applied sciences

MDPI

Article

Cross-Modality Interaction-Based Traffic Accident Classification

Changhyeon Oh and Yuseok Ban *

School of Electronics Engineering, Chungbuk National University, 1 Chungdae-ro, Seowon-gu, Cheongju 28644, Republic of Korea; toddud08@chungbuk.ac.kr
* Correspondence: ban@cbnu.ac.kr; Tel.: +82-43-261-2475

Abstract: Traffic accidents on the road lead to serious personal and material damage. Furthermore, preventing secondary accidents caused by traffic accidents is crucial. As various technologies for detecting traffic accidents in videos using deep learning are being researched, this paper proposes a method to classify accident videos based on a video highlight detection network. To utilize video highlight detection for traffic accident classification, we generate information using the existing traffic accident videos. Moreover, we introduce the Car Crash Highlights Dataset (CCHD). This dataset contains a variety of weather conditions, such as snow, rain, and clear skies, as well as multiple types of traffic accidents. We compare and analyze the performance of various video highlight detection networks in traffic accident detection, thereby presenting an efficient video feature extraction method according to the accident and the optimal video highlight detection network. For the first time, we have applied video highlight detection networks to the task of traffic accident classification. In the task, the most superior video highlight detection network achieves a classification performance of up to 79.26% when using video, audio, and text as inputs, compared to using video and text alone. Moreover, we elaborated the analysis of our approach in the aspects of cross-modality interaction, self-attention and cross-attention, feature extraction, and negative loss.

Keywords: traffic accident classification; video highlight detection; cross-modality interaction; augmented reality; moment retrieval

Citation: Oh, C.; Ban, Y. Cross-Modality Interaction-Based Traffic Accident Classification. *Appl. Sci.* **2024**, *14*, 1958. https://doi.org/10.3390/app14051958

Academic Editor: Luís Picado Santos

Received: 9 February 2024
Revised: 22 February 2024
Accepted: 24 February 2024
Published: 27 February 2024

1. Introduction

As of 2020, in terms of traffic accident fatalities per 100,000 population among OECD countries, the United States had the highest rate at 11.6, followed by Mexico at 10.6, Chile at 9.3, South Korea at 6, and Switzerland at 2.6 [1]. In this context, the probability of death occurring on the road is higher in secondary traffic accidents than in primary ones, making it crucial to quickly identify accidents on the road to prevent subsequent secondary accidents. Consequently, in the field of artificial intelligence, technologies are being actively developed to quickly detect traffic accidents or accurately classify types of accidents [2–6].

The technology for classifying accidents, which determines the occurrence of an accident in a given image or a specific segment of a video, involves a process of anomaly detection that identifies abnormal events from normal ones [7–11]. The predictive model used for this purpose is crucial for its ability to distinguish abnormal parts within the video sequences that make up the video. The definition of what constitutes an abnormal event can vary and may include a car invading on a sidewalk or vehicles colliding, depending on the training environment.

On the other hand, with the growth of platforms for video content such as YouTube and the rapid expansion of the related market, there has been an explosive increase in video data, and while the amount and duration of video content on these platforms are increasing, there is a growing demand for content summarized in shorter forms, emphasizing the need for technology to identify meaningful segments. Within video content, the sections of interest can vary among users. To address this, new technologies have been developed that allow users to input a query, enabling the predictive model to selectively find related

segments in the video. A query refers to a sentence that defines the meaning of a video segment from the user's perspective. For example, "Cars collide" could be such a query. The video segment related to the user-defined query is then used as a video highlight, serving as training data for the predictive model. Notable technologies for video highlight detection include Moment-DETR [12], UMT [13], and QD-DETR [14].

In this paper, we propose to apply the video highlight detection networks on the traffic accident dataset in order to identify accident segments within video sequences and to investigate the main aspects that should be considered during the application process. To achieve this, we introduce the concept of cross-modality, leveraging the interaction between video, text, and audio data. The video highlight detection networks are utilized to predict video segments relevant to user-defined text queries. These use the video and audio features extracted from the video and text features of the query as inputs to generate outputs that consist of highlighted time intervals, defined by start and end times, along with saliency scores.

Existing video highlight detection models, such as Moment-DETR, UMT, and QD-DETR [12–14], have never been applied to the classification of accident videos. These models are typically used with datasets composed of activities like cooking and travel (e.g., QVHighlights [12]), aiming to categorize specific daily life events. Our work introduces the first application of video highlight detection models for the task of traffic accident video classification.

We generate necessary supplementary information to facilitate the use of existing traffic accident datasets for video highlight detection. This additional information comprises annotations of the start and end segments of traffic accidents, queries of accident types, and saliency scores. These datasets include a range of weather conditions such as clear, snowy, and rainy, as well as various types of traffic accidents. Therefore, they can prevent traffic accident video classification models from becoming biased towards detecting accidents in specific situations.

Furthermore, we create accident classification models using three distinct primary video feature extraction methods [15] and analyze the accident detection performance of each model.

In summary, the contributions of this paper are as follows:

- For the first time, we introduce a video highlight detection network to the task of traffic accident video classification.
- We generate necessary additional information to make the existing traffic accident dataset more conducive to video highlight detection.
- We analyze the performance of traffic accident video classification from the perspectives of cross-modality interaction, self-attention and cross-attention, feature extraction, and negative loss.

