-MD [32]	Movie	94	73.6	68,375	9.6	✓
ActivityNet Captions [16]	Open	20,000	849.0	73,000	13.5	✓
DiDeMo [3]	Open	10,464	144.2	41,206	7.5	✓
VATEX [45]	Open	41,250	114.6	825,000	15.2	✓
MSRVTT [47]	Open	7,180	41.2	200,000	9.3	✓
MAD [37]	Movie	650	1207.3	384,600	12.7	✓
UCA (Ours)	Surveillance	1,854	121.9	23,542	20.15	✓

2.1. Surveillance Video Datasets

The majority of surveillance video datasets have some limitations on the number of videos or the degree of reality, such as UCSD Ped1 [20], UCSD Ped2 datasets [20], Avenue dataset [25], Subway dataset [2], ShanghaiTech Campus dataset [26], NWPU [5], *etc*. Differently, Sultani et al [38] constructed a real-world surveillance video dataset, called UCF-Crime. The dataset consists of 1,900 surveillance videos that present 13 real-world anomalies, such as Abuse, Burglary, Explosion, etc. However, the annotation information of UCF-Crime only includes abnormal categories, which can only be used for abnormal detection. Thus, in the surveillance video research field, more complex multimodal learning tasks, such as moment retrieval and video captioning lack available datasets.

To address this issue, we contribute new annotations on the largest surveillance video dataset UCF-Crime. We manually annotated the event content and event occurrence time for 1,854 videos from UCF-Crime, called **U**CF-**C**rime **A**nnotation (UCA). Additionally, we have noted that the paper SAVCHOI [27] mentions annotations on the UCF-Crime dataset for monitoring suspicious activities. However, SAVCHOI, which only annotated summaries for 300 videos and lacked detailed temporal information on activities, is not directly comparable with our UCA. The annotations from SAVCHOI are not included in our UCA dataset.

2.2. Multimodal Video Datasets

Recently, numerous video datasets have been released for different video-language-understanding tasks, such as video caption, video dense caption, temporal sentence grounding in videos (TSGV), *etc*. We review some video datasets widely utilized in various video-understanding tasks, and list the comparison of the statistics of our UCA and other seven conventional video datasets for multimodal learning tasks in Table 1. All of these video datasets consist of high-quality videos and have been annotated with detailed descriptions, which have played indispensable roles in video and language tasks. The main difference between our dataset and other multimodal datasets lies in the difference in domain. Our dataset is specifically designed for the surveillance field. We have provided detailed descriptions of abnormal behaviors in surveillance videos, which are not typically present in general datasets. Additionally, there are also differences in aspects such as video numbers, duration, queries, *etc*. It is worth noting that our UCA has the longest word length.

2.3. Multimodal Video Learning

In the following, we briefly review some mainstream multimodal video learning tasks. For more completed reviews, we refer readers to [1, 8, 19, 23, 34].

- Temporal sentence grounding in videos (TSGV) is a recent multimodal task [4, 17, 23, 49], which learns the temporal activity localization from a given video with respect to a given language query, *i.e*., the goal of the task is localizing the start and end times for the described activity inside the video. TSGV can be applied in surveillance video to retrieve the event and moment.
- Video captioning (VC) is a multimodal task of producing a natural-language utterance, which aims at describing the visual content of a video. It plays an important technical role in the demand for automatic visual understanding and content summarization [1, 19], which is also important in surveillance video understanding.
- Dense video captioning (DVC) aims to acquire the temporal localization and captioning of all events in an untrimmed video. Surveillance video analysis can also benefit from Dense video captioning. It is more challenging than standard video captioning, which aims to generate a single caption for a video clip [1, 16].
- Multimodal anomaly detection (MAD) is proposed by Chen *et al.* [9], which fused video caption features with original visual and temporal features for anomaly detection in surveillance videos. But the video caption information is derived from Swinbert [22] trained on open-domain dataset VATEX [45]. We aim to introduce surveillance-domain VC models to generate captions to enhance the multimodal anomaly detection task.

With the manually fine-grained annotations, our UCA dataset can be used in various multimodal video learning tasks, including but not limited to the above four tasks. We select SOTA methods of these tasks as baselines in the following experiments, by considering various factors, including the novelty, the differences between methods, the stability of experimental results and the openness of the code, *etc*. These baselines are detailed in **Sec. 4**.

Table 2. Data split and video statistics in our UCA.

