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

Supplementary Material

Outline

The Supplementary Material consists of four parts:

- Sec. 1: This section presents the details of dataset construction, which includes video collection, annotation details, annotation guidance, and analysis of annotation agreement.
- Sec. 2: This section provides experiment details for the submitted manuscript, including a thorough explanation of the experiment settings and visualizations for three multimodal learning tasks.
- Sec. 3: In this section, additional experiments and results are provided, covering the performance on normal and anomalous videos, performance with gender-neutral annotations, and anomaly detection.
- Sec. 4: This section clarifies the license details and accessibility of our UCA dataset.

1. Dataset Construction Details

1.1. Video Collection

All the videos in our UCA dataset are sourced from UCF-Crime [11], a real-world surveillance dataset released at CVPR 2018. The dataset encompasses a wide range of event categories, including Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, Road Accidents, Robbery, Shooting, Shoplifting, Stealing, Vandalism, and Normal Videos.

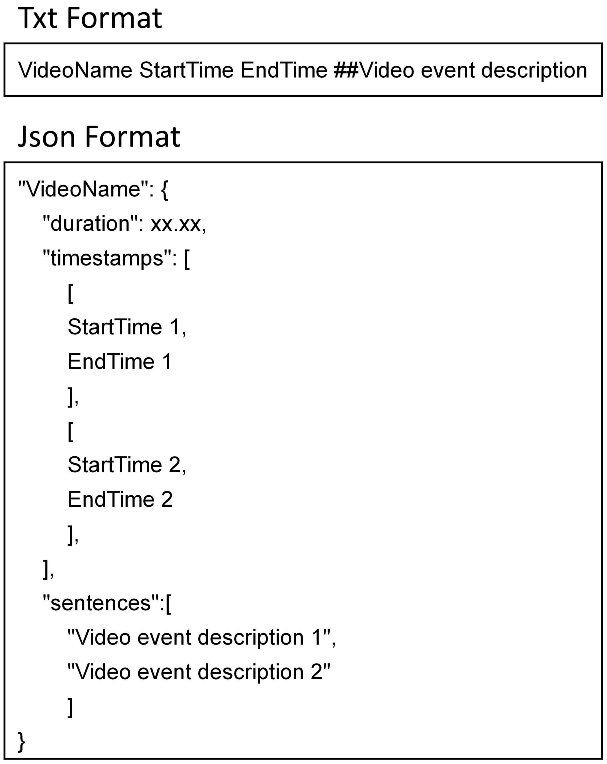
Out of the original 1900 videos, we excluded 46 low-quality videos, which resulted in a final collection of 1854 videos. The criteria for removing these low-quality videos included severe occlusion, blurry content, excessively fast playback speeds, and the presence of duplicate videos. The presence of such low-quality videos poses challenges in subsequent annotation tasks, as they are difficult to identify through manual inspection. The statistics of UCF-Crime and our UCA in terms of the number of videos in different categories are shown in Table 1.

Moreover, following the partitioning approach of the original UCF-Crime dataset, UCA is categorized into two main groups: "Abnormal" and "Normal" videos, as shown in Table 2. In this context, "Abnormal" videos denote those containing scenes with exceptional occurrences or criminal activities present within the original videos.

1.2. Annotations

For videos in UCF-Crime, we provide fine-grained annotations that describe the events occurring within the video. These annotations are highly beneficial for tasks such as video understanding, video temporal localization, and video caption generation. Each annotation also includes the precise start and end times of the events, accurate up to 0.1 seconds.

To meet the needs of different researchers and enhance the convenience of data processing, we have uploaded two versions of the annotation files in our dataset project at https://xuange923.github.io/Surveillance-Video-Understanding, namely txt and json formats. The original annotation data was collected in txt format, which is simple and easy to read, facilitating preliminary data collection and quick reference. Additionally, considering the needs for research and experimentation, we have converted these txt annotations into json format for ease of use in subsequent experiments.



Txt Format

```
VideoName StartTime EndTime ##Video event description
```

Json Format

```
"VideoName": {
  "duration": xx.xx,
  "timestamps": [
    [
    StartTime 1,
    EndTime 1
    ],
    [
    StartTime 2,
    EndTime 2
    ],
  ],
  "sentences":[
    "Video event description 1",
    "Video event description 2"
    ]
}
```

Figure 1. Comparison of Two Versions of Annotation Formats

In the txt files, we record several pieces of information for each annotation, including the corresponding video name, event start time, event end time, and annotation content.