2. Background

Several methods have been proposed for effectively extracting visual features from video data. The Slowfast network [16] is a neural architecture designed for video classification, capable of detecting objects in images or videos and identifying specific actions or scenes within video data. This approach was first introduced by Facebook AI Research in 2019 and has since achieved first place in the CVPR2019 AVA Challenge for action recognition due to its exceptional performance. The 2D ResNet network [17] is a deep neural network structure used for image classification, object detection, and region segmentation within videos, proposed by Microsoft Research Asia in 2015, which performed well in the large-scale ImageNet image recognition competition. The MIL-NCE pre-trained S3D network [18] extracts features from videos to correct misalignments in narrated videos. The S3D, used as the backbone, implements 3D convolutions with 2D convolutions to reduce computational costs [19].

The PANNs [20] method has been proposed for the effective extraction of audio features from audio data. This network is trained on a large-scale AudioSet to recognize

audio patterns and is designed with systems such as Wavegram-CNN and Wavegram-Logmel-CNN, which allow for the effective extraction of diverse audio data characteristics. The Wavegram-CNN, a system for audio tagging in the time domain, incorporates a feature known as Wavegram to learn frequency information and employs 2D CNNs to capture time-frequency invariant patterns. The Wavegram-Logmel-CNN combines the frequency information of Wavegram with the traditionally used log mel spectrogram in audio analysis to extract a wide range of information across time and frequency domains.

A variety of methods has been suggested for learning the relationship between the video and text. Among them is the CLIP network [21], which is designed to identify the text that best describes a specific video. CLIP is a network for multi-modal zero-shot model training that was developed and made public by OpenAI in 2021. It is a notable example of the integration of computer vision and natural language processing. The multi-modal structure of CLIP uses more than one type of data to train for a specific objective function and employs a zero-shot learning strategy, which trains the model to classify data that it has not previously seen.

As previously mentioned, among the leading technologies for detecting highlights in videos are Moment-DETR [12], UMT [13], and QD-DETR [14]. Moment-DETR utilizes an encoder–decoder transformer architecture, while UMT is trained with multi-modal architecture that encompasses both video and audio data. QD-DETR employs cross-attention layers and a negative loss function to enhance the relevance between queries and video clips and can be trained on video-only or video–audio multi-modal data. The primary challenges in video highlight detection are moment retrieval and highlight detection tasks.

Moment retrieval is the process of finding moments in a video that are relevant to a provided natural language query. In the video highlight detection network, this involves pinpointing the start and end times of a segment within the temporal domain of the video sequence. Queries are typically about specific activities, and datasets often have a strong bias toward the beginning of the video rather than the end. The QVHighlights dataset, introduced in 2021, was designed to counter this bias by creating queries for multiple moments within the videos, effectively breaking the bias of focusing on the beginning of the videos. The results of a model's moment retrieval are dependent on the highlight annotations in the training data. If the training data's highlight segments are annotated as two seconds each, the model's predictions for moment retrieval will also reflect two-second intervals.

Highlight detection is the process of identifying interesting or important segments within a video [22–25]. Traditional datasets [26] do not provide personalized highlights as they lack queries related to specific video segments. However, video highlight detection models such as Moment-DETR, UMT, and QD-DETR [12–14] derive meaningful information from saliency score annotations, which rate the relevance between clips and queries. Thus, humans annotate the relevance between a query and a clip with saliency scores on an integer scale from 0 to 4 (Very Bad, Bad, Fair, Good, Very Good) [12].

Multi-modal learning is a method that utilizes data with diverse characteristics, not just one type of data. This is similar to the way humans use multiple sensory organs to solve problems. Typically, modalities refer to various data such as images, videos, text, audio, and more. Multi-modal learning is utilized in various application areas, with research being conducted to improve performance in autonomous vehicles, music analysis, speech recognition, natural language processing, and image caption generation, among others.

In short, the adopted methods for feature extraction from the training data include video feature extraction, audio feature extraction, and text feature extraction, all of which commonly utilize pre-trained models to extract features. The aforementioned video highlight detection methods employ a transformer-based network and share the common characteristic of multi-modal learning.

In this paper, we leverage the previously mentioned video highlight detection models and the interaction of various modalities. Moment-DETR learns from video and text data, while UMT and QD-DETR consider interactions between video, audio, and text data to

build robust video highlight detection models. Models such as Moment-DETR, UMT, and QD-DETR share a common characteristic of detecting segments related to queries about daily life, such as travel and cooking, within a video. However, for the first time, we apply these existing video highlight detection models to the task of traffic accident video classification and analyze the characteristics of the detailed design considerations during this application process.

3. Proposed Method

3.1. Cross-Modality Augmented Accident Classification

We compare and analyze performance in terms of cross-modality interaction in traffic accident video classification. QD-DETR [14] can be trained for accident video classification using video and text or video, audio, and text as inputs. We evaluate the performance of QD-DETR trained only with video and text against the performance when trained with video, audio, and text on our proposed CCHD. We explore the impact on performance when not only video and text but also audio modality are included as input in accident video classification.

