Human-Centric Behavior Description in Videos: New Benchmark and Model

Lingru Zhou, Yiqi Gao, Manqing Zhang, Peng Wu, Peng Wang, and Yanning Zhang, *Senior Member, IEEE*

***Abstract*—In the domain of video surveillance, describing the behavior of each individual within the video is becoming increasingly essential, especially in complex scenarios with multiple individuals present. This is because describing each individual's behavior provides more detailed situational analysis, enabling accurate assessment and response to potential risks, ensuring the safety and harmony of public places. Currently, video-level captioning datasets cannot provide fine-grained descriptions for each individual's specific behavior. However, mere descriptions at the video-level fail to provide an in-depth interpretation of individual behaviors, making it challenging to accurately determine the specific identity of each individual. To address this challenge, we construct a human-centric video surveillance captioning dataset, which provides detailed descriptions of the dynamic behaviors of 7,820 individuals. Specifically, we have labeled several aspects of each person, such as location, clothing, and interactions with other elements in the scene, and these people are distributed across 1,012 videos. Based on this dataset, we can link individuals to their respective behaviors, allowing for further analysis of each person's behavior in surveillance videos. Besides the dataset, we propose a novel video captioning approach that can describe individual behavior in detail on a person-level basis, achieving state-of-the-art results.**

***Index Terms*—Human-centric caption, Behavior description, Deformable transformer, Video anomaly detection.**

[9.15, 27.95]A man wearing a black jacket and black trousers ran toward the entrance of a supermarket. He was holding a plastic bag in his right hand. He touched the door with his left hand. He transferred the plastic bag to his left hand. He raised his left arm to defend himself from 2. He was pushed against the wall by 2. He stood straight and looked at 2. He looked at the plastic bag and held it with both hands when he was released by 2. He then ran out of the supermarket.
[10.79, 22.94]A man wearing a white long-sleeved shirt and black trousers pushed 1 against the wall at the entrance of a supermarket. He held 1's collar with his right hand. He talked to 1. He looked to his left. He then let go of 1 and walked away.

Fig. 1. This is an example from the UCF-crime captioning dataset we collected. To facilitate in-depth research, we annotated the bounding boxes for the first frame in which each individual appears and recorded the time intervals of their appearance and disappearance in the captions, such as [9.15, 27.25] and [10.79, 22.94]. In these intervals, the first number in each bracket represents the time stamp at which the individual appears in the video, while the second number indicates the time stamp of their disappearance. Additionally, we provided objective descriptions of the behavior of each individual appearing in our dataset.

I. Introduction

With the rapid development of security technology, vision applications based on video surveillances have become the focus of many scholars and the industrial community [1]. So far, research and datasets related to surveillance videos mainly focus on anomaly detection [2], [3], [4], [5], [6], [7], [8], [9], [10]. Undoubtedly, this is an important research field; however, it overlooks some more complex scenarios. For example, in the real world, we need not only to detect abnormal events but also to analyze individual abnormal behaviors in surveillance videos [11], [12] to prevent the occurrence of abnormal events or stop ongoing criminal activities from worsening. Existing research does not fully encompass these scenarios because they demand a human-centric behavioral video captioning dataset to describe individuals, aiding the analysis of individual behaviors, a simple example of which is illustrated in Fig. 1. Currently available datasets, as shown in Fig. 2, mainly describe entire videos or divide videos into several events for description, failing to meet this demand. Obtaining this data requires a large amount of human resources, posing significant challenges in data collection.

To address this issue, we propose a human-centric behavioral video captioning dataset: the UCF-crime captioning dataset (UCCD) as shown in Fig. 1. The UCCD dataset includes 1,012 videos and descriptions of the behavior of 7820 individuals, covering various scenarios, including normal and abnormal events. It not only solves the problem of analyzing individual behaviors in surveillance videos but also contains some unique features. For each individual in a video, we detect them in the frame where they first appear and mark them with different colored

Manuscript received 16 September 2023; revised 15 January 2024, 8 March 2024, and 17 May 2024; accepted 7 June 2024. Date of publication 2 July 2024; date of current version 14 November 2024. This work was supported in part by National Science and Technology Major Project under Grant 2020AAA0106900, in part by the National Natural Science Foundation of China under Grant U19B2037, in part by Shaanxi Provincial Key R&D Program under Grant 2021KWZ-03, and in part by the Natural Science Basic Research Program of Shaanxi under Grant 2021JCW-03. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hanli Wang. *(Corresponding authors: Peng Wu; Peng Wang.)*

Lingru Zhou, Yiqi Gao, Peng Wu, Peng Wang, and Yanning Zhang are with the School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xdwupeng@gmail.com; peng.wang@nwpu.edu.cn).

Manqing Zhang is with the School of Software, Northwestern Polytechnical University, Xi'an 710129, China.

We released the dataset at https://github.com/lingruzhou/UCCD to facilitate future research.

Digital Object Identifier 10.1109/TMM.2024.3414263

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

(a) Video caption

[0.05, 0.55]A man in a room holds a bike and talks to the camera.
[0.55, 1.47]The man adjusts a and talks off the front tire, and folds the bike in half on itself.
[1.47, 2.19]The man unfolds the bike and puts the tire back on.

(b) Dense video caption

[0, 12.52]Outdoor, a lady wearing a black jacket and black jeans was walking on the pavement. She was hit from the back by 2 and fell face first on the pavement. She laid there unconscious.
[0.66, 7.74]Outdoor, a man in a black jacket and black pants was walking after 1 on the pavement. He was holding something in his right hand. He ran to 1 and hit 1's head with his right hand. He then turned around and jogged away.

(c) UCCD

Fig. 2. The figure are respectively examples of the video caption, dense video caption, and UCF-crime caption. The video caption, which was proposed first, is a description of the video. The dense video caption divides a video into several time frames, assigning temporal segmentation to events, and then describing each event. The UCF-crime caption identifies the time period when a person appears and disappears, and describes the behavior of each person within this period.

(a) Abuse023_x264.mp4

(b) Shooting053_x264.mp4

Fig. 3. Subfigure (a) and (b) are videos under different scenarios. In both videos, different individuals are marked in sequence of appearance with vivid yellow, sky blue, emerald green, purple, etc.

bounding boxes, tracking them until they disappear. Fig. 3 represents an example of the color of the boundary boxes appearing in the same order in different videos. This not only enhances the captioning task but also provides a more diversified dataset for research.

Based on the rich annotation information of our dataset, we propose a human-centric video captioning method. This method extracts frame and human features from the deformable transformer and generates a caption for each individual's behavior in the video through the localization head and captioning head. By introducing a module for person detection and tracking, our method can accurately divide the video into different segments according to different people, thereby fully understanding the video content while avoiding information omission and repeated caption generation caused by unreliable estimates of the number of people.

The contributions of this paper are three-fold:

- We construct a human-centric video surveillance captioning dataset across 1,012 videos, detailing the dynamic behaviors of 7,820 individuals. Specifically, we label various aspects such as location, clothing, and interactions with other elements in the scene, greatly enriching the comprehension of human interactions in complex scenarios.
- To the best of our knowledge, we first propose the video surveillance captioning task, enabling the understanding of human actions in videos and the output of descriptions of human behavior, marking a novel direction in the field of video surveillance.
- We propose a novel video captioning approach that is designed on a person-level basis, achieving state-of-the-art results.