In the json files, we record the timestamp and description of each annotation as a list. Additionally, we include the du-

Table 1. The comparison of UCF-Crime and UCA in the video numbers of different categories.

Video numbers	Abuse	Arrest	Arson	Assault	Burglary	Explosion	Fighting	Road Accidents
UCF-Crime	50	50	50	50	100	50	50	150
Our UCA	50	50	50	48	100	50	50	148
Video numbers	**Robbery**	**Shooting**	**Shoplifting**	**Stealing**	**Vandalism**	**Training_ Normal_Videos**	**Testing_ Normal_Videos**	**Summary**
UCF-Crime	150	50	50	100	50	800	150	1900
Our UCA	149	50	50	100	49	764	146	1854

Txt:

Normal_Videos001_x264 00:00.0 00:11.3 ##A fat man in a blue shirt stood behind the glass, picked up a white item from the bookshelf next to him and looked at it

Normal_Videos001_x264 00:04.2 00:10.5 ##A woman wearing a hat and a pink coat is sitting at the table, typing in front of the lit computer screen

Normal_Videos001_x264 00:09.1 00:15.7 ##A man wearing a white shirt with a black lanyard around his neck sat in front of the desk. The man held up the chair with his hands, and then held the table with his hands to pull his body closer to desk

Json:

```
"Normal_Videos001_x264": {
    "duration": 18.14,
    "timestamps": [
        [
            0.0,
            11.3
        ],
        [
            4.2,
            10.5
        ],
        [
            9.1,
            15.7
        ]
    ],
    "sentences": [
        "A fat man in a blue shirt stood behind the glass, picked up a white item
from the bookshelf next to him and looked at it",
        "A woman wearing a hat and a pink coat is sitting at the table, typing in
front of the lit computer screen",
        "A man wearing a white shirt with a black lanyard around his neck sat in
front of the desk. The man held up the chair with his hands, and then held the
table with his hands to pull his body closer to desk"
    ]
}
```

Figure 2. Detailed Example of the Annotation Format.

Table 2. Abnormal and normal video splits of our UCA dataset.

Statistics	Train		Val		Test	
	Abnormal	Normal	Abnormal	Normal	Abnormal	Normal
#Video	576	589	162	217	206	104
Video length(h)	20.94	54.59	5.88	15.29	6.85	18.38

ration information of the original video, which will facilitate subsequent experimental tasks.

Figure 1 illustrates two versions of the annotation format, while Figure 2 provides a specific example.

1.3. Annotation Guidelines

Here are the main annotation guidelines we established prior to commencing the labeling work, aimed at ensuring accuracy and consistency throughout the data annotation process:

- Fine-grained annotation principle: we emphasize the granularity of the data, considering each event with changes (including changes in human or object states) as an individual data point.
- Time precision: time should be accurately recorded up to 0.1 seconds, allowing for overlaps in the start and end times of multiple samples.
- Rich sentence descriptions: we encourage the use of rich sentence descriptions that enhance semantic understanding, employing techniques such as the use of adjectives to provide detailed coverage of the objects present in the scene.
- Region of Interest (ROI) descriptions: in cases where multiple regions in a single frame experience simultaneous actions or events, annotators may use terms like "top-left corner" or "middle" to differentiate and describe the

changes in the states of people or objects within those specific regions.
- Handling intense changes: For scenes with rapid changes, annotators should combine short-term and long-term descriptions. Initially, they can describe these changes over shorter intervals to capture the micro-variations in the video content. Then, these scattered short-term descriptions should be integrated into a coherent and comprehensive narrative. This layered, multi-granularity annotation method can not only improve the accuracy of the descriptions but also ensure a complete representation of the event, playing a crucial role in video anomaly detection and content analysis.
- Clear action descriptions: If actions in a video are so subtle that they are hard to discern by the human eye, unclear action descriptions can affect the overall quality of the annotation. Hence, we recommend ensuring clarity and accuracy in action descriptions during annotation. For video segments where actions are not obvious or difficult to discern, we decide whether to include them in the dataset based on their importance. In some cases, if these segments' actions are of little value to the study or their ambiguity could lead to misunderstandings, we might choose to omit these annotations or, in extreme cases, remove these videos from the dataset. Under this condition, we encourage annotators to communicate with our reviewer team.
- Complex environment descriptions: In complex environments, where there may be a lack of obvious main events or prominent actions, extracting key information from videos becomes challenging. Therefore, we encourage annotators to provide an overall overview to facilitate a comprehensive understanding of the environment.