Video Statistics	UCF-Crime	UCA	UCA train	UCA val	UCA test
#Video	1,900	1,854	1,165	379	310
Video length	127.5h	121.9h	75.5h	21.2h	25.2h
Annotated video length	—	110.7h	73.7h	16.4h	20.6h

3. The UCA Dataset

Our dataset is based on the UCF-Crime dataset, which is a real-world surveillance video dataset containing 13 real-world anomalies and some normal videos. To extend UCF-Crime to a multimodal dataset, we conducted a fine-grained language annotation on UCF-Crime that recorded each event/change of the videos with detailed descriptions and time stamps. The result of our annotation is a novel dataset named UCA, which is the first large-scale multimodal surveillance video dataset for TSGV, VC, DVC, and MAD.

In this section, we provide a comprehensive outline of the UCA dataset, covering the aspects of data collection and annotation, dataset analysis, comparison with existing datasets, application scope, and ethical considerations. By presenting these details, we aim to provide a concise understanding of the UCA dataset and its potential for advancing research in the field of surveillance VALU.

3.1. Collection and Annotation

During the video collection, we filtered some low-quality videos in the original UCF-Crime to ensure the quality and fairness of our UCA. These low-quality videos encompassed instances of repetition, severe occlusion, and excessively accelerated playback, which affected the clarity of manual annotations and the precision of event timing localization. Therefore, we removed 46 videos from the original UCF-Crime dataset, resulting in 1,854 videos for UCA. The data split in UCA is shown in Table 2.

When labeling a video from UCF-Crime, our goal is to make fine-grained annotations, *i.e*., making a detailed description of each event that can be described in language as much as we can, regardless of whether it is an abnormal event. We also record the starting and ending time for each event in an individual video and the recorded time is 0.1-second interval. The annotated video length of UCA is 110.7 hours, accounting for 86.8% of the total video duration of UCA. Figure 1 presents some annotation examples in UCA.

During the dataset annotation process, we recruited 10 volunteers with computer backgrounds as annotators and formed a review team consisting of 3 AI researchers. To ensure the accuracy and consistency of the annotations, we provided the annotators with comprehensive annotation in-

structions (The instructions are shown in **Sec. 1** of Supplementary Material (SM).). These guidelines were designed to ensure that the annotators utilize linguistically informative language when describing the events depicted in the videos, thereby ensuring clarity and accuracy. Additionally, the instructions emphasized the accurate recording of the start and end times.
Before commencing the annotation work, we provided all annotators with comprehensive training. Throughout the annotation process, annotators were required to watch the videos repeatedly to ensure accurate positioning and description of events. In order to maintain the annotation quality, we implemented a validation process conducted by the review team, wherein annotators' work was reviewed every 100 instances. The focus of this validation was on the quality and consistency of annotations provided by different annotators. Following the completion of all annotations (23,542 sentence-level queries), reviewers conducted a further review of annotated data. The entire annotation and review process required approximately two months. A detailed description of the annotation procedure is shown in SM Sec. 1.

3.2. Dataset Analysis

As shown in Table 2, the UCA dataset comprises a total of 1854 videos, which is divided into three subsets: Train, Validation, and Test sets according to the video lengths and original video categories. These abnormal instances encompass a spectrum of 13 real-world anomalies, including *Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, RoadAccidents, Robbery, Shooting, Shoplifting, Stealing, and Vandalism* [38]. Details for video numbers of different categories in UCA are shown in SM Sec. 1. UCA has little difference from the original UCF-Crime dataset in the modality of video. Therefore, we focus on analyzing statistical information on the language modality.

It can be seen that the total duration of the annotated video is 110.7 hours as shown in Table 2. Figure 2 (a) shows the distribution of the time length of each annotated event in our dataset. The distribution indicates that the video length of the event with the most number is 5-10 seconds, with the majority of events concentrated within 30 seconds and a few events exceeding 40 seconds. Within the UCA dataset, the average duration of each video is 236.8s, and the average duration of each event annotation is 16.9s.

Table 3 shows the number of query descriptions of the events we labeled and the average number of words per query. It can be seen that the average number of words in our annotation is around 20 words. The distribution of the number of words in annotated queries of UCA is shown in Figure 2 (b). The highest number of sentences contains 10-20 words, followed by sentences containing 20-30 words. Additionally, we perform the vocabulary statistics in UCA queries, which can be referred to Table 3. We can find the ratio of nouns, verbs, and adjectives in all sentences of Train, Val, and Test is approximately 2:2:1. This indicates a consistent annotation distribution across different videos.