The rest of our paper is organized as follows: In Section II, we introduce the relevant work concerning video captioning datasets, video captioning methods, and video anomaly detection. In Section III, we introduce the UCCD dataset. In Section IV, we propose a new method for captioning individual behavior in video surveillance s. In Section V, we validate that our proposed method is superior to existing methods, emphasizing the value of the UCCD dataset in the development of advanced video captioning and surveillance technologies. In Section VI, we conclude the paper.

II. Related Work

Video Captioning Datasets: In recent years, numerous datasets have been compiled in the field of video captioning [13], [14], [15], [22], [23], [24]. These datasets vary in size, scope, and focus, but all provide valuable data for training and evaluating video captioning models. For example, the Microsoft

Research Video Description Dataset (MSR-VTT) [13] offers a large-scale video dataset to researchers. It includes about 10,000 video clips and 200,000 textual descriptions, covering approximately 200 categories, sourced from online video sharing platforms. Similarly, the HowTo100 M dataset [14] provides around 1.36 million YouTube videos with their corresponding spoken descriptions, catering to approximately 1.3 billion video clips, primarily aimed at facilitating how-to tasks such as action recognition, object recognition, and scene recognition. The Microsoft Video Description Corpus (MSVD) [15] offers about 2,000 YouTube short videos with multilingual descriptions. The WebVid dataset [16] contains approximately 500,000 videos scraped from the internet along with their annotations. The bilingual VATEX dataset [17] encompasses around 41,250 videos. The TGIF dataset [18] covers about 100,000 GIF animations with their English descriptions. The TV show Clip Captioning Dataset (TVC) [19] is a video captioning dataset that includes around 15,000 TV show clips with their descriptions. Beyond the realm of vision, the VALOR-1 M dataset developed by Chen et al. [20] comprises 1 million video clips from AudioSet [21], each paired with annotated audiovisual captions.

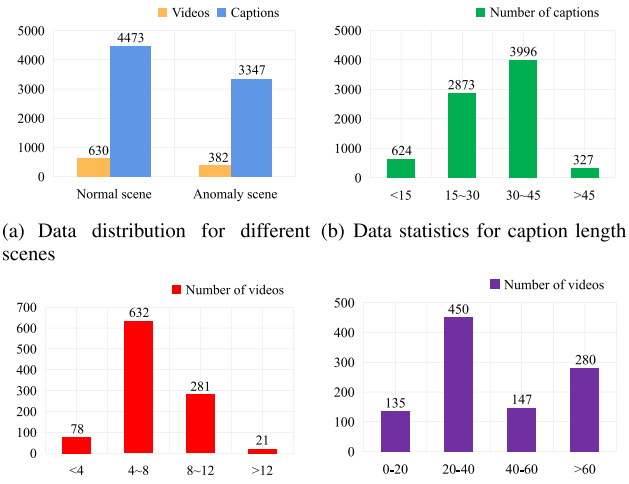
ActivityNet Captions [22] is another dataset, comprising approximately 20,000 YouTube videos along with their linguistic descriptions. This dataset particularly emphasizes the description of the temporal context of videos [23]. In a similar vein, the YouCook2 dataset [24] specializes in cooking videos, containing around 2,000 YouTube videos related to cooking, each accompanied by event-level descriptions. On a different note, the UCF-crime dataset, constructed by Sultani et al. [5], focuses on real-world surveillance footage. It consists of 1,900 videos that capture 13 types of anomalous events, such as abuse, burglary, and explosions. Although this dataset includes annotations specifying the types of anomalies, its primary application has been in anomaly detection. It falls short for more intricate multimodal learning tasks, such as moment retrieval and video captioning. To bridge this gap, Yuan et al. [25] enriched the UCF-Crime dataset by providing more comprehensive annotations for 1,854 of its videos. This effort led to the creation of the UCF-Crime Annotation (UCA), which encompasses detailed annotations regarding the content and timing of the events depicted in the videos.

Existing datasets such as MSR-VTT [13], HowTo100M [14], MSVD [15], WebVid [16], VATEX [17], TGIF [18], and TVC [19] primarily provide video-level descriptions. In contrast, ActivityNet Captions [22], YouCook2 [24] and UCA [25] are dense video caption datasets, segmenting a single video into several events and describing each event at the event-level.

Compared to these existing datasets, our UCCD bears similarity to dense video caption datasets but adopts a unique approach. UCCD focuses on instance-level descriptions of individual behaviors within a video, segmenting the video based on individuals and providing detailed descriptions of the behavior of each person. The rationale for constructing UCCD is that existing video-level and event-level descriptions are insufficient to intricately portray the actions of each individual in a video, thus making it challenging to accurately identify bystanders, victims, or perpetrators in video surveillance. Such granular descriptions are essential for a deeper understanding and analysis of video content.

Video Captioning Methods: In the realm of urban surveillance, significant advancements have been made in action recognition, object tracking, and video captioning. Krishna et al. [22] pioneered the multifaceted task of video captioning with a dense model, integrating a multi-scale proposal [26] module for localization and an attention-driven Long Short-Term Memory (LSTM) for context-aware caption generation [27] [28]. This innovation has sparked further developments, such as the research by Ghaderi et al. [29] introduced a temporal-spatial attention module to improve the accuracy of action recognition. In contrast, Wang et al. [30] presented a comprehensive multi-stage framework that prioritizes precise action identification. The exploration of synergies between video captioning sub-tasks has also provided valuable insights. For example, Li et al. [31] introduced a proxy task to predict language rewards of generated sentences, optimizing the localization module. In a similar vein, Wang et al. [32] presented PDVC, a model that capitalizes on inter-task interactions by sharing intermediate features. Building on these advances, we propose a novel task: generating sequential captions for each individual throughout a video, thereby unifying the fields of action recognition, object tracking, and video captioning into a holistic, individual-centric approach.

Video Anomaly Detection: Video Anomaly Detection (VAD) refers to the identification and detection of events deviating from normal behaviors, widely applied in video surveillance scenarios. Depending on the developmental stages of the algorithms, they can be classified into three categories: traditional machine learning methods, hybrid methods of traditional machine learning and deep learning, and deep learning methods. Most studies employ traditional handcrafted features, such as histogram of oriented gradients (HOG) [33], histogram of optical flow (HOF) [34], local binary pattern (LBP) [35], etc., to represent crowd appearance and motion information, then detect anomalies using conventional machine learning techniques. Given that deep features exhibit stronger descriptive ability than handcrafted features, during the hybrid phase, algorithms use deep features to replace handcrafted ones, followed by anomaly detection using traditional machine learning methods. Discrimination models that saw significant advancements during this stage mainly include point models, and applications focused primarily on cluster discrimination [36], [37], [38], reconstruction discrimination [39], [40], and others [41]. In the phase of deep learning methods, algorithms combine feature extraction steps with model training steps, conducting anomaly detection through end-to-end methods [6], [42], [43], [44], [45], [46], [47]. Currently, video anomaly detection faces multiple challenges, such as the vagueness in anomaly event definition, the lack of clear delineation between normal and abnormal samples [48], and the scene-dependence of anomaly event definition [49]. The same event under different scenes may present different anomalous attributes. In response to these issues, we propose a new task, which includes captioning videos under normal and abnormal scenarios, offering a new foundation and direction for video anomaly detection.



(a) Data distribution for different scenes

(b) Data statistics for caption length

(c) Data statistics for number of people

(d) Duration distribution of videos

Fig. 4. Statistical information on normal and anomaly scene data, caption length, number of people appearing in each video, and distribution of video durations.

III. Dataset

In this section, we provide the UCF-crime captioning dataset, focusing on the data collection process and annotations. The UCF-crime dataset serves as the foundation for our work, where we have augmented it with detailed captions for each individual.