In CCHD, UMT [13] and QD-DETR are compared and analyzed for their performance in detecting accident segments and predicting saliency scores by learning from video, audio, and text queries. UMT utilizes a bottleneck transformer structure in its cross-modal encoder to compress and expand video and audio features, reducing computational costs, and employs text queries in its transformer decoder. QD-DETR simply concatenates each of the video and audio features, utilizing a cross-attention layer to enhance their correlation with text queries. Experiments are conducted to determine how effectively fusing features from video, audio, and text queries can lead to superior performance in traffic accident detection.

3.2. Feature Extraction

The single-image-based Slowfast [16] does not use optical flow. Structurally, it captures semantic information through one pathway, while another pathway is used to detect fast movements within the video. The 2D Resnet [17] architecture has the advantage of preventing gradient vanishing even in neural networks of considerable depth. The depth of the network impacts accuracy, and a relatively deep network with 152 layers, the 2D Resnet, was utilized. The MIL-NCE pre-trained S3D [18] consists of separable 3D CNN, which uses 3D convolutions intended to separate spatial and temporal features. 3D CNNs are useful for linking temporal and spatial information. This paper applies these three video feature extraction methods to serve as input data for the video highlight detection network. Additionally, we examine the dependency of each model's performance on the methods of feature extraction from the input video data. The video features were extracted using pre-trained models. CS represents the features extracted through the CLIP and Slowfast models, while CR denotes the features obtained via the CLIP and 2D Resnet models. CM refers to the features extracted using the CLIP and MIL-Pre-trained S3D models.

3.3. Self-Attention and Cross-Attention

We evaluate the performance of highlight accident classification by training Moment-DETR [12] and QD-DETR [14] on DAD [27] and CCHD. For this purpose, both Moment-DETR and QD-DETR learn from video features. Additionally, they learn from the features of text queries. In Moment-DETR, the concatenated video and text features are inputted through self-attention. QD-DETR is an augmentation of the Moment-DETR framework with an added cross-attention layer. This experiment explores whether these enhancements contribute to improved performance in traffic accident classification. Moreover, we compare and analyze the performance for a specific query across all the datasets used in the experiments.

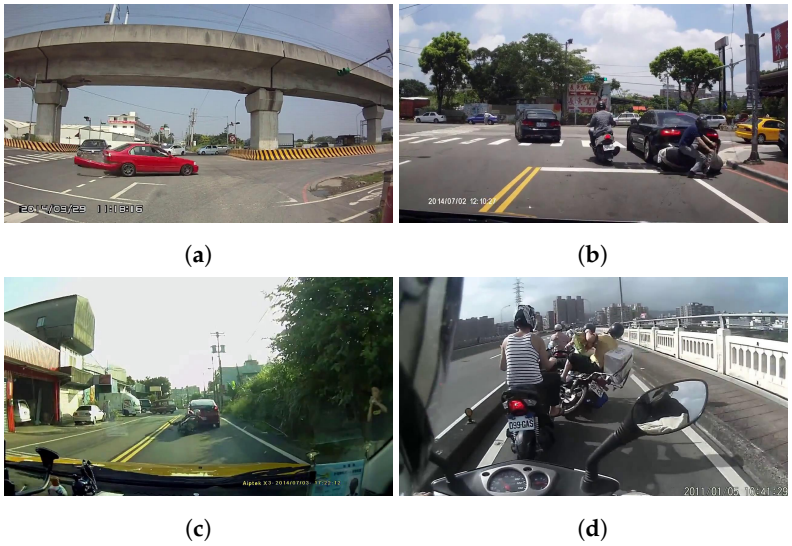
3.4. Negative Loss

We analyze the influence of QD-DETR's negative loss [14] on the traffic accident dataset through a performance comparison between QD-DETR and Moment-DETR [12]. This training scheme, which includes negative loss, prompts the model to consider the relevance to the query, thereby preventing predictions based purely on the inter-relationships among video clips [14]. In other words, the model should classify segments differently depending on whether the query is positive or negative. QD-DETR is an extension of the Moment-DETR structure with the addition of negative loss. We compare the performance of Moment-DETR and QD-DETR, both trained with video and text as inputs on DAD [27] and CCHD, in predicting the saliency scores of highlights. We investigate whether this negative loss improves performance by predicting lower saliency scores for segments irrelevant to the query.

4. Experiment

4.1. Augmented Reality-Driven Dataset

These datasets are devised for an augmented reality scenario where both video and audio, as well as video and text, are interactive. Through DAD [27], a total of 621 videos were compiled for the training dataset. Each video was set to 100 frames based on 25 fps to align with the experimental environment, and the presence of accidents was identified for each segment. Referring to the DADA [28] dataset, we organized four representative types of accidents. Figure 1 displays the accident scenes from DAD.



(**a**) (**b**)

(**c**) (**d**)

Figure 1. Example of accident scenes in DAD (queries are defined according to the situation of an accident scene). (**a**) Query: "Car hits car"; (**b**) query: "Car hits motorcycle"; (**c**) query: "Motorcycle hits car"; (**d**) query: "Motorcycle hits motorcycle".