Table 3. Statistics of annotation in our UCA dataset, including the number of queries (#Query), the number of words per query (#Word/Query), as well as the numbers of nouns (#Nouns), verbs (#Verbs), and adjectives (#Adj).

Annotation Statistics	#Train	#Val	#Test	#Summary
#Query	15,677	3,534	4,331	23,542
#Word/Query	20.45	18.82	20.13	20.15
#Nouns	25,333	5,633	6,748	37,714
#Verbs	26,817	5,569	6,959	39,345
#Adj	12,880	2,963	3,571	19,414

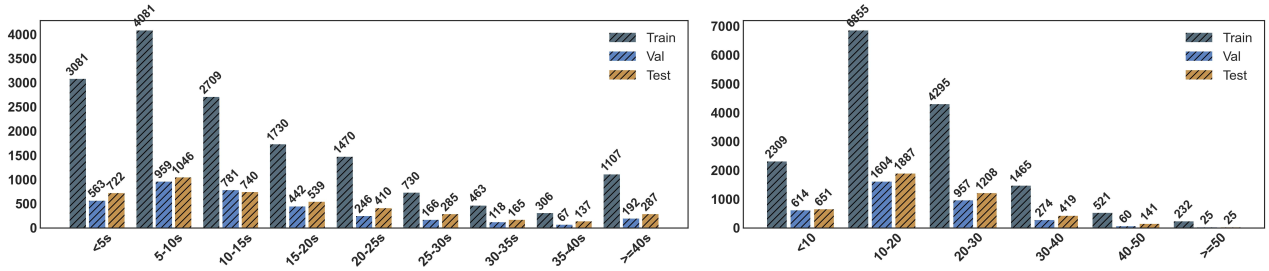
3.3. Comparison with Existing Datasets

Difference with video-language datasets. Table 1 presents a comparison between our UCA dataset and video-language datasets for multimodal learning tasks. Our UCA is for the surveillance domain, which distinguishes it from other datasets. Through this comparison, we observe that our dataset has a moderate number of annotated sentences. However, in terms of the average number of words per sentence, our dataset has the highest number of annotations, indicating that our annotated sentence descriptions are more specific than those of other datasets. **The main differences between UCA and other conventional video datasets lie in the domain of videos (the domain can be seen in Table 1) and video quality (such as uneven image quality, complex backgrounds, and complex events in surveillance videos).** Consequently, when faced with the same learning task, the learning difficulty on our dataset is often much higher than that on a conventional video dataset. The subsequent experiments will demonstrate the experimental results of some state-of-the-art learning methods tested on our dataset, indicating the challenges in multimodal surveillance video learning.

Difference with abnormal video datasets. Our work highlights a key difference from abnormal video datasets [26, 38]: we conducted the unprecedented large-scale sentence-level annotation of events in real-world surveillance scenarios. In contrast to the simple category annotation, we provide specific event descriptions in a fine-grained way. As a result, our dataset can be used to conduct various VALU tasks in surveillance scenarios. To our knowledge, research on multimodal surveillance learning is still blank, and it holds great practical significance.

3.4. Application Scope

UCA is the first multimodal surveillance video dataset and can be applied to the research on intelligent public security. In particular, it can provide a fundamental dataset for tasks related to multimodal surveillance video comprehension, which is a novel area within the surveillance domain.



(a) Event time length distribution.

(b) Distribution of the number of words in sentences.

Figure 2. The duration of annotated events and the number of words in the annotated queries of the UCA dataset.

Based on UCA, researchers can explore retrieving detailed event queries with temporal information, captioning surveillance videos, and multimodal anomaly detection, to improve technical investigative capabilities.

3.5. Ethical Considerations

The videos utilized in our research are sourced from a publicly available dataset called UCF-Crime [38]. This dataset does not explicitly mention any privacy or ethical concerns. Our work exclusively involves sentence-level annotations to describe the events depicted in these videos. Throughout our annotation process, we discovered that the primary videos did not exhibit any particular focus on race or gender. To further minimize any potential risks related to data annotation, we implemented multiple measures, such as annotation guidance and review procedures. Our annotation examples are shown in Figure 1, and more examples are provided in SM Sec. 1. Additionally, we also provide gender-neutral annotations by substituting terms like "man" and "woman" in our annotations with the term "person". Without loss of generality, the experiments presented in this paper were conducted without using gender-neutral annotations.