As shown in Fig. 4, we conducted a detailed analysis of the UCCD dataset. In the dataset, the number of videos in normal scenes far exceeds that of anomaly scenes. The length of the captions mainly ranges from 30 to 45 words, with an average of 34 words per caption. The number of people appearing in the videos is usually between 4 to 8, but in some abnormal scenes, such as fights, road accidents, and attacks, the number of people can increase, with some videos even featuring more than 30 people. Considering our task, the videos we selected are mostly around 40 seconds in duration.

A. Data Collection

In our endeavor, we engaged 20 native speakers to manually caption the videos, dedicating a substantial 200 hours to provide targeted training prior to formal annotation. During the labeling process, each video required simultaneous annotation by five individuals. The specific tasks involve initially selecting videos from the UCF-crime dataset that meet our project requirements. The criteria for selection are as follows: the presence of individuals in the video, a maximum duration of 10 minutes per video, and relatively clear video quality to enable accurate observation of the individuals' physical appearance, attire, and specific actions. Such a selection process is critical to ensure that the videos in our dataset possess the necessary detail and clarity for accurate observation and analysis of the behaviors and characteristics of the individuals, thereby enhancing the quality of the dataset and the precision of the annotations. Subsequently, in the selected videos, we detect the individuals present and mark a bounding box around them in the first frame of their appearance. They are then assigned sequential numbers based on the order of their appearance. Following this, we provided an objective description of their actions throughout the process from their appearance to their disappearance in the video. To ensure the accuracy of the caption and the objectivity of the action description, once the five individuals had annotated each video, a seasoned caption annotation expert would sift through their work, selecting the caption most fitting to the setting. After over 5,000 hours of annotation, we ultimately collected 1,012 videos with a total duration of 112 hours, containing a total of 7,820 captions, at a cost of approximately 6,017 dollars.

B. Annotation

1) Person Bounding Box Annotation: In each video, four corners of every person appearing in the first frame are manually marked, and the smallest rectangular bounding box encompassing all four corners is computed and stored. Since multiple individuals typically appear in most videos, to better discern the order of each individual's appearance, we have assigned a color palette consisting of 30 different colors to the bounding boxes. This way, the sequence of appearances of different individuals within the same video is noted according to the color palette, and the sequence of individuals appearing in different videos employs the same color scheme. We have invested 300 man-hours in completing this phase of annotation.

2) Captioning Individual Actions: The most labor-intensive aspect of our annotation process pertains to the captioning of individual actions. Given that numerous individuals appear within a single video, meticulous descriptions of each person's actions necessitate multiple video reviews, which is further complicated by the provision of bounding boxes for only the initial appearance frame of each individual. The captioning of each individual's actions commences from their first appearance-the frame marked with a bounding box and concludes when they vanish from the video. The content of each caption primarily consists of the scene in which the individual is located, their attire, a fine-grained objective description of their actions, and interactions with other individuals within the video. Generally, the length of a caption ranges between 15 to 45 words, but for extended videos, this can increase to approximately 65 words. Among all depicted actions, the most frequently occurring verbs include "walk", "turn", and "look". Given the prevalence of anomalous behavior in the UCF-crime dataset, men are predominantly represented, with the most common scenes being "streets" and "indoors". All annotations are performed by trained native speakers.

C. Comparison With Other Datasets

Table I compares our dataset, based on UCF-crime annotations and centered around human behavior, with other caption datasets. We compare UCCD with existing video caption datasets in several aspects: domain, video source, average time, caption length, caption target, and target type. As can be seen from the table, most videos are sourced from YouTube, while

TABLE I
COMPARISON WITH OTHER DATASETS

Datasets	Domain	Video Source	Average Time	Caption Length	Caption Target	Target Type
MSR-VTT [13]	Open	YouTube	14.8s	9	video	generic event
MSVD [15]	Open	YouTube	9.0s	8	segment	generic event
VATEX [17]	Open	YouTube	10.0s	15	video	action
ActivityNet Captions [22]	Action	YouTube	120.0s	13.5	segment	action
UCCD	Security	Video surveillance	42.3s	34	individual	individual's behaviors

We compare UCCD with existing video caption datasets in several aspects: domain, video source, average time, caption length, caption target, and target type.We discovered that the UCCD dataset is a video caption dataset based on human-centric units, originating from video surveillances.

ours originate from real-world video surveillance. Due to our detailed, granular behavior descriptions of people, we have the longest caption length. Furthermore, the target type of other videos is generic events or actions, while ours is focused on providing detailed descriptions of individuals from multiple perspectives. Our dataset has certain unique features distinguishing it from other datasets, and these are summarized below:

1) Data Source: Unlike datasets like MSR-VTT, MSVD, ActivityNet Captions, etc., which are sourced from YouTube, our dataset is based on UCF-crime. Notably, our dataset includes a plethora of anomaly scene captions in addition to regular scenarios, setting it apart from conventional datasets.

2) Video Integrity: Our dataset provides an extensive video surveillance scene, detailing individual behaviors throughout the video with fine-grained descriptions. This is different from other video caption datasets that typically provide descriptions for short videos or carry out dense video captioning by breaking a video into several segments. Thus, our dataset boasts superior scene and temporal continuity.

3) Behavioral Complexity: In contrast to video captions in normal scenarios like those in HowTo100M, YouCook2, and other existing video caption datasets, our dataset includes descriptions for various anomaly scenarios. These anomaly scenarios are known to entail more complex human behaviors, enhancing the richness and challenge of our dataset.

4) Annotation Challenge: Throughout the captioning process, since only the bounding box of the first frame where a person appears is provided, the annotator must constantly track the individual and describe their interactions with others. This enhances the difficulty of annotation, but simultaneously improves the quality of the dataset.

In summary, our dataset excels in terms of data source, video integrity, behavioral complexity, and annotation challenge. This makes our dataset highly valuable to researchers.

IV. APPROACH

In this section, we introduce a video captioning algorithm that takes advantage of the extensive annotations in our proposed dataset, capable of accurately describing the behavior of each individual appearing in the video. The details of our video captioning algorithm will be elaborated in the following content.

A. Overall Framework

The introduced method, depicted in Fig. 5, consists of two primary components. The initial component deals with feature encoding; this process begins with frame extraction from the video, followed by the utilization of pretrained visual models [50], [51], [52] for frame-level feature extraction. The second stage involves feeding the extracted features, along with their respective positional embeddings [53], into a deformable encoder [54]. Then the YOLOv7 [55] with StrongSORT [56] with OsNet [11] is used to detect and track individuals in the video, cropping the video during the time each individual appears and disappears according to the bounding boxes and extracting frames. The frames that contain only individual's information are used to extract features with the same pretrained visual models [50], [51], [52], and the obtained features of each individual are input into the deformable decoder [54] as a query.

The second component involves decoding, where the features of the individuals in the video are extracted and combined with the frame features in the encoder, before being placed into the decoder. The decoder outputs the queried features [57] connected to a localization head and a captioning head for generating each person's caption. The loss function includes localization loss and captioning loss, used respectively for calculating the timing of the appearance and disappearance of individuals, and for comparing the generated captions with the real captions.

B. Feature Encoding

Our initial step towards exploiting the comprehensive spatio-temporal characteristics within a video is to utilize pretrained visual models to perform feature extraction at the frame level. We employ a consistent frame rate of 30 fps to uniformly sample frames from the video. Each frame, represented as $x_{img} \in R^{3\times H_0\times W_0}$, is processed through I3D to extract features represented as $f \in R^{C\times H\times W}$, with our conventional values being $C = 1024$ and $H, W = H_0/32, W_0/32$.