In our annotation project, experts initially annotated the video data, collecting 100 examples. These cases not only laid the groundwork for our preliminary version of the annotation guidelines but also became a resource for subsequent annotators to refer to. Throughout the annotation process, we encourage annotators to communicate with our review team whenever they encounter uncertainties or questions. Through this continuous exchange and feedback, we have constantly refined and fine-tuned these annotation principles. These meticulously designed principles provide clear guidance to annotators throughout the process, ensuring high quality and consistency of the annotated data.

1.4. Analysis on Annotation Agreement

Ensuring agreement among different annotators is a critical aspect of the annotation process. 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 achieve this, we impose specific constraints and guidelines to maintain a high level of consistency throughout the annotation process. These constraints include:

- Detailed Annotation Guidelines: we provide annotators with comprehensive guidelines that outline the specific criteria and principles for annotation.
- Training and Familiarization: before starting the annotation work, annotators undergo training sessions where they are familiarized with the annotation guidelines and are given the opportunity to clarify any doubts or concerns.
- Ongoing Communication: throughout the annotation process, annotators are encouraged to engage in regular communication with designated reviewers. This allows them to seek clarification on any ambiguous aspects of the annotation task and receive feedback on their work.
- Review and Validation: the work of annotators undergoes thorough review and validation by designated reviewers. This process involves cross-checking annotations against the guidelines and comparing them with the work of other annotators. During the review process, particular attention has been given to addressing disagreement in annotation styles among annotators. Any discrepancies or inconsistencies are identified and addressed through feedback and clarification.

By implementing these constraints and measures, we aim to ensure a high level of agreement among annotators, which enhances the reliability and quality of the annotated dataset.

2. Experiment Details

To ensure consistency, all experiments are conducted using an RTX3090 GPU with 24GB of memory. The experimental environment is configured with CUDA 11.4, Python 3.9, PyTorch 1.12.0, and TensorFlow 2.12.0. However, due to version compatibility issues in the source code of TDA-CG, a Tesla T4 GPU with 16GB of memory is utilized for running this specific experiment. The environment for this experiment is configured with CUDA 10.1, Python 2.7, and TensorFlow 1.14.0.

In the following, we will present the experimental setup specifics for different experimental tasks. Simultaneously, to compare and demonstrate the efficiency differences among various methods, we also record the runtime of each model.

2.1. Experimental Details of TSGV

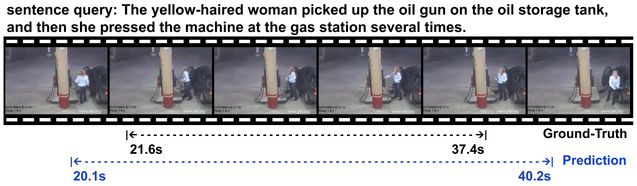

sentence query: The yellow-haired woman picked up the oil gun on the oil storage tank, and then she pressed the machine at the gas station several times.

Figure 3. Visualization results of TSGV

In the Temporal Sentence Grounding in Videos (TSGV) task, we opt to use the C3D network pre-trained on the Sports1M dataset for extracting visual features. As this network is used for visual feature extraction in all experimental models across various datasets, we chose the C3D network for visual feature extraction on the UCA dataset as well, to facilitate experimentation and enable easy comparison with results from other datasets.

For CTRL [3], we set the sliding window size to 128 or 240 frames, equivalent to about 5 or 10 seconds of content in the original video, with an overlap of 0.5. The number of context clips is set to 1. The text encoder employs the Skip-thought model, producing text features of 4800 dimensions. During training, the batch size is set to 32, using the Adam optimizer.

For SCDM [17], the input video clip length is set to 512 frames, longer videos are truncated, and shorter ones are padded with zero vectors. The settings for the temporal convolutional layers are five temporal dimensions {128, 64, 32, 16, 8}. The maximum length for input text is limited to 50, with word embedding sequences obtained through Glove. During training, the batch size is set to 8, using the Adam optimizer.