A total of 478 videos, including audio, were compiled for the dataset using CCHD. The videos were structured to even lengths based on a 30 fps standard to accommodate the experimental environment, and the presence of accidents was identified for each segment. Figure 2 shows the accident scenes from CCHD. The types of queries defined are "Cars collide", "Car and motorcycle collide", "Car and bike collide", and "Accident". The types of accidents in the videos are identified, and the queries are defined accordingly. If a video's accident segment included audio of car horns, collision sounds, or people screaming, higher saliency scores (2 to 4) are assigned. Conversely, segments where an animal runs onto the road, a person is standing on the crosswalk, or a car slides on ice, which do not involve a direct collision with another object, are assigned lower saliency scores (0 to 1).

CCHD consists of 445 accident videos and 33 non-accident videos. To prevent imbalanced training when training models with this dataset, the train and test datasets were

structured to contain accident and non-accident data at a ratio of 8:2. By resolving the data imbalance, the predictive performance on new data was improved and bias in the traffic accident detection model was reduced. The CCHD and the raw data are available at Link_GoogleDrive_CCHD, https://drive.google.com/file/d/1XHsAsYV7YkqjKfvTT9oJIeWZ7ZZJHxmb/view?usp=sharing, accessed on 30 August 2023, and Link_Kaggle_Car CrashDataset, https://www.kaggle.com/datasets/sivoha/car-crash-dataset-russia-2022-2023, accessed on 15 June 2023.

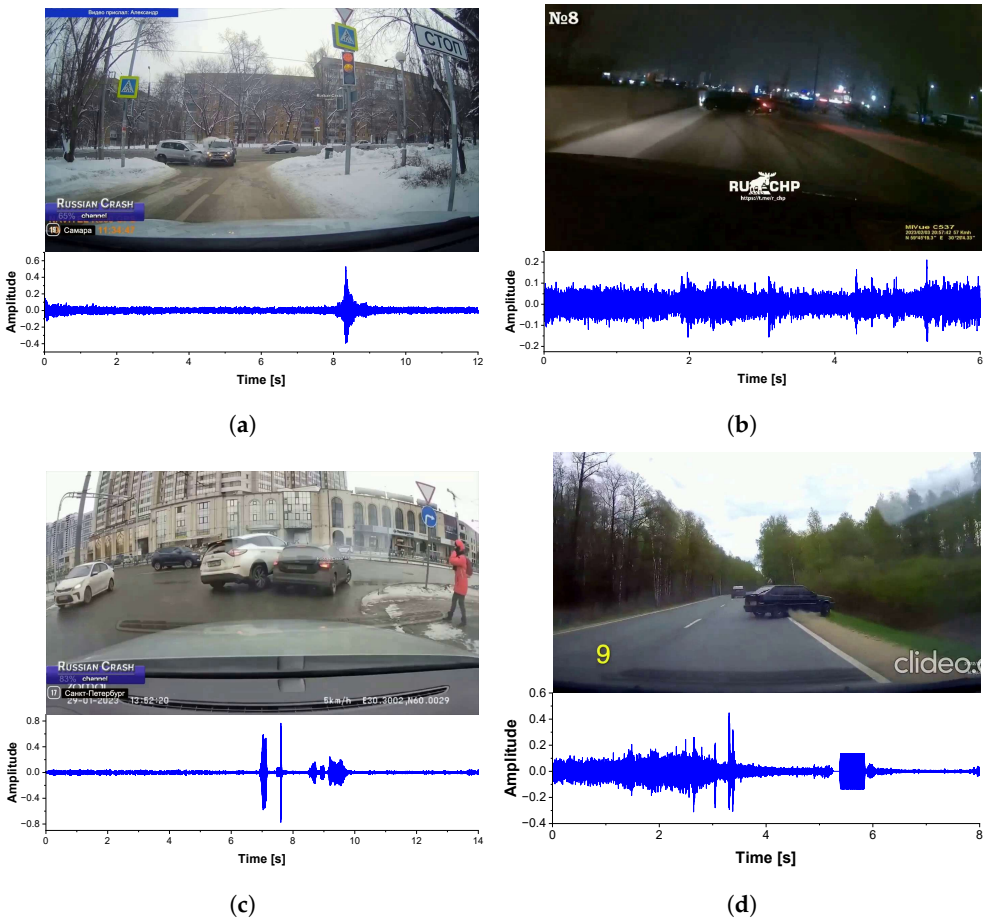


Figure 2. Example of accident scenes in CCHD; both video and audio waveform are presented (queries are defined based on the type of an accident). (**a**) Query: "Cars collide"; (**b**) query: "Accident"; (**c**) query: "Cars collide"; (**d**) query: "Accident".

The additional information that facilitates the classification of traffic accidents was referenced from the format of the QVHIGHLIGHTS dataset [12]. In Figure 3 and Table 1, QID refers to the identifier for the query describing the video, while VID is the identifier for the video itself. The duration indicates the length of the video, and relevant windows refer to the time intervals for the defined moments. Relevant clip IDs consist of indices of clips, which are split into 2 s segments, which are relevant to the query. Saliency scores represent the relevance of the clips indexed in relevant clip IDs to the query as integers ranging from 0 to 4, with higher values indicating greater relevance.