We handle the video as a series of frames, and for a temporal sequence length of t, we achieve a set of frame features denoted as $X_1^f, X_2^f, ..., X_t^f$, where X_i^f denotes the feature vector for the i^{th} frame. Following this, we employ a 1×1 convolution to reduce the channel dimension from C to a smaller dimension d in the high-level activation map f, generating a new feature map $z_0 \in R^{d\times H\times W}$. Given that the encoder expects a sequence as input, we collapse the spatial dimensions of z_0 into one dimension, forming a feature map of $d \times HW$. Each layer within the encoder is designed with a standard architecture, composed of a multi-head self-attention module and a feed-forward network. Since the transformer architecture is permutation-invariant, it is supplemented with fixed positional encodings which are introduced to the input of each attention layer.

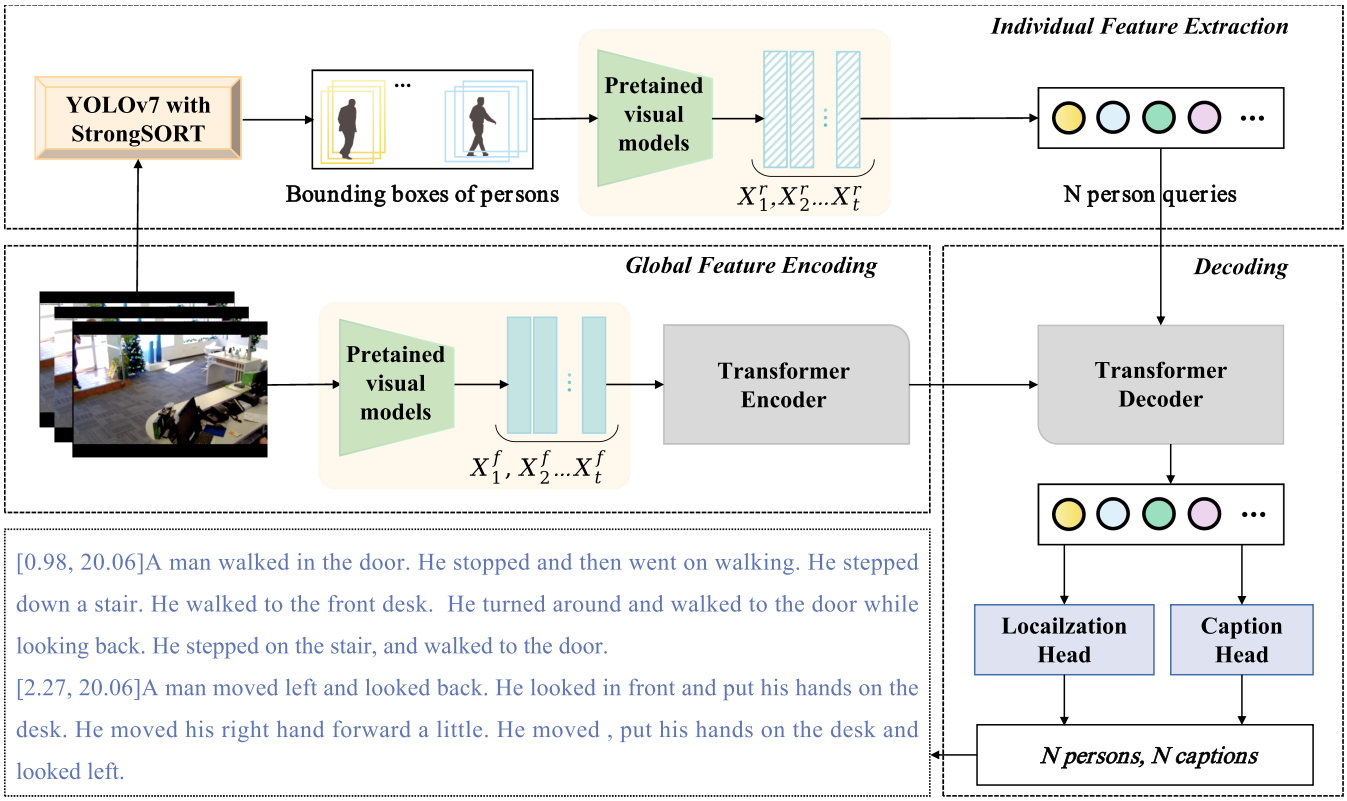


Fig. 5. The overall structure of our proposed model consists of three main components. The first component is feature extraction, in which we separately employ the pretrained visual models to extract global and individual features. The second component is feature processing, where we use a deformable transformer to concentrate attention on the times when individuals appear, thereby enhancing the performance of the model's captioning. The final component consists of a localization head and a captioning head, which generate the captions. The localization component is responsible for identifying the specific timestamps of the appearances and disappearances of individuals. Concurrently, the captioning component, which includes elements like LSTM networks, facilitates the generation of descriptive captions.

C. Person Feature Extraction

Our approach diverges from traditional video captioning methods, which typically focus on depicting specific events. Instead, our strategy centers on comprehensively narrating all actions performed by participants in the video, requiring exhaustive feature extraction at the individual level. During the detection phase, we use YOLOv7 with StrongSORT with OsNet [11] for detecting and tracking individuals. This algorithm incorporates reid [58] capabilities, addressing issues such as losing original identification recognition of the individual after individuals meet or reappear after a period of absence. We obtain frames labeled with individual bounding boxes, denoted as F_i. Subsequently, we segregate regions containing these bounding boxes, yielding the set R_i. In order to more accurately capture the actions of each participant during the interaction process via features, we employ a frame feature extraction technique to identify distinct attributes of each individual. We apply uniform frame sampling, extracting 64 frames from each person, with an input size of 224×224, and then input them into I3D to obtain individual features, ultimately generating the feature set $X_1^r, X_2^r, ..., X_t^r$. Upon completing the pooling process, we obtain an output dimension of $1024 \times t$. These extracted features are then passed through a fully connected(FC) layer to be converted into a compact 256-dimensional format. Each query stands for a single individual's features, and then these N distinct queries are fed into the Transformer decoder's multi-head self-attention layer.

D. Decoding

The decoding network comprises three main components: a Deformable Transformer Decoder, and two parallel heads - a captioning head for generating captions, and a localization head designed to predict human boundaries. The Deformable Transformer is an encoder-decoder architecture based on multi-scale deformable attention (MSDAtt), which mitigates the slow convergence problem of the self-attention in Transformer when processing image feature maps, by attending to a sparse set of sampling points around reference points. Given multi-scale feature maps X, where $X \in R^{C \times H \times W}$, a query element q_j and a normalized reference point $p_j \in [0, 1]^2$, MSDAtt outputs a context vector [59] by the weighted sum of K sampling points across feature maps at L scales.

The goal of the decoder is to query frame-level features of human features under the condition of N human features $q_j{}_{j=1}^{N}$ and their corresponding scalar reference points p_j. It is worth noting that q_j is predicted by linear projection p_j and using a Sigmoid activation function. Human features and reference points serve as initial guesses for human features and positions,

and they interact with each other at each decoding layer. The output query features and reference points are denoted as q_j and p_j, respectively.