For A2C [4], when setting normalized start and end points, the maximum frame length T_{max} is set to 10. In the observation network, the output size of the fully connected layer for encoding description features is set to 1024. The state vector s(t) at step t is also 1024 dimensions. The text encoder uses the Skip-thought model. During training, the batch size is 32, using the Adam optimizer.

For LGI [7], the model uniformly samples 128 segments from each video, with the maximum text length set to 50, and λ , which controls the extent of overlap between query attention distributions, is set to 0.2. During training, the batch size is 64, using the Adam optimizer.

For 2D-TAN [19], the number of sampled clips N in visual features is set to 16. Non-maximum suppression with a threshold of 0.5 is applied during inference. The network structure uses an 8-layer convolutional network with a kernel size of 5. During training, the batch size is 32, using the Adam optimizer.

For MMN [16], the number of sampled clips N for visual features is also set to 16, the dimension of the joint feature space d^H is 256, and the temperature parameter is set to 0.1. During training, the batch size is 8, using the Adam optimizer.

For MomentDiff [5], the maximum length of visual features is set to 1000, the maximum text length to 32, and the number of random spans N_r to 5. During training, the batch size is 16, using the Adam optimizer.

In all experiments, the CTRL model exhibits the longest runtime, requiring about 22 hours per training and testing epoch. This is mainly due to the original model processing a large number of clip-sentence pairs. Each training and testing epoch of the A2C model takes about 50 minutes. The epochs for 2D-TAN and SCDM take about 1 hour each. The MomentDiff model requires approximately 20 minutes per epoch. Among all the models, MMN performs the most efficiently, completing each training and testing round within 10 minutes.

2.2. Experimental Details of Video Captioning

In the video captioning (VC) task, we utilize the pre-trained models library built on PyTorch for extracting visual features.

For S2VT [12], we uniformly sample 80 frames from each video, employing pre-trained Inception V4 and VGG16 BN as the visual encoders. During training, the batch size is set to 32, optimized with the Adam optimizer.

For RecNet [13], the visual encoder also utilizes Inception V4, sampling 80 frames uniformly from each video. The maximum sentence length is limited to 50, with excess being truncated. In training, both local and global parts have a batch size of 32, using the AMSGrad optimizer.

For MARN [8], we similarly sample 80 frames from each video, with pre-trained Inception V4 and ResNext101 as visual encoders, subsequently linearly transformed to 512 dimensions. The maximum sentence length is set to 50. Initially, 100 epochs of training are conducted on the Attention-based Recurrent Decoder, followed by the integration of the Attended Memory Decoder. During training, the batch size is 32, using the Adam optimizer.

For SGN [9], 50 frames are uniformly sampled from each video, using pre-trained ResNet 101 and ResNext101 as visual encoders to extract 2D and 3D features. The maximum sentence length is set to 30, with word embedding sequences obtained through GloVe. The model's similarity threshold τ is set to 0.2, and the coefficient λ for Contrastive Attention loss is 0.16. During training, the batch size is 16, using the Adamax optimizer.

For SwinBERT [6], in the preparation stage, each video is segmented into 32 frames, employing an end-to-end training approach. VidSwin is initialized with pre-trained weights from Kinetics-600, and the multimodal transformer encoder is randomly initialized. Due to GPU memory constraints, the batch size during training is set to 4, using the Adam optimizer.

For CoCap [10], before training and testing, the video's minimum edge size is adjusted to 240, and it is compressed using H.264 encoding. The visual encoder is initialized with pre-trained weights from CLIP, while the other encoders and the multimodal decoder are randomly initialized. During training, the batch size is 4, using the Adam optimizer.

In all experiments, the SwinBERT model, due to its end-to-end training mechanism, requires the longest time per epoch for training and testing, approximately 80 minutes.

The CoCap model significantly improves training efficiency by using compressed videos, taking about 40 minutes per epoch. The MARN model, after the integration of the Memory Decoder in the training phase, requires about 60 minutes per epoch. The remaining models complete each training and testing epoch in under 20 minutes.

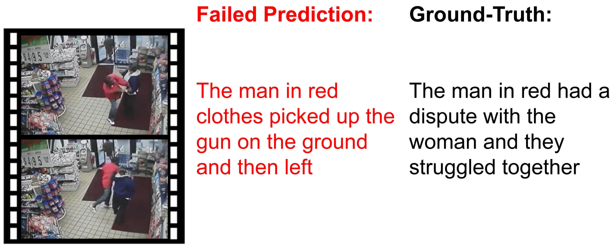
Ground-Truth: The woman in pink clothes took several things from the bed, when she raised her feet, the baby grabbed her skirt and then let go, the woman put the things aside.