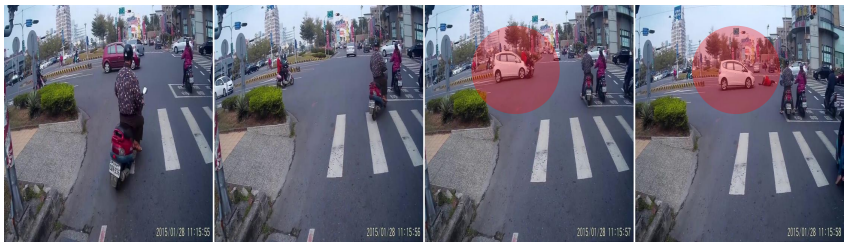


Figure 3. The four images are ordered by the sequence of occurrence from left to right (areas where the accident appears are highlighted with a red circle).

Table 1. An example of data annotation.

Component	Value
QID	8
Query	"Car hits motorcycle"
Duration	4
VID	000008
Relevant Clip IDs	[1]
Saliency Scores	[[4, 3, 4]]
Relevant Windows	[[2, 4]]

4.2. Experimental Setup

For the first time, the video highlight detection models [12–14] are applied to the field of traffic accident detection, and their usefulness in detecting accidents is demonstrated. The mAP (mean average precision) for moment retrieval is calculated using 10 predictions and the ground truth, while the mAP for highlight detection is computed with a single prediction and the ground truth. In order to evaluate the performance of highlight accident segment classification according to each model's feature extraction method, an accuracy metric for moment retrieval is added. Accuracy is defined as follows 1.

$$Accuracy = TP + TN/(TP + TN + FP + FN) \quad (1)$$

In this context, TP is the number of accurately predicted highlighted segments, FP is the number of incorrectly predicted highlighted segments, TN is the number of correctly predicted non-highlighted segments, and FN is the number of incorrectly predicted non-highlighted segments. The prediction and the ground truth are classified into video segment clips of two-second intervals, with one indicating a highlight and zero indicating a non-highlight. Accuracy assesses the degree to which these two-second video clips are correctly predicted.

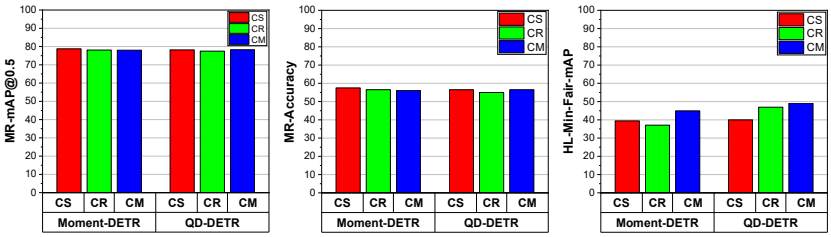
For the experiments, a GPU: GeForce RTX 3090, CPU: Intel(R) Core(TM) i9-10900K @ 3.70 GHz, running on Ubuntu 22.04, with CUDA version 1.15.0 was utilized. For Moment-DETR, UMT, and QD-DETR, the default settings were used for the main parameters, and the training was conducted for a maximum of 100 epochs.

4.3. Result

Firstly, we experimented to assess the performance of the traffic accident detection of Moment-DETR [12] and QD-DETR [14]. These models utilize video and text features as inputs and are evaluated on DAD [27]. The results, which are presented in Tables 2–4, show the performance of Moment-DETR and QD-DETR with three different feature extraction methods. Figures 4 and 5 visualize these results.

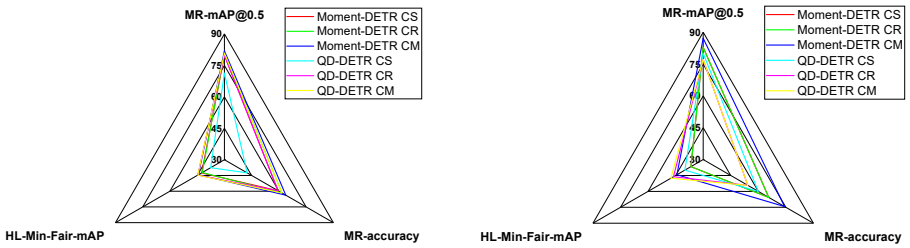
Table 2. Comparing Moment-DETR and QD-DETR with video and text data (DAD).

Metric	Moment-DETR			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	78.75%	78.08%	78.00%	78.17%	77.50%	78.25%
MR-Accuracy	57.50%	56.50%	56.00%	56.50%	55.00%	56.50%
HL-Min-Fair-mAP	39.42%	37.08%	44.92%	40.00%	46.92%	49.08%



(**a**) (**b**) (**c**)

Figure 4. Visualizing the performances of Moment-DETR and QD-DETR (DAD). (**a**) MR-mAP@0.5; (**b**) MR-Accuracy; (c) HL-Min-Fair-mAP.



(**a**) (**b**)

Figure 5. Visualizing the performances of Moment-DETR and QD-DETR in terms of evaluation metrics, MR-mAP@0.5, MR-Accuracy, and HL-Min-Fair-mAP (DAD). (**a**) Query: "Car hits car"; (**b**) query: "Motorcycle hits motorcycle".

Table 3. Comparing Moment-DETR and QD-DETR using video and text data (query: "Car hits car", DAD).