Users are expected to adhere to our data license agreement, which can be accessed in SM Sec. 4. We emphasize that the usage of the dataset is strictly limited to academic and research purposes. We encourage the research community to provide suggestions for enhancing the quality and ethical reliability of the dataset.

4. Experiments

We proceed with the experiments of four multimodal tasks on the UCA dataset using an RTX3090 GPU. In each task, we first describe the corresponding task and evaluation metric, along with baseline methods, and then present the performance of our selected baselines. We followed the basic experimental settings of these various methods in their respective published papers and public code repositories. Our codebase including data processing and baselines will be open-source to encourage researchers to integrate their new models. For readability, more detailed experiment settings, complementary experiment results, and more visualization examples can be found in SM Sec. 2 and Sec. 3.

Table 4. Benchmarking of TSGV baselines on our UCA dataset.

Method	IoU=0.3		IoU=0.5		IoU=0.7	
	R@1	R@5	R@1	R@5	R@1	R@5
CTRL [12]	6.35	20.09	3.00	10.88	0.74	3.32
SCDM [48]	6.49	10.72	3.65	7.39	1.66	4.78
A2C [14]	4.25	-	1.78	-	0.44	-
2D-TAN [51]	7.76	19.67	4.16	11.54	2.22	6.00
LGI [28]	7.71	-	3.26	-	1.18	-
MMN [46]	8.68	21.38	4.66	11.98	2.12	5.89
MomentDiff [18]	8.15	-	4.46	-	1.71	-

4.1. Temporal Sentence Grounding in Videos

Task. Temporal sentence grounding in videos (TSGV) aims to learn the temporal activity localization from a given video with respect to a given language query [23].

Metric. In existing methods [12, 48, 52], R@K for IoU=θ is commonly adopted as the evaluation metric to measure the performance in TSGV. It is defined as the percentage of at least one of the top-K predicted moments that have IoU with ground-truth moment larger than θ [12]. In the following, we set R@K for IoU= θ with $K = 1, 5$ and $\theta = 0.3, 0.5, 0.7$ as the evaluation metric.

Baselines. We benchmark UCA using 7 different methods (ranging from 2017 to 2023), comparing the novelty, differences, and reproducibility of experiments. CTRL [12] represented a traditional sliding window-based framework of TSGV. SCDM [48] introduced the Semantic Conditioned Dynamic Modulation mechanism, which is a standard anchor-based Method. A2C [14] applied reinforcement learning to this field. For 2D-Map anchor-based methods, 2D-TAN [51] proposed a unique two-dimensional temporal matrix, while MMN [46] further introduced the idea of metric learning. LGI [28] proposed a local-global interaction (LGI) framework as a regression-based method. Momentdiff [18] applied a diffusion model to video moment retrieval.

Implementation Settings. We use C3D [39] pretrained on Sports1M to extract video features for all experiments and the 4,096 dimensions output of the FC6 layer is represented as the video feature. For CTRL, we employ skip-thought [15] as the sentence encoder, resulting in sentence features of 4,800 dimensions. For SCDM, the length of input video clips is set as 512 to accommodate the temporal convolution. For A2C, we also obtain sentence features via skip-thought. For 2D-TAN, the number of video clips is set to 16, and the scaling thresholds t_{min} and t_{max} are set to 0.5 and 1.0. We chose the stacked convolution approach to extract moment-level features. For LGI, the number of sampled segments per video is 128, and the maximum length of each query is 50. The settings of MMN are similar to 2D-TAN, except that max-pooling is used when obtaining moment-level features. The text encoder of Momentdiff utilizes Glove, with a maximum text length set to 32.

Results and Analysis. The experimental results from Table 4 indicate that the TSGV task presents significant challenges within the domain of surveillance videos. In the R@1, IoU = 0.3 metrics, all evaluation results are generally below 10%. This underscores the difficulty in achieving precise event localization in surveillance videos. The videos in UCA have various video durations, but existing models do not adequately consider the handling of long-term temporal information, thereby further affecting the comprehensive fusion of text and video feature information. Among the numerous benchmarking methods, the A2C model, which employs reinforcement learning mechanisms, exhibits relatively inferior performance on the UCA dataset. However, two models employing 2D Temporal Adjacent Networks, namely 2D-TAN and MMN, demonstrate superior overall performance compared to other methods. This advantage can be attributed to their capability to perceive richer contextual information from video data, thereby better comprehending and capturing these temporal characteristics.

4.2. Video Captioning

Task. The goal of video captioning (VC) is understanding a video clip and describing it with language [1].