Prediction: The woman in the pink top walked to the side of the bed then stood up and walked.

Ground-Truth: A woman wearing white clothes appears on the right and walks into the store on the right.

Prediction: A man in white walks out from the right to the right.

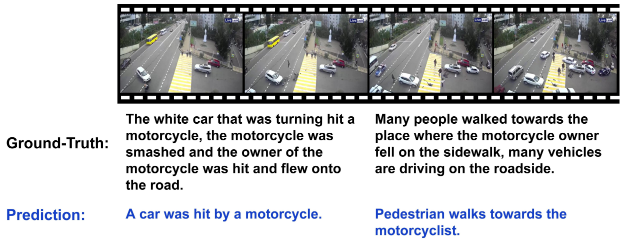
Figure 4. Visualization results of Successful Video Captioning



Failed Prediction:	**Ground-Truth:**
The man in red clothes picked up the gun on the ground and then left	The man in red had a dispute with the woman and they struggled together

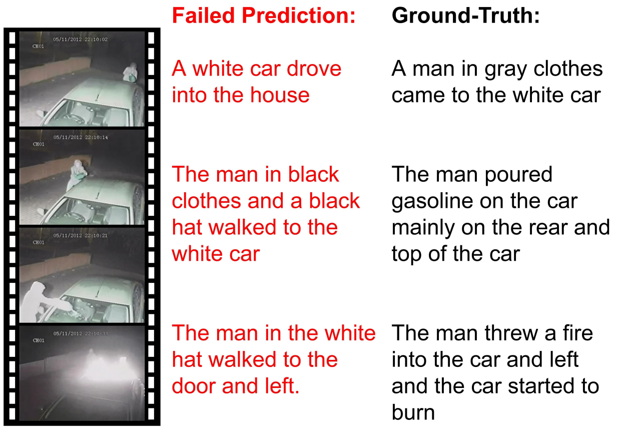
Figure 5. Visualization results of Failed Video Captioning

2.3. Experimental Details of Dense Video Captioning



Ground-Truth:	The white car that was turning hit a motorcycle, the motorcycle was smashed and the owner of the motorcycle was hit and flew onto the road.	Many people walked towards the place where the motorcycle owner fell on the sidewalk, many vehicles are driving on the roadside.
Prediction:	A car was hit by a motorcycle.	Pedestrian walks towards the motorcyclist.

Figure 6. Visualization results of Successful Dense Video Captioning



Failed Prediction:	**Ground-Truth:**
A white car drove into the house	A man in gray clothes came to the white car
The man in black clothes and a black hat walked to the white car	The man poured gasoline on the car mainly on the rear and top of the car
The man in the white hat walked to the door and left.	The man threw a fire into the car and left and the car started to burn

Figure 7. Visualization results of Failed Dense Video Captioning

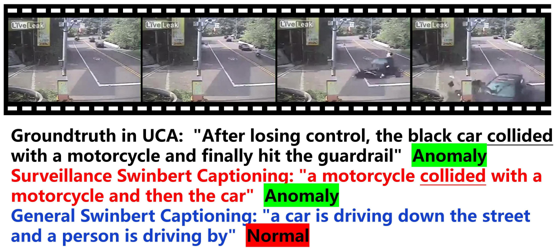
In this dense video captioning (DVC) task, The C3D features used remain consistent with those used in TSGV. The I3D features are directly obtained from an external GitHub repository at https://github.com/tianyu0207/RTFM, which was established by other researchers.

For TDA-CG [14], each feature sequence corresponds to 64 frames of content in the video. The maximum sentence length is limited to 30. During training, the Adam optimizer is used.

For PDVC [15], we set the number of event queries to 100, and conduct experiments using the standard PDVC model with the LSTMDSA captioner. The LSTM hidden layer dimension in the caption head is set to 512. The Adam optimizer is used during training.

For UEDVC [18], the maximum length of video frames is set to 200, and the maximum sentence length is 50. The number of layers in the independent encoder is set to 1, while the cross encoder has 4 layers. This model's training also employs the Adam optimizer.