Query	"Car Hits Car"					
Metric	Moment-DETR			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	79.79%	80.85%	81.91 %	71.28%	79.79%	80.85%
MR-Accuracy	59.57%	61.70%	63.83%	42.55%	59.57%	61.70%
HL-Min-Fair-mAP	42.55%	42.20%	44.33%	37.59%	44.68%	44.68%

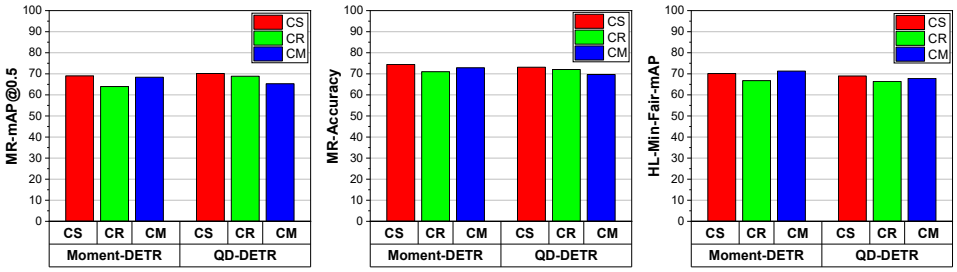
Table 4. Comparing Moment-DETR and QD-DETR using video and text data (query: "Motorcycle hits motorcycle", DAD).

Query	"Motorcycle Hits Motorcycle."					
Metric	Moment-DETR			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	82.86%	82.86%	87.14%	80.00%	77.14%	77.14%
MR-Accuracy	65.71%	65.71%	74.29%	60.00%	54.29%	54.29%
HL-Min-Fair-mAP	36.67%	36.67%	44.29%	40.00%	46.19%	47.62%

Secondly, we conduct an experiment where we apply Moment-DETR and QD-DETR to CCHD, and while we continue to use video and text features as inputs, the dataset change allows us to observe their accident classification performance in a different context. Tables 5–7 present the performance of Moment-DETR and QD-DETR with three different feature extraction methods, and Figures 6 and 7 provide corresponding visual representations.

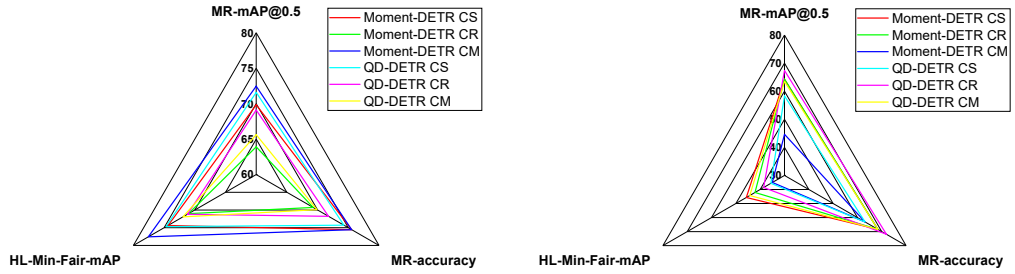
Table 5. Comparing Moment-DETR and QD-DETR using video and text data (CCHD).

Metric	Moment-DETR			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	69.02%	63.96%	68.37%	70.17%	68.85%	65.31%
MR-Accuracy	74.47%	71.01%	72.87%	73.14%	72.07%	69.68%
HL-Min-Fair-mAP	70.13%	66.78%	71.30%	68.99%	66.39%	67.85%



(**a**) (**b**) (**c**)

Figure 6. Visualizing the performances of Moment-DETR and QD-DETR using video and text data (CCHD). (**a**) MR-mAP@0.5; (**b**) MR-Accuracy; (**c**) HL-Min-Fair-mAP.



(**a**) (**b**)

Figure 7. Visualizing the performances of Moment-DETR and QD-DETR in terms of evaluation metrics, MR-mAP@0.5, MR-Accuracy, and HL-Min-Fair-mAP (CCHD). (**a**) Query: "Cars collide"; (**b**) query: "Accident".

Table 6. Comparing Moment-DETR and QD-DETR using video and text data (query: "Cars collide", CCHD).

Query	"Cars Collide"					
Metric	Moment-DETR			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	69.84%	63.91%	72.47%	71.59%	69.10%	65.62%
MR-Accuracy	75.55%	69.28%	75.55%	74.29%	71.79%	69.91%
HL-Min-Fair-mAP	74.36%	71.04%	77.58%	74.70%	71.20%	71.93%

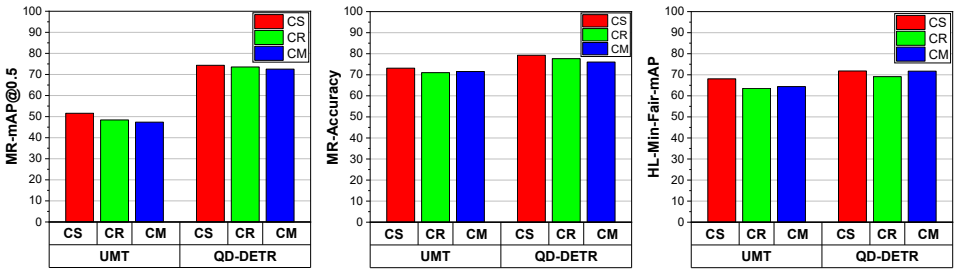
Table 7. Comparing Moment-DETR and QD-DETR using video and text data (query: "Accident", CCHD).