To distinguish characters with overlapping characteristics, a localization head is trained, which performs box prediction and binary classification for each unique character feature. Box prediction aims to predict the 2D relative offset of ground truth segments, corresponding to specific reference points. The goal of binary classification is to generate foreground confidence scores for each character query. The mechanisms for box prediction and binary classification are both facilitated by a multilayer perceptron. As such, a set of tuples $\{t_j^s, t_j^e, c_j^{loc}\}_{j=1}^N$ are obtained, representing the start time, end time, and location of the detected characters. Here, c_j^{loc} represents the location confidence of character query $\tilde{q_j}$.

$$MSDAtt(q_j, p_j, X) = \sum_{l=1}^{L}\sum_{k=1}^{K} A_{jlk} W x^l_{\tilde{p}_{jlk}} \tag{1}$$

$$\tilde{p}_{jlk} = \phi_l(p_j) + \Delta p_{jkl} \tag{2}$$

For creating descriptive content for video captions, a different task setup than traditional methods [60] is adopted. A new method is proposed, which uses LSTM hidden state h_{jt} to predict the word w_{jt} after applying a fully connected layer and softmax activation, instead of inputting character-level representation q_j into a standard LSTM at each timestamp. The standard captioning model, considering only character-level representation q_j lacks dynamic interaction between linguistic cues and frame features. To rectify this, a mechanism based on soft attention, known as Deformable Soft Attention (DSA), is introduced. This mechanism effectively enforces soft attention weights to concentrate in a smaller region around the reference point. When generating the t-th word wt, K sampling points are first created on each f^l using linguistic query h_{jt} and character query q_j, following (1), where h_{jt} represents the hidden state within the LSTM. Then, the K sampling points are considered as key values, and $[h_{jt}, q_j]$ are considered as the query inside the soft attention. Considering that sampling points are distributed around reference point p_j, the output feature z_{jt} of DSA is constrained within a relatively small region. LSTM takes concatenated context features z_{jt}, character query features q_j, and previous word $\{w_j, t-1\}$ as input. By applying softmax activation to h_{jt}, the probability of the subsequent word w_{jt} is obtained. As LSTM proceeds, a sentence $S_j = w_{j1}, \ldots, w_{jMj}$ is generated, where M_j indicates the length of the sentence.

E. Loss Function

In the course of training, our model generates a set of actions for N individuals, encompassing both location and description. To align predicted events with actual data in the global scheme, we employ the Hungarian algorithm as per [61] to determine the optimal binary matching outcome. The matching cost is defined as $C = \alpha_{giou}\mathcal{L}_{giou} + \alpha_{cls}\mathcal{L}_{cls}$, where $\mathcal{L}_{giou}$ indicates the generalized IOU [62] between predicted and actual time segments, while $\mathcal{L}_{cls}$ refers to the focal loss [63] between the predicted classification score and actual data labels. In the cost ratio for bipartite matching, we set $\alpha_{giou} : \alpha_{cls} = 2 : 1$, highlighting the greater importance of the generalized IOU loss relative to the classification loss in the calculation of matching cost. The chosen pairs are used to calculate the set prediction loss, which is a weighted sum of gIOU loss, classification loss and caption loss:

$$\mathcal{L} = \beta_{giou}\mathcal{L}_{giou} + \beta_{cls}\mathcal{L}_{cls} + \beta_{cap}\mathcal{L}_{cap} \tag{3}$$

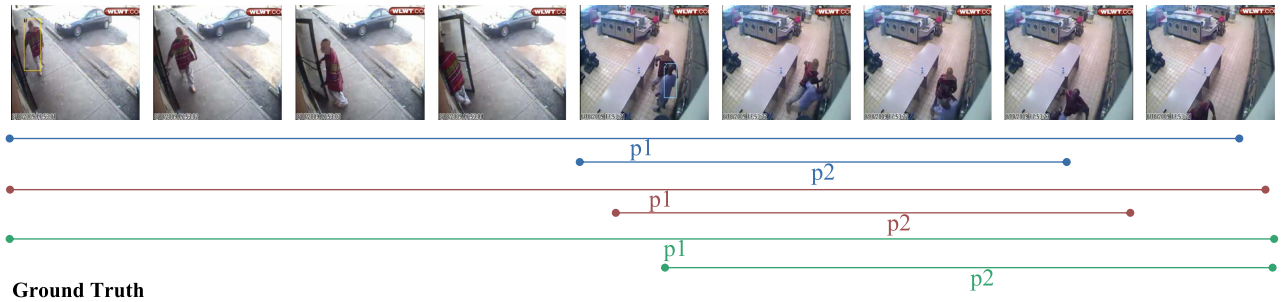
Here, $\mathcal{L}_{cap}$ measures the cross-entropy between predicted word probabilities and actual values, normalized by caption length. The β represents the weights of various losses, such as gIOU loss, classification loss, and caption loss. The loss ratio $\beta_{giou} : \beta_{cls} : \beta_{cap} = 2 : 1 : 1$ clearly demonstrates the relative importance of different types of losses in the overall loss function. Importantly, we adhere to [54], [61] in adding a prediction head at every layer of the transformation decoder. The final loss is the sum of the set prediction losses across all decoder layers.

V. Experiments

In this section, we present the trial outcomes of our proposed video captioning technique on our UCCD. This includes ablation studies, comparative trials against both baseline methods and the latest advancements in video captioning techniques. Additionally, we conduct an evaluation of human performance to gauge the potential of our dataset. For the sake of clarity, we initially provide the evaluation procedures and details of implementation.

A. Evaluation Protocols and Implementation Details

The UCCD is randomly divided into training, validation, and testing subsets. The training subset encompasses 584 videos with 4,014 captions, the validation subset contains 205 videos and 1,842 captions, and the testing subset includes 223 videos with 1,964 captions. In addition to the captions, each video has time locations corresponding to the number of captions, denoting the time the person first appears in the video in each caption. Our method's captioning performance is evaluated using BLEU4 [64], METEOR [65], CIDER [66], and ROUGE-L [67], calculating the average precision of matches between the generated captions and the benchmark true captions at IOU thresholds 0.3, 0.5, 0.7, 0.9. However, these scorers do not consider narrative quality or how well the generated captions cover the entire video story. Hence, we further employ SODA_c [68] for comprehensive evaluation. In addition to this, for evaluating the accuracy of individual behavior descriptions, we also relied on qualitative assessment. Specifically, we measured the accuracy of the captions through human evaluation. We invited a group of 20 evaluators who assessed 20 randomly selected videos and their corresponding captions from our UCCD dataset. During the evaluation process, the evaluators focused on how well the behavior descriptions in the captions matched the actual behaviors observed in the videos. They were asked to rate each caption in terms of its accuracy, relevance, and contextual consistency on a scale of up to 100 points. This was to determine how accurately the captions reflected the individual behaviors in the videos. During the tracking and detection phase, we experiment with the YOLOv5 with DeepSORT and YOLOv7



Ground Truth

p1:[0, 7.92]Outdoor, a man in a red t-shirt with white stripes was walking on the pavement with his left hand inside his pants pocket. He pulled open a gate with his right hand and walked inside. Inside a laundromat, he swung his right hand and punched 2. He swung his left hand and punched 2 again. As 2 stumbled backward, he backed 2 towards the row of washing machines and kept punching 2. He then disappeared.
p2:[4.35, 6.98]Inside the laundromat, a man in a blue T-shirt was standing in front of a row of washing machines. He was being punched in the face by 1. He stumbled backward from the impact of the punch and hit his back on a washing machine. He had his arms stretched out in front of him towards 1 as he was being hit and backed away. He then disappeared.