2.4. Experimental Details of MAD



Groundtruth in UCA: "After losing control, the black car collided with a motorcycle and finally hit the guardrail" Anomaly

Surveillance Swinbert Captioning: "a motorcycle collided with a motorcycle and then the car" Anomaly

General Swinbert Captioning: "a car is driving down the street and a person is driving by" Normal

Figure 8. Examples of different video captioning results in MAD.

We will introduce the experimental details of the Multimodal Anomaly Detection (MAD) task.

For TEVAD [1], we use the same visual features as the original model. The video is divided into non-overlapping segments of 16 frames, and 2048-dimensional visual features are obtained through the I3D feature extractor. The

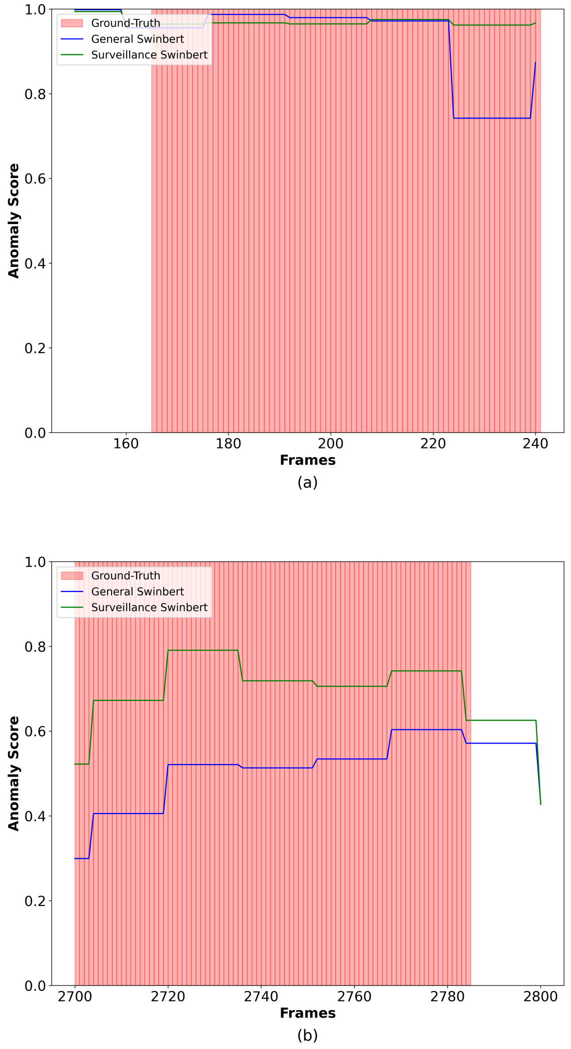


Figure 9. Comparison of MAD results using different text branches. The magnitude of curves represents the anomaly score.

hyperparameter λ , used to adjust the weights of the loss components, is set to 0.0001. During training, the batch size is 32, using the Adam optimizer.

2.5. Visualization Results

This section presents visualizations of the experimental outcomes outlined within the paper. Figure 3 showcases instances where the IoU is greater than 0.7 between predicted and ground truth results in the TSGV task. Figures 4 and 5 respectively depict accurate predictions and erroneous predictions in the Video Captioning task. Furthermore, Figures 6 and 7 display successful and unsuccessful predictions in the Video Dense Captioning task. By visualizing and analyzing these instances of failure, we identify challenges that existing models may encounter when handling surveillance videos, including inaccuracies in color recognition, difficulty in identifying intricate scenes, and struggles in capturing subtle movements.

Figure 8 shows the visualization results in the MAD task. From Figure 8, we can find Surveillance SwinBERT can generate anomaly captions for videos, which is more similar to ground truth in UCA. However, General SwinBERT generates normal captions, which are different from ground truth in UCA. Figure 9 illustrates the impact of different text branches on anomaly detection result scores in MAD. The two subplots extract frames 150-240 and 2,700-2,800 from the video in UCF-Crime, with the red region indicating frames labeled as anomalies in the ground truth. The two curves represent the results of different text branches obtained using General SwinBERT and Surveillance SwinBERT. Clearly, when using Surveillance SwinBERT to generate descriptive statements, anomalous video frames obtain higher anomaly scores, thereby improving the accuracy of anomaly detection. This further demonstrates the effectiveness of introducing the UCA dataset for the anomaly detection task.