Metric. We use the metrics as in [22, 33]. The evaluation metrics of correctness include Bilingual Evaluation Understudy (BLEU) [B@n, n=1,2,3,4] [29], Metric for Evaluation of Translation with Explicit Ordering (METEOR) [M] [11], Recall Oriented Understudy of Gisting Evaluation (ROUGE-L) [R] [21], and Consensus-based Image Description Evaluation (CIDEr) [C] [40].

Baselines. We select 6 methods (ranging from 2015 to 2023) for video captioning as our baselines. These methods have various motivations for learning VC: S2VT [41] is a classic work that first proposed an end-to-end model for generating video descriptions, providing important insights for subsequent research. RecNet [42] introduced the encoder-decoder-reconstructor architecture, using the idea of forward and backward flows for video caption generation. MARN [30] improved the traditional encoder-decoder framework and proposed a Memory-Attended Recurrent Network. SGN [33] emphasized the importance of decoded captions and used semantic groups as information units to describe videos. SwinBERT [22] is the first end-to-end video caption generation model fully based on Transformer. CoCap [35] introduced an end-to-end video captioning approach based on compressed videos, achieving faster runtime speeds.

Table 5. Benchmarking of VC baselines on our UCA. B, M, R, and C represent BLEU, METEOR, ROUGE-L, and CIDEr, respectively.

Method	Features	B1	B2	B3	B4	M	R	C
S2VT [41]	Inception V4	28.27	16.34	9.62	5.56	10.35	25.38	15.44
S2VT [41]	VGG 16	23.53	13.17	7.72	4.53	10.39	24.09	13.92
RecNet_global [42]	Inception V4	25.81	15.06	9.14	5.78	10.42	25.12	16.90
RecNet_local [42]	Inception V4	26.89	15.73	9.52	6.00	10.55	25.52	16.70
MARN [30]	Inception V4 ResNext101	26.16	15.61	10.22	6.63	9.67	23.85	15.33
SGN [33]	ResNet101 ResNext101	29.17	16.93	10.32	6.28	11.73	26.25	18.95
SwinBERT [22]	VidSwin	25.02	15.60	9.98	6.33	11.15	27.16	25.29
CoCap [35]	CLIP	28.53	17.34	10.75	6.57	11.43	28.14	21.24

Implementation Settings. The video feature extractor settings are shown in Table 5. For SwinBert, the VidSwin is initialized with Kinetics-600 pre-trained weights [24]. For SGN, we uniformly sample 50 frames and clips from each video, and the visual encoder utilizes pre-trained ResNet and 3D-ResNext [13] models. ResNext in MARN and SGN corresponds to the motion feature, which focuses on the dynamic variation information between video frames.

Results and Analysis. The performance of video captioning baselines is presented in Table 5. Among these benchmarks, SGN has achieved a higher accuracy in phrase matching due to its semantic group design, effectively integrating annotation information with video frames. Two Transformer-based models, SwinBERT and CoCap, have demonstrated superior performance across multiple metrics, particularly in this CIDEr metric. This substantiates their capability to capture key textual information more effectively. Overall, surveillance videos have distinctive characteristics, such as blurry and low-resolution visuals, which will pose more challenges for VC models to learn and align the visual modality and the text modality.

4.3. Dense Video Captioning

Task. Dense video captioning (DVC) generates the temporal localization and captioning of dense events in an untrimmed video.

Table 6. Event localization of DVC on the UCA dataset.

Method	Features	Recall				Precision			
		0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
TDA-CG [43]	C3D	32.54	16.64	8.40	2.49	58.79	27.16	8.95	1.44
PDVC [44]	C3D	48.03	28.48	11.80	2.50	67.33	36.19	12.20	1.94
TDA-CG [43]	I3D	31.72	17.09	8.97	2.89	57.88	29.69	12.45	1.91
PDVC [44]	I3D	49.53	28.59	12.56	2.11	68.84	35.36	12.84	1.73

Table 7. DVC on the UCA dataset with predicted proposals.

Method	Features	Predicted proposals						
		B1	B2	B3	B4	M	C	SODA_c
TDA-CG [43]	C3D	5.50	2.23	0.84	0.38	2.72	4.44	0.98
PDVC [44]	C3D	7.89	3.93	1.65	0.55	3.78	8.58	2.07
TDA-CG [43]	I3D	6.35	2.84	1.22	0.37	3.12	5.51	0.98
PDVC [44]	I3D	8.02	4.22	2.00	0.71	4.06	8.78	2.21

Metric. We have assessed the performance from two perspectives. For localization performance, we employed the evaluation tools provided by the ActivityNet Captions Challenge 2018 and used the common metrics [43, 50], like different IoU thresholds (0.3, 0.5, 0.7, 0.9), classic caption evaluation metrics: BLEU, METEOR, CIDEr, and the performance in describing video stories: SODA_c score.