Query	"Accident"					
Metric	Moment-DETR			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	64.29%	64.29%	44.64%	58.33%	67.38%	63.51%
MR-Accuracy	68.42%	68.42%	63.16%	63.16%	71.93%	68.42%
HL-Min-Fair-mAP	45.71%	42.14%	35.00%	35.60%	38.57%	44.29%

Lastly, we perform an experiment with UMT [13] and QD-DETR on CCHD. This experiment utilized the interaction of video, audio, and text features. Tables 8–10 display the performance of UMT and QD-DETR with three different feature extraction methods, and Figures 8 and 9 visually depict these results.

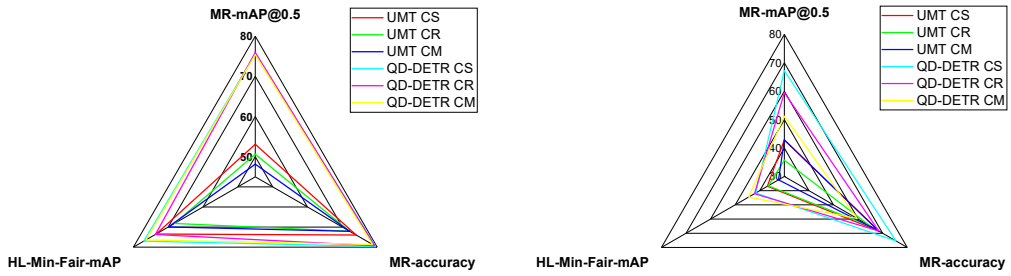
Table 8. Comparing UMT and QD-DETR using video, audio, and text data (CCHD).

Metric	UMT			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	51.58%	48.42%	47.37%	74.34%	73.55%	72.54%
MR-Accuracy	73.14%	71.01%	71.54%	79.26%	77.66%	76.06%
HL-Min-Fair-mAP	68.07%	63.55%	64.40%	71.80%	69.13%	71.73%



(**a**) (**b**) (**c**)

Figure 8. Visualizing the performances of UMT and QD-DETR using video, audio, and text data (CCHD). (**a**) MR-mAP@0.5; (**b**) MR-Accuracy; (**c**) HL-Min-Fair-mAP.



(**a**) (**b**)

Figure 9. Visualizing the performances of UMT and QD-DETR in terms of evaluation metrics, MR-mAP@0.5, MR-Accuracy, and HL-Min-Fair-mAP (CCHD). (**a**) Query: "Cars collide"; (**b**) query: "Accident".

Table 9. Comparing UMT and QD-DETR using video, audio, and text data (query: "Cars collide", CCHD).

Query	"Cars Collide"					
Metric	UMT			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	53.09%	50.62%	48.15%	75.57%	75.97%	75.41%
MR-Accuracy	73.98%	72.10%	72.10%	79.62%	79.31%	78.68%
HL-Min-Fair-mAP	73.46%	68.22%	69.94%	77.01%	73.69%	76.42%

Table 10. Comparing UMT and QD-DETR using video, audio, and text data (query: "Accident", CCHD).

Query	"Accident"					
Metric	UMT			QD-DETR		
	CS	*CR*	*CM*	*CS*	*CR*	*CM*
MR-mAP@0.5	42.86%	35.71%	42.86%	67.26%	59.52%	51.19%
MR-Accuracy	68.42%	64.91%	68.42%	75.44%	68.42%	59.65%
HL-Min-Fair-mAP	36.90%	36.55%	32.38%	41.67%	42.14%	44.52%

5. Discussion

5.1. Cross-Modality Interaction

We designated QD-DETR(Video) as the model that uses video and text features as inputs, and QD-DETR(Video+Audio) as the model that takes video, audio, and text features as inputs. In Tables 5 and 8, the QD-DETR(Video+Audio) outperforms QD-DETR(Video) across all evaluation metrics. Notably, QD-DETR(Video+Audio) CM shows a 7.23% higher performance in MR-mAP@0.5, a 6.38% increase in MR-Accuracy, and a 3.88% improvement in HL-Min-Fair-mAP compared to QD-DETR(Video) CM.

In Figure 10, the accident segments of the video contain sounds such as horns, crashes, and human screams. Therefore, such audio features from the accident segments, when combined with video and text features, enhance the performance of traffic accident segment prediction. Moreover, it has been demonstrated that training QD-DETR with video and audio features from the existing QVHIGHLIGHTS dataset [12] results in better performance than training with only video features [14].

In Table 8, QD-DETR CS outperforms UMT CS with a 22.76% higher MR-mAP@0.5, a 6.12% increase in MR-Accuracy, and a 3.73% improvement in HL-Min-Fair-mAP. QD-DETR CR exceeds UMT CR by 25.13% in MR-mAP@0.5, by 6.65% in MR-Accuracy, and by 5.58% in HL-Min-Fair-mAP. Furthermore, QD-DETR CM surpasses UMT CM with a 25.17% higher MR-mAP@0.5, a 4.52% increase in MR-Accuracy, and a 7.33% improvement in HL-Min-Fair-mAP.