Ours

p1:[0, 8.24]A man was walking on the road with his left hand down. He opened a door with his right hand and walked inside. He took his right hand and beat. He beat again. He moved forward and kept beating. He then was no longer visible.
p2:[4.68, 7.58]A man was standing in front of a machines. He was being beaten in the face by others. He moved backward and hit on a machine. He had his arms towards others when he was being beaten and backed away. He then was no longer visible.

Vid2Seq

p1:[0, 9.75]A man in a red T-shirt was walking on the road. He opened the door and entered. He beat a man again and again. He then went out.
p2:[5.25, 9.75]A man in a blue T-shirt was standing in front of cabinet. He was being beaten by others. He moved backward and hit on a machine. He beat the man and then couldn't visible.

Fig. 6. The figure shows the visualization results of video captioning. p_1 and p_2 represent two individuals generating two captions. Different colors indicate different models. The horizontal line above represents the timeline, which allows for a more intuitive observation of the model's localization performance.

with StrongSORT with OsNet methods. For UCF-crime caption, we tested our model using three distinct feature extraction methods: C3D [50], I3D [51], and CLIP [52]. We used the C3D pre-trained on Sports1M [69] to extract frame-level features. For I3D, we used a model pre-trained on the Kinetics dataset [51]. For CLIP, we downloaded a pre-trained model released officially by OpenAI. We used a two-layer deformable transformer with multi-scale (4-level) deformable attention. The deformable transformer uses a hidden size of 512 in the MSDAtt layer and 2,048 in the feed-forward layer. After each person is detected and tracked, their features are separately extracted using the corresponding feature extraction method and fed as a query into the deformable transformer. Finally, we employed an LSTM as the caption generator with a hidden dimension of 512. We use the Adam optimizer [70] with an initial learning rate of 5e-5, with each mini-batch sized to one video. Two NVIDIA A100 GPUs facilitate the training process.

B. Comparison With State-of-The-Art Methods

Overall Introduction: Although specialized methods for annotating individual actions across diverse scenarios are limited, we are dedicated to addressing this issue. In our approach, we utilize the I3D model pre-trained on the Kinetics dataset to extract human behavioral features, as Kinetics is a large-scale human action video dataset that is more suitable for our task. We employ YOLOv7 with StrongSORT with OsNet to detect and track each individual's features [71], using I3D to extract them, and interact with the overall conditional features to enhance the quality of individual subtitle generation [60]. As shown in Fig. 6, we have visualized the subtitle results outputted by our model and compared them with existing methods.

Method Comparison: We compared our algorithm with nine leading technologies applied to the most commonly used datasets across different scenarios, such as ActivityNet Captions, MSR-VTT, and VATEX. The results, as shown in Table II, indicate that our algorithm outperforms the other nine state-of-the-art methods, demonstrating our algorithm's outstanding performance. In terms of BLEU-4, CIDER, METEOR, ROUGE-L metrics, we have respectively achieved improvements of 4.2, 3.8, 3.6, and 1.3 over the current sota, highlighting our method's advantages in character positioning and action description.

Reason Analysis: Currently, the cutting-edge technology for dense video subtitling is Vid2Seq [72], and the most advanced methods for the MSR-VTT dataset include mPlug-2 [73], VAST [74], GIT2 [75], VLAB [76], and VALOR [20], with VALOR and VAST leading in the VATEX dataset. Next, we analyze the reasons why these methods do not perform well on the UCCD dataset. Vid2Seq [72] enhances the language model through special time marking but performs poorly in fine-grained description. mPlug-2 [73] resolves the entanglement between video-text modalities but falls short in tracking individuals. VAST [74] achieves state-of-the-art results in video subtitling tasks but struggles to capture actions effectively. Similarly, GIT2 [75] and VALOR [20] demonstrate strong generalization abilities in pre-training but face difficulties in extracting individual actions. We also employed counters from MT [77], BMT [78], PDVC [32] to distinguish different individuals in our task, and the experimental results showed that their effects were not as good as the detection and tracking of people in our methods. Therefore, the performance of these methods on the UCCD dataset is not very satisfactory.

TABLE II
COMPARE OUR MODEL WITH OTHER BASELINES AND VARIOUS INTERMEDIATE MODELS

Method	Features	BLEU-4	CIDER	METEOR	ROUGE-L	Extra Data Training
Vid2Seq [72]	CLIP ViT-L/14	40.5	71.4	26.7	57.9	✓
VALOR [20]	CLIP/VideoSwin Transformer	34.6	62.3	19.4	48.7	✓
mPlug-2 [73]	Dual-vision Encoder	31.7	61.4	16.2	43.6	✗
VAST [74]	Vision Transformer	34.1	64.2	-	-	✓
GIT2 [75]	Florence	29.6	57.8	15.2	41.1	✓
VLAB [76]	CLIP	29.4	56.7	14.5	41.3	✓
MT [77]	TSN	32.8	60.2	16.3	44.5	✗
BMT [78]	I3D	34.3	62.7	18.6	46.8	✗
PDVC [32]	TSP	36.5	65.4	20.7	49.7	✗
Ours	C3D	43.6	73.8	28.5	57.1	✗
Ours	CLIP	44.2	74.6	29.2	57.9	✗
Ours	**I3D**	**44.7**	**75.2**	**30.3**	**59.2**	✗

We can find that I3D has greater contribution for improving performance

TABLE III
EXPERIMENTAL RESULTS FOR VIDEO CAPTIONING IN DIFFERENT SCENARIOS

Method	Normal Scenarious				Anomaly Scenarious			
	BLEU-4	CIDER	METEOR	ROUGE-L	BLEU-4	CIDER	METEOR	ROUGE-L
VALOR [20]	34.6	62.3	19.4	48.7	33.6	59.3	16.2	47.7
VAST [74]	34.1	64.2	-	-	32.1	60.3	-	-
OURS	44.7	75.2	30.3	59.2	44.2	74.8	30.6	58.7

In both normal and anomaly scenarios, our method achieved the best results across all metrics.

C. Anomalies Captioning

Regarding the peculiarities of our dataset, it is worth noting that it is annotated based on the UCF-crime dataset, which is divided into normal videos and videos showcasing anomaly scenarios. These comprise 13 types of real-life anomalies, namely, abuse, arrests, arson, assaults, road accidents, burglary, explosions, fights, robbery, shooting, theft, shoplifting, and vandalism. These particular anomaly behaviors were chosen due to their detrimental impact on public safety. To validate our method's capacity to generate captions for human behavior under anomaly scenarios, we proceeded to re-segment the UCCD. We designated all normal videos to the training set, amassing a total of 630 videos, while all anomaly scenarios were allocated to the validation set, accounting for 382 videos. As illustrated in Table III, from our experimental results, in normal scenarios, the various metrics have been improved by 10.1, 11.0, 10.9, and 10.5, respectively, compared to the state-of-the-art methods benchmarked on the VATEAX dataset. Even under anomaly scenarios, the metrics of our method have been increased by 10.6, 14.5, 14.4, and 11.0, respectively, achieving commendable results.

D. Person Tracking Detection

For the experiment of person tracking detection, we utilize the pre-trained models YOLOv5 with DeepSORT and YOLOv7 with StrongSORT with OsNet on the COCO dataset to perform human detection and tracking. Given that Yolov5 and Yolov7 are designed for multi-type object detection, whereas tracking systems can only track one type of object, we limit the number of detection types to a single class for tracking purposes, i.e., classes 1 for humans.

Upon detecting and tracking the presence of humans, the content within each person's bounding box is cropped and saved as a new video. The original frames are cropped based on each individual's bounding box and then saved directly to the corresponding videowriter. As each person's bounding box size varies, the resolution of each created video differs as well. To maintain consistency in feature extraction, the methodology of extracting individual features remains the same as that for global feature extraction.