Baselines. We select three methods (ranging from 2018 to 2022) for DVC with different input features (to increase the diversity of comparisons) as our baselines. These methods have distinct designs and are reproducible. TDA-CG [43] proposed a bidirectional proposal method that effectively utilizes past and future context for proposal prediction. PDVC [44] enhanced the representation of event queries and inputs the localization head and caption head in parallel. UEDVC [50] fully utilized the interdependence between events to detect more diverse and consistent events.

Table 8. DVC on the UCA dataset with Ground-Truth proposals.

Method	Features	Ground-Truth proposals					
		B1	B2	B3	B4	M	C
PDVC [44]	C3D	23.09	11.63	5.38	1.95	10.27	19.16
UEDVC [50]	C3D	28.91	16.54	9.58	5.49	10.59	14.64
PDVC [44]	I3D	23.64	12.58	5.90	2.52	10.56	22.65
UEDVC [50]	I3D	26.00	15.47	9.53	5.71	10.62	18.83

Implementation Settings. For the three models used here, we utilize pre-trained C3D [39] and I3D [6] models for video feature extraction. In the TDA-CG model, the maximum frame length is set to 300, and any frames beyond this length will be truncated. For the PDVC model, the number of event queries is 100. For the UEDVC model, the maximum frame length is set to 200.

Results and Analysis. The localization performance of the model on the UCA dataset is illustrated in Table 6. Overall, the model exhibits a tendency for relatively higher recall and relatively lower precision. Across different IoU thresholds, the PDVC method consistently outperforms the TDA-CG method, implying the PDVC method's notable competence in target localization. Dense captioning performance is shown in Table 7 and Table 8. Using predicted proposals, PDVC surpasses TDA-CG across most metrics, especially in BLEU and METEOR scores, implying superior n-gram matching and semantic alignment in PDVC-generated captions. However, all methods exhibit suboptimal CIDEr performance, suggesting room for enhancing caption diversity. In terms of the SODA_c metric, PDVC's captions better capture the overall video story to some extent compared to TDA-CG. With ground-truth proposals, PDVC outperforms UEDVC in the CIDEr indicator, suggesting a better similarity between its predictions and the ground truth. DVC tasks in UCA encounter challenges from both TSGV and VC, including learning comprehensive temporal information, retaining crucial video event features, and exploring text information.

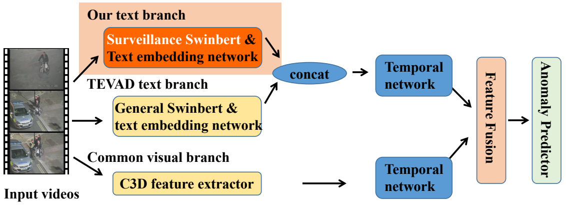


Figure 3. Our enhanced TEVAD framework for multimodal anomaly detection.

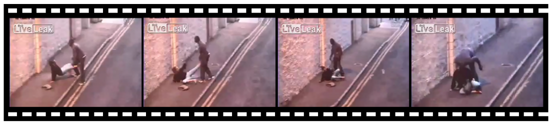
4.4. Multimodal Anomaly Detection

Task. Anomaly detection is one of the important learning tasks in video surveillance. Previous studies have predominantly focused on spatial and temporal features, however, in many complex real-world surveillance videos, these visual features are insufficient in capturing rich semantic meanings. Therefore, Chen *et al.* [9] proposed the learning task of multimodal anomaly detection (MAD), which utilizes captions generated by a video captioning model(*i.e.*, Swinbert [22]) as a text feature source to improve the performance of traditional anomaly detection.

Metric. As in TEVAD [9], we also employ the micro-averaged AUC (Area Under the ROC curve) by concatenating all frames of the videos to obtain the AUC scores.