These findings contribute to the exploration of the potential value of the new UCA dataset in enhancing anomaly detection model performance.

3. Additional Experiments

3.1. Multimodal Task Performance on Normal and Anomalous videos

The video data in UCF-Crime can be classified into two major categories: abnormal videos and normal videos. To delve into the performance differences between these two types of videos, we conduct a series of tests using pre-trained models, evaluating the best-performing methods for each task.

In the TSGV task, as shown in Table 3, the experimental results indicate that normal videos significantly underperform compared to abnormal videos. This finding aligns with our expectations. The TSGV task focuses on locating specific video segments within untrimmed videos. Given that abnormal videos contain more distinctive segments that starkly contrast with regular scenes, they are easier to identify and locate. We also analyzed the differences in video length between abnormal and normal videos, finding that normal videos contain more long-duration videos, which is a primary reason for the difficulty in event localization in these videos.

In the VC task, the experimental results are presented in Table 4. This task requires generating textual descriptions for video segments, where the original length of the video does not influence the outcome. Abnormal videos often contain specific descriptive terms like ’explosion,’ ’collision,’

Table 3. Performance of TSGV on UCA Dataset

Method	Split	IoU=0.3		IoU=0.5		IoU=0.7	
		R@1	R@5	R@1	R@5	R@1	R@5
MMN [16]	Normal	4.02	10.73	2.14	5.63	1.09	3.11
	Anomalous	16.15	38.48	8.70	22.15	3.78	10.32

Table 4. Performance of VC on UCA Dataset.

Method	Split	B1	B2	B3	B4	M	R	C
CoCap [10]	Normal	27.55	17.27	11.04	6.98	11.79	29.41	21.75
	Anomalous	29.55	16.46	9.2	4.82	10.57	26.10	18.62

Table 5. Performance of DVC on UCA Dataset.

Method	Split	Predicted proposals						
		B1	B2	B3	B4	M	C	SODA_c
PDVC [15]	Normal	8.94	4.97	2.66	1.12	4.21	7.88	1.80
	Anomalous	7.56	3.85	1.67	0.50	3.99	9.23	2.69
Method	Split	Ground-Truth proposals						
		B1	B2	B3	B4	M	C	
PDVC [15]	Normal	22.16	12.73	7.15	3.61	10.32	17.44	
	Anomalous	24.33	12.48	5.28	1.97	10.68	25.26	

etc., which are less frequent in the vocabulary. Precisely describing anomalous events in surveillance videos remains a challenge for existing models, hence normal videos with more generic descriptions perform better in tests.

In the DVC task, as shown in Table 5, the results suggest that normal videos outperform abnormal videos in terms of subtitle generation accuracy. However, since this task requires complete video features as input, the difference in video length between normal and abnormal videos imposes limitations on the models in capturing key video information and narrating the video story.

3.2. Multimodal Task Performance with Gender-Neutral Annotations

To minimize the impact of gender on experimental results, we generated gender-neutral annotations, as detailed in Table 6. For three distinct multimodal tasks, we selected one model for experimentation using gender-neutral annotations.

The experimental results for the TSGV task are presented in Table 7. It can be observed that, at low precision (IoU=0.3), gender-neutral annotations show a slight improvement compared to the regular version. However, as the task difficulty increases, there is no significant difference between the results of gender-neutral and regular versions. This further underscores the challenge of accurately pinpointing "what a person is doing" on the UCA dataset.

The experimental results for the VC and DVC tasks are presented in Tables 8 and Tbale 9. Both tasks require generating captions based on video content. Substituting different gender-specific terms with neutral terms in the vocabulary contributes to a more uniform vocabulary, leading to slightly higher experimental results using gender-neutral annotations compared to the regular version.

Table 6. Replacement vocabulary for gender-neutral annotations.

Regular	Gender-neutral
woman, man, she, he, him	person
herself, himself	themself
her, his	the person's
policeman	police
salesman	salesperson
postman	mail carrier
doorman	doorperson
fireman	firefighter
gunman	person with a gun
repairman	mechanic
cameraman	photographer

Table 7. Performance of TSGV on UCA Dataset with Gender-Neutral Annotations.