Based on the different feature extraction techniques employed subsequently, the videos are resized to the corresponding resolution. For instance, with C3D for feature extraction, the resolution is set at 112×112, with a frame rate of 30fps, and every 16 frames are selected as a person's feature. When using I3D for feature extraction, the resolution is 224×224, with a frame rate of 30fps, and every 64 frames are selected as a person's feature. When using CLIP for feature extraction, the resolution is 224×224, with a frame rate of 30fps, and every 40 frames are chosen as a person's feature.

Apart from these methods, we also tried two other techniques. The first method involves cropping the content in the bounding box of the first frame in which each person appears and then performing feature extraction using the corresponding feature extraction method. The second method involves encoding all the locations where a person has been tracked into the model as a query. Table IV displays the impact of different person tracking detection algorithms on model performance.

E. Ablation Study

This section aims to assess the efficiency of the proposed approach and showcase the contributions of each component of our suggested model towards the final performance. We depict the performance of action captioning by comparing it with four influential ablation studies. For assessment, the generated caption is evaluated using METEOR and SODA_c metrics, as detailed in Table V.

TABLE IV
RESULTS OF THE MODEL USING DIFFERENT PERSON TRACKING DETECTION ALGORITHMS

Method	Features	BLEU-4	CIDER	METEOR	ROUGE-L
BBOX	C3D	42.1	73.4	26.3	56.2
POS	C3D	42.8	73.3	27.6	56.8
YOLOv5	C3D	38.5	69.3	21.6	50.4
YOLOv7	C3D	43.6	73.8	28.5	57.1

Under the condition of using the same feature extraction method, the experimental results of YOLOv7 are the best, indicating that it is more suitable for our task.

TABLE V
ABLATION STUDIES ON THE UCF-CRIME CAPTIONING DATASET VALIDATION SET

Tracking Detection			Transformer		Localization Head		Captioning Head			METEOR	SODA_c
YOLOv7 with StrongSORT	YOLOv5 with DeepSORT	w/o	Vanilla	Deformable	w	w/o	LSTM	SA	DSA		
	✓			✓	✓		✓		✓	27.1	21.5
		✓		✓	✓		✓		✓	26.4	20.3
✓			✓		✓		✓		✓	28.7	22.8
✓				✓		✓	✓		✓	29.8	26.4
✓				✓	✓		✓			29.6	26.1
✓				✓	✓		✓	✓		29.3	23.2
✓				✓	✓		✓		✓	**30.3**	**26.8**

We test the components' impact on our experimental results from four aspects: tracking detection, transformer, localization head, and captioning head.

1) Tracking Detection: Our first experiment was focused on examining the impact of person detection and tracking algorithms on the performance of our model. The results indicate that employing various methods for detecting, tracking, and extracting features for each individual, along with the interaction with global features, significantly affects the overall performance of the model. Given that individuals in surveillance videos often overlap during encounters and reappear frequently, the performance of YOLOv5 with DeepSORT, which incorporates detection and tracking, is not as effective as YOLOv7 with StrongSORT with OSNet. The latter maintains precise localization and re-identification capabilities even in cases of person overlap.

2) Transformer: Transitioning from a deformable transformer to a vanilla one also impacts the model's performance. We observed that incorporating locality into the transformer effectively aids in the extraction of temporally-sensitive features, which is crucial for localization-aware tasks.

3) Localization Head: We further add a localization head in our model, which has shown to improve the determination of the start and end times of events. In other words, it accurately tracks the appearance and disappearance times of each individual.

4) Captioning Head: Regarding caption generation, LSTM is primarily employed, but we enhanced it by adding two distinct attention mechanisms. This addition helps LSTM in generating more accurate descriptions of human behavior. Our results clearly show that focusing on a small segment around the proposals, instead of the entire video, significantly optimizes behavior captioning.

5) Loss Function: In the ablation study shown in Table VI, we conducted detailed tests on various components of the loss function to assess their impact on the final model's performance. By comparing different loss weight ratios, we discovered that these ratios significantly affect the model's performance. Specifically, the model performed best on multiple key performance metrics when we used a ratio of 2:1:1.

TABLE VI
THE IMPACT OF THE WEIGHT OF EACH LOSS RATIO ON THE EXPERIMENTAL RESULTS

Loss ratio of $\beta_{giou} : \beta_{cls} : \beta_{cap}$	BLEU-4	CIDER	METEOR	ROUGE-L
2:1:1	44.7	75.2	30.3	59.2
1:2:1	40.2	68.4	26.5	55.3
1:1:2	42.5	72.1	27.4	56.8
1:1:1	41.4	68.6	26.9	56.1

TABLE VII
COMPARISON BETWEEN HUMAN EVALUATORS AND OUR METHOD

	Annotator 1	Annotator 2	Annotator 3	Ours
Accuracy	92%	90%	91%	76%

Our findings reveal that the average accuracy of human assessments stands at 91%, outperforming our algorithm by a significant 15%. This sizable gap underscores the substantial potential for further enhancing our algorithm's performance.

F. Human Performance Evaluation

In order to explore the complexity of our dataset and the performance difference between human evaluators and our algorithm, we conducted a human performance assessment on our dataset. In this experiment, we randomly selected 10 videos under anomaly scenarios, each video containing an average of about 5 individuals, each appearing for a duration varying from 5 to 30 seconds. Three well-trained annotators participated in this experiment, and their performances are displayed in Table VII. According to the data from the table, the accuracy rate of human annotators is on average 15% higher than that of our model. Despite human annotators providing precise descriptions, the time cost is significantly high. Annotating each video requires an average of 13 minutes per video, while our trained model can generate captions within a few seconds.

VI. CONCLUSION

In this paper, we have gathered the UCF-crime captioning dataset, the first of its kind, to our knowledge, that offers captions for anomalous videos in real-life surveillance scenarios. We furnish the bounding box for the first frame each individual appears

in and proceed to provide an objective description of the person's entire behavior in the video. Consequently, our dataset can also be applied to other vision tasks, such as action recognition and anomalyity detection in videos. Moreover, during the captioning of individuals, we have also included some information about their interactions with other people; how to capture this information is worthy of further exploration. In addition, we have conducted thorough experiments to fully exploit the rich annotations. Drawing on the abundant annotation data in our dataset, we have proposed a novel method for captioning people's behaviors in videos. This method can detect and track individuals at every time point in the video and provide an objective description of their behaviors. Experimental results show that our proposed method significantly outperforms five state-of-the-art video captioning methods evaluated on our dataset.

REFERENCES

[1] G. Shidik et al., "A systematic review of intelligence video surveillance: Trends, techniques, frameworks, and datasets," *IEEE Access*, vol. 7, pp. 170457–170473, 2019.

[2] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2138–2148, Aug. 2020.

[3] R. Morais et al., "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. On Comput. Vis. Pattern Recognit.*, 2019, pp. 11996–12004.

[4] N. Li, F. Chang, and C. Liu, "Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Trans. Multimedia*, vol. 23 pp. 203–215, 2021.

[5] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6479–6488.

[6] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2609–2622, Jul. 2019.

[7] P. Wu et al., "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Proc. Comput. Vis.–ECCV 2020, 16th Eur. Conf.*, 2020, pp. 322–339.

[8] P. Wu and J. Liu, "Learning causal temporal relation and feature discrimination for anomaly detection," *IEEE Trans. Image Process.*, vol. 30 pp. 3513–3527, 2021.