Baselines and Ours. We choose TEVAD [9] as the baselines of MAD, which fused multi-modal features, including visual features, temporal features, and caption features. We also propose a new framework for anomaly detection with captions on the basis of TEVAD, which is denoted as enhanced TEVAD, and the framework is shown in Figure 3. TEVAD [9] deploys Swinbert [22] pre-trained on VATEX [45] (General Swinbert in Figure 3) as the video caption extractors. VATEX is a large-scale and general video dataset, which provides a general video captioning capacity. Differently, our pre-trained Swinbert on UCA can provide more capacity for captioning the surveillance videos. On the basis of TEVAD, our enhanced TEVAD further introduces our pre-trained Swinbert on UCA (Surveillance Swinbert in Figure 3, as a plug-and-play module) as a domain-specific feature branch for anomaly detection.

Implementation Settings. We use the original experiment settings in TEVAD, and deploy feature concat to fuse multimodal features. The main difference between TEVAD and our enhanced TEVAD is that we introduce a new Surveillance Swinbert trained on UCA.



Groundtruth in UCA: "A powerful and tall man threw a woman to the ground" 🟩 **Anomaly**
Surveillance Swinbert Captioning: "the man in black was knocked down the man in black" 🟩 **Anomaly**
General Swinbert Captioning: "man is standing on the ground and another man is standing on the ground" 🟥 **Normal**

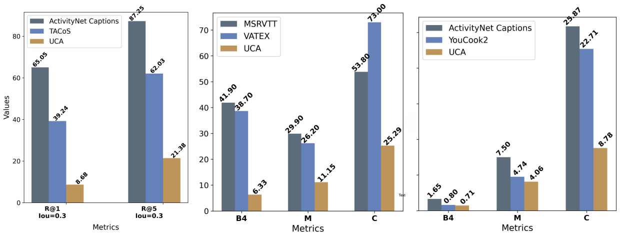
Figure 4. Examples of different video captioning results in MAD.

Results and Analysis. The results in Table 9 show our enhanced TEVAD outperforms others. It indicates that our pretrained Swinbert on UCA can improve anomaly detection accuracy by providing surveillance-domain-specific feature information. In the example of Figure 4, we can find that our Surveillance Swinbert can generate captions with anomalous descriptions, which is useful in multimodal anomaly detection. This also demonstrates the effectiveness of our UCA and our trained multimodal learning models, which can also play an important role in refining existing surveillance tasks.

Table 9. Multimodal Anomaly detection (MAD) on UCF-Crime, where "visual" denotes visual features.

method	visual	General Swinbert	Surveillance Swinbert	AUC
Visual	✓	✗	✗	83.1%
TEVAD [9] (2023)	✓	✓	✗	84.9%
Enhanced TEVAD (Ours)	✓	✓	✓	85.3%

4.5. Discussion



a. Comparison of MMN results in TSGV
b. Comparison of Swinbert results in VC
c. Comparison of PDVC results in DVC

Figure 5. Illustrations of the performance gaps lie in our UCA and other conventional datasets.

We have conducted experiments on mainstream multimodal video understanding tasks. It is the first time that these tasks are oriented to surveillance video. Through these experiments, we can provide a benchmark for surveillance VALU. By analyzing the experiment results, we give the following discussions.

- SOTA methods oriented to conventional videos cannot perform well on our new UCA, which demonstrates the challenges of surveillance VALU. Several instances of these performance gaps lie in multimodal learning between UCA and conventional video datasets are shown in Figure 5. For TSGV tasks, surveillance videos have long video lengths, leading to the difficulties of temporal grounding. For VC and DVC tasks, due to different domains of videos having particularities, a general model cannot well adapt to a surveillance domain. Moreover, we also found that multimodal models trained from general videos could not capture the abnormal behaviors and generate suboptimal descriptions missing critical abnormal information as shown in Figure 4 and SM Sec. 2.5.
- We also recognize that the basic learning model architecture must be modified and designed based on the unique characteristics of surveillance video datasets. We suggest that researchers pay more attention to specific aspects such as video-and-language modality alignment, video-temporal information feature fusion, critical abnormal behaviors grounding, and more comprehensive semantic information mining in the field of surveillance videos. These areas hold significant potential for improving the understanding of surveillance video content.
- For the multimodal anomaly detection task, we can find that surveillance video captioning can play an auxiliary role in anomaly detection by fusing the caption features. The model trained on our UCA can complement the specific capacity of surveillance video understanding, compared with the model trained on conventional videos. This highlights the effectiveness of our UCA. Furthermore, a well-trained multimodal surveillance video learning model can become a promising factor in lifting existing surveillance and security tasks. In the future, we will pay more attention to mining more potentials of our annotations.