Method	IoU=0.3		IoU=0.5		IoU=0.7	
	R@1	R@5	R@1	R@5	R@1	R@5
MMN [16]	9.03	21.77	4.13	12.42	2.08	5.82

Overall, the impact of using gender-neutral annotations on these multimodal tasks is relatively minor. As mentioned in the paper, different tasks still face significant challenges on the surveillance video dataset. We provide gender-neutral annotations in the repository, and researchers considering ethical considerations are encouraged to utilize this version of annotations.

Table 8. Performance of VC on UCA Dataset with Gender-Neutral Annotations.

Method	Features	B1	B2	B3	B4	M	R	C
CoCap [10]	CLIP	30.04	18.71	11.58	6.89	12.22	29.48	20.41

Table 9. Performance of DVC on UCA Dataset with Gender-Neutral Annotations.

Method	Features	Predicted proposals						
		B1	B2	B3	B4	M	C	SODA_c
PDVC [15]	I3D	9.08	5.13	2.71	1.04	4.72	10.06	2.50
Method	Features	Ground-Truth proposals						
		B1	B2	B3	B4	M	C	
PDVC [15]	I3D	25.18	14.05	7.12	2.84	11.58	23.06	

3.3. Anomaly detection

In this section, we conduct additional anomaly detection experiments, aimed at validating the effectiveness of textual information in enhancing anomaly detection results. We choose two models, MGFN [2] and UR-DMU [20], and perform experiments under the environment of Python 3.9, CUDA 11.6, and PyTorch 1.13. For comparison convenience, we adopt the same 2048-dimensional I3D visual features as

Table 10. Comparative Results of Anomaly Detection Accuracy Using Multimodal and Visual-Only Features. Visual* indicates results obtained from models retrained with unified visual features. Multimodal represents using our provided SwinBERT trained on UCA dataset.

Method	Visual*	Multimodal	AUC
MGFN [2]	✓	✗	82.42%
	✓	✓	83.06%
UR-DMU [20]	✓	✗	83.14%
	✓	✓	84.16%

in TEVAD. Moreover, we generate caption texts using the SwinBERT model trained on the UCA dataset, then obtain 768-dimensional sentence embeddings through the supervised SimCSE pretrained on bert-base-uncased. Finally, we concatenate the visual and textual features to serve as the multimodal input features for the models.

For MGFN, we set the random seed to 2023 and retrain the model with both single visual features and multimodal features. During training, the batch size is set to 16 and the learning rate to 0.001. For UR-DMU, the input dimension of visual features is set to 2048, different from the 1024 dimensions used in the original paper. In training, the batch size is set to 32 and the learning rate to 0.0001. Other parameter settings remain the same as in the original paper.

The experimental results are shown in Table 10. It is evident that using multimodal features effectively enhances the accuracy of anomaly detection in surveillance videos. Notably, the results of the experiments with the single visual feature branch are obtained by retraining the model, differing from the data provided in the original paper. This part of the experiments aims to highlight the important role of multimodal information in anomaly detection, and the results convincingly demonstrate the effectiveness and significance of our UCA dataset in improving the accuracy of anomaly detection tasks.

4. License Details and Accessibility

The UCA dataset is released under a Apache License 2.0. In accordance with the Apache License 2.0, users are free to use, modify, and distribute this dataset, but must include the original copyright and license notices. This means that any derivative works or distributed versions based on this dataset should retain the original copyright and license information.

Please note that our dataset is intended solely for academic and research purposes. We encourage the academic community and researchers to use this dataset to advance the field. If you have any questions about the use of the dataset, please contact us directly. We warmly welcome and look forward to your feedback and usage experiences.

The content of the UCA dataset can be accessed at the following link: https://anonymous.4open.science/r/UCA-dataset. We are committed to providing accessible and user-friendly resources to contribute to the advancement of the field of multimodal surveillance video datasets. We sincerely hope this dataset becomes a valuable resource in your research endeavors.

References

[1] 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. 5

[2] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 387–395, 2023. 7, 8

[3] 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. 4

[4] 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. 4

[5] 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. 4

[6] 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. 4

[7] 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. 4

[8] 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. 4

[9] 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. 4

[10] 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. 4, 7

[11] 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

[12] 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. 4

[13] 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. 4

[14] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018. 5

[15] 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. 5, 7

[16] 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. 4, 7

[17] 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. 4

[18] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 363–379. Springer, 2022. 5

[19] 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. 4

[20] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. *arXiv preprint arXiv:2302.05160*, 2023. 7, 8