[9] Z. Yang, J. Liu, Z. Wu, P. Wu, and X. Liu, "Video event restoration based on keyframes for video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14592–14601.

[10] P. Wu et al., "Towards video anomaly retrieval from video anomaly detection: New benchmarks and mode," *IEEE Trans. Image Process.*, vol. 33, pp. 2213–2225, 2024.

[11] Q. Hongxu and S. Ablameyko, "Multi-object tracking by using strong SORT tracker and YOLOv7 network," Информационные системы и технологии, pp. 76–79, 2022.

[12] Y. Bi, H. Jiang, Y. Hu, Y. Sun, and B. Yin, "See and learn more: Dense caption-aware representation for visual question answering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 1135–1146, Feb. 2024.

[13] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5288–5296.

[14] A. Miech et al., "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE/CVF Int. Conf. On Comput. Vis.*, 2019, pp. 2630–2640.

[15] D. Chen and W. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 190–200.

[16] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1728–1738.

[17] X. Wang et al., "VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4581–4591.

[18] Y. Li et al., "TGIF: A new dataset and benchmark on animated GIF description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4641–4650.

[19] J. Lei, L. Yu, T. Berg, and M. Bansal, "TVR: A large-scale dataset for video-subtitle moment retrieval," in *Proc. Comput. Vis.—ECCV 2020, 16th Eur. Conf.*, 2020, pp. 447–463.

[20] S. Chen et al., "VALOR: Vision-audio-language omni-perception pretraining model and dataset," 2023, *arXiv:2304.08345*.

[21] J. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 776–780.

[22] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 706–715.

[23] Y. Wang et al., "GEB : A. benchmark for generic event boundary captioning, grounding and retrieval," in *Proc. Comput. Vis.–CCV 2022, 17th Eur. Conf.*, 2022, pp. 709–725.

[24] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[25] T. Yuan et al. "Towards Surveillance Video-and-Language Understanding: New Dataset, Baselines, and Challenges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 2024, pp. 22052–22061.

[26] Y. Liu, H. Li, J. Cheng, and X. Chen, "MSCAF-Net: A general framework for camouflaged object detection via learning multi-scale context-aware features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4934–4947, Sep. 2023.

[27] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7190–7198.

[28] D. Yang and C. Yuan, "Hierarchical context encoding for events captioning in videos," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 1288–1292.

[29] Z. Ghaderi, L. Salewski, and H. Lensch, "Diverse video captioning by adaptive spatio-temporal attention," in *Proc. Pattern Recognit., 44th DAGM German Conf.*, 2022, pp. 409–425.

[30] L. Wang et al., "Multi-stage tag guidance network in video caption," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4610–4614.

[31] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly localizing and describing events for dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7492–7500.

[32] T. Wang et al., "End-to-end dense video captioning with parallel decoding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6847–6857.

[33] M. Lewandowski, D. Simonnet, D. Makris, S. Velastin, and J. Orwell, "Tracklet reidentification in crowded scenes using bag of spatio-temporal histograms of oriented gradients," in *Proc. Pattern Recognit., 5th Mex. Conf.*, 2013, pp. 94–103.

[34] Y. Cong, J. Yuan, and J. Liu, "Abnormal event detection in crowded scenes using sparse representation," *Pattern Recognit.*, vol. 46, pp. 1851–1864, 2013.

[35] J. Xu, S. Denman, S. Sridharan, C. Fookes, and R. Rana, "Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes," in *Proc. Joint ACM Workshop Model. Representing Events*, 2011, pp. 25–30.

[36] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 56–62.

[37] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1992–2004, Apr. 2017.

[38] M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, and R. Klette, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Comput. Vis. Image Understanding*, vol. 172, pp. 88–97, 2018.

[39] J. Sun, X. Wang, N. Xiong, and J. Shao, "Learning sparse representation with variational auto-encoder for anomaly detection," *IEEE Access*, vol. 6, pp. 33353–33361, 2018.

[40] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 246–255, Jan. 2019.

[41] R. T. Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2895–2903.

[42] J. Feng, C. Zhang, and P. Hao, "Online learning with self-organizing maps for anomaly detection in crowd scenes," in *Proc. IEEE 20th Int. Conf. Pattern Recognit.*, 2010, pp. 3599–3602.
[43] Y. Fan et al., "Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder," *Comput. Vis. Image Understanding*, vol. 195, 2020, Art. no. 102920.
[44] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 2018, *arXiv:1802.06360*.
[45] M. Hasan, J. Choi, J. Neumann, A. Roy-Chowdhury, and L. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 733–742.
[46] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder," *Electron. Lett.*, vol. 52, pp. 1122–1124, 2016.
[47] T. Wang et al., "Generative neural networks for anomaly detection in crowded scenes," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 5, pp. 1390–1399, May 2019.
[48] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, pp. 1–58, 2009.
[49] Y. S. Chong and Y. H. Tay, "Modeling representation of videos for anomaly detection using deep learning: A review," 2015, *arXiv:1505.00523*.
[50] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
[51] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.
[52] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[53] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
[54] X. Zhu et al., "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[55] C. Wang, A. Bochkovskiy, and H. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.
[56] Y. Du et al., "StrongSORT: Make DeepSORT great again," *IEEE Trans. Multimedia*, vol. 25, pp. 8725–8737, 2023.
[57] X. Sun, J. Gao, Y. Zhu, X. Wang, and X. Zhou, "Video moment retrieval via comprehensive relation-aware network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5281–5295, Sep. 2023.
[58] G. Zhang, H. Zhang, W. Lin, A. Chandran, and X. Jing, "Camera contrast learning for unsupervised person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4096–4107, Aug. 2023.
[59] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4257–4268, Aug. 2023.
[60] X. Yang, F. Lv, F. Liu, and G. Lin, "Self-Training Vision Language BERTs with a Unified Conditional Model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3560–3569, Aug. 2023.
[61] N. Carion et al., "End-to-end object detection with transformers," in *Proc. Comput. Vis.–ECCV 2020, 16th Eur. Conf.*, 2020, pp. 213–229.
[62] H. Rezatofighi et al., "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[63] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[64] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[65] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Meas. Mach. Transl. Summarization*, 2005, pp. 65–72.
[66] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4566–4575, 2015.
[67] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[68] S. Fujita, T. Hirao, H. Kamigaito, M. Okumura, and M. Nagata, "SODA: Story oriented dense video captioning evaluation framework," in *Proc. Comput. Vis.–ECCV 2020, 16th Eur. Conf.*, 2020, pp. 517–531.
[69] A. Karpathy et al., "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1725–1732.
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[71] H. Li, M. Liu, Z. Hu, F. Nie, and Z. Yu, "Intermediary-guided bidirectional spatial-temporal aggregation network for video-based visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4962–4972, Sep. 2023.
[72] A. Yang et al., "Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10714–10726.
[73] H. Xu et al., "mPLUG-2: A modularized multi-modal foundation model across text, image and video," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2023, pp. 38728–38748.
[74] S. Chen et al., "VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36.
[75] J. Wang et al., "GIT: A generative image-to-text transformer for vision and language," 2022, *arXiv:2205.14100*.
[76] X. He et al., "VLAB: Enhancing video language pre-training by feature adapting and blending," 2023, *arXiv:2305.13167*.
[77] L. Zhou, Y. Zhou, J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8739–8748.
[78] V. Iashin and E. A. Rahtu, "Better use of audio-visual cues: Dense video captioning with bi-modal transformer," 2020, *arXiv:2005.08271*.