However, obtaining such a well-trained surveillance learning model needs the support of datasets. Because the collection of surveillance video data is still challenging, our UCA based on UCF-Crime, also suffers from insufficient data volumes. In the future, we can make more annotations on the newly released surveillance videos, *e.g.*, NWPU [5].

5. Conclusions

In the current AI field, there is a significant gap in research on multimodal surveillance video datasets, despite their potential for contributing to social security and daily life. To bridge this gap, we propose UCA, as the first multimodal surveillance video dataset, which is derived from re-annotating UCF-Crime. Our annotation focuses on providing detailed event descriptions and start-stop time marks for normal or abnormal events occurring in surveillance videos, resulting in the creation of a multimodal surveillance video dataset with 23,542 sentence-level descriptions and frame-level time records. Moreover, we conducted experiments on 4 multimodal tasks using the UCA dataset, evaluating the performance of 17 benchmark methods. Based on the evaluations, we identify significant opportunities and challenges in this field and emphasize the need for further research in both the dataset and model and aspects.

References

[1] Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, et al. A review of deep learning for video captioning. *arXiv preprint arXiv:2304.11431*, 2023. 1, 2, 3, 6

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008. 2

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[4] Wayner Barrios, Mattia Soldan, Alberto Mario Ceballos-Arroyo, Fabian Caba Heilbron, and Bernard Ghanem. Localizing moments in long video via multimodal guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13667–13678, 2023. 3

[5] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20392–20401, 2023. 2, 8

[6] 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. 7

[7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2

[8] Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. Deep learning for video captioning: A review. In *IJCAI*, page 2, 2019. 2, 3

[9] Weiling Chen, Keng Teck Ma, Zi Jian Yew, Minhoe Hur, and David Aik-Aun Khoo. Tevad: Improved video anomaly detection with captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2023. 2, 3, 7, 8

[10] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641, 2013. 2

[11] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 6

[12] 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. 5

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 6

[14] Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8393–8400, 2019. 5

[15] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, 28, 2015. 6

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 2, 3

[17] Xiaohan Lan, Yitian Yuan, Xin Wang, Zhi Wang, and Wenwu Zhu. A survey on temporal sentence grounding in videos. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–33, 2023. 1, 2, 3

[18] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. *arXiv preprint arXiv:2307.02869*, 2023. 5

[19] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312, 2019. 3

[20] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36 (1):18–32, 2013. 1, 2

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[22] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 3, 6, 7

[23] Meng Liu, Liqiang Nie, Yunxiao Wang, Meng Wang, and Yong Rui. A survey on video moment localization. *ACM Computing Surveys*, 55(9):1–37, 2023. 3, 5

[24] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 6

[25] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 1, 2

[26] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 4

[27] Ansh Mittal, Shuvam Ghosal, and Rishibha Bansal. Savchoi: Detecting suspicious activities using dense video captioning with human object interactions. *arXiv preprint arXiv:2207.11838*, 2022. 2

[28] Jonghwan Mun, Minsu Cho, and Bohyung Han. Local-global video-text interactions for temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2020. 5

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[30] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019. 6

[31] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 1, 2

[32] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 1, 2

[33] Hobin Ryu, Sunghun Kang, Haeyong Kang, and Chang D Yoo. Semantic grouping network for video captioning. In *proceedings of the AAAI Conference on Artificial Intelligence*, pages 2514–2522, 2021. 6

[34] Erkan Şengönül, Refik Samet, Qasem Abu Al-Haija, Ali Alqahtani, Badraddin Alturki, and Abdulaziz A Alsulami. An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. *Applied Sciences*, 13(8):4956, 2023. 3

[35] Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, and Libo Zhang. Accurate and fast compressed video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15558–15567, 2023. 6

[36] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference Computer Vision, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 1

[37] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022. 2

[38] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1, 2, 4, 5

[39] 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. 5, 7

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[41] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 6

[42] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 6

[43] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018. 6, 7

[44] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857, 2021. 6, 7

[45] Xin Wang, Jiawei Wu, Junkun Chen, Li Lei, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4580–4590, 2019. 2, 3, 7

[46] Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2613–2623, 2022. 5

[47] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[48] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[49] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *arXiv preprint arXiv:2201.08071*, 2022. 3

[50] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *European Conference Computer Vision, 2022, Proceedings, Part XXXVI*, pages 363–379. Springer, 2022. 7

[51] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020. 5

[52] Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-scale 2d temporal adjacency networks for moment localization with natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9073–9087, 2021. 5