In Table 9, QD-DETR demonstrates superior performance in detecting traffic accidents with the query ("Cars collide") across all evaluation metrics when compared to UMT. QD-DETR CR boasts a 75.97% MR-mAP@0.5, QD-DETR CS shows a 79.62% in MR-Accuracy, and a 77.01% in HL-Min-Fair-mAP. In Table 10, QD-DETR CS exhibits superior performance in moment retrieval metrics for the detection of "Accident" traffic accidents when compared to UMT CS. Specifically, QD-DETR CS shows an improvement of 24.4% in MR-mAP@0.5, 7.02% in MR-Accuracy, and 4.77% in HL-Min-Fair-mAP over UMT CS.

In video highlight detection, multi-modal learning, which efficiently fuses data with different characteristics, is crucial. In common, UMT [13] and QD-DETR [14] have a cross-attention structure where video and audio features are defined as query and text features as key and value. The two models differ in how they process video and audio features. Figure 11 shows the predictions from UMT and QD-DETR. Instead of combining video and audio features produced by self-attention within UMT's bottleneck transformer,

QD-DETR's simple concatenation approach to processing video and audio features shows superior performance. Encoding video and audio features allows the module that simply concatenates video and audio to automatically activate audio information [29]. Although the bottleneck transformer of UMT reduces computational costs by compressing and expanding video and audio features through multi-head attention, the loss of feature information could be a cause of performance degradation.

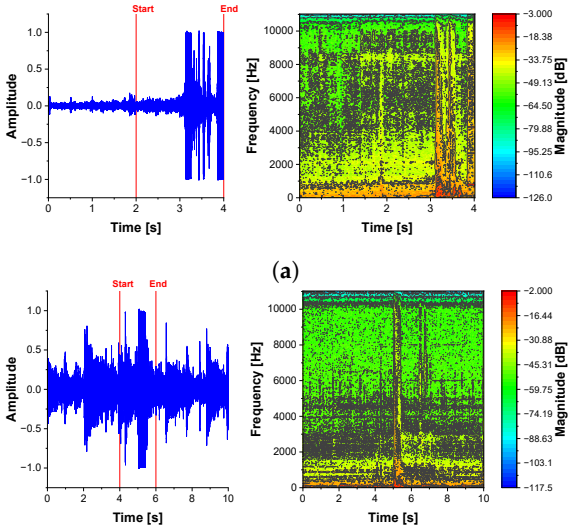


Figure 10. Visualizing the waveform of audio samples in CCHD. The start line represents the beginning time of a traffic accident, and the end line represents the ending time of the traffic accident. (**a**) Video ID: 000239; (**b**) video ID: 000242.

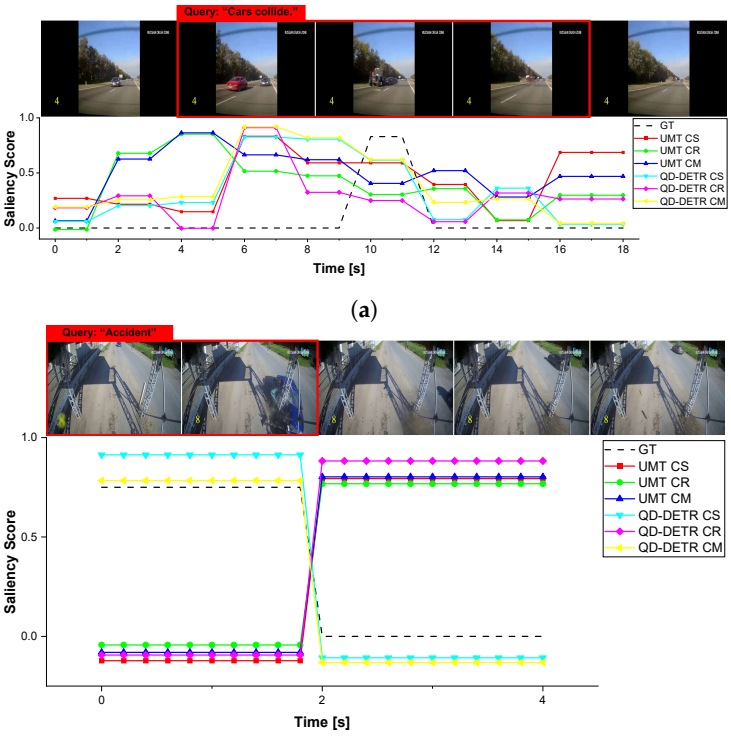


Figure 11. Visualizing the prediction results of UMT and QD-DETR using video, audio, and text features (CCHD). (**a**) Query: "Cars collide"; (**b**) query: "Accident".

5.2. Feature Extraction

The methods for extracting features from video and text play a crucial role in improving the performance of traffic accident detection. Notably, CLIP [21] is essential as it effectively combines visual and linguistic information by leveraging the interaction between images and the text.

In Table 2, Moment-DETR CS achieves 78.75% in MR-mAP@0.5 and 57.50% in MR-Accuracy, which is slightly higher, by 0.75% in MR-mAP@0.5 and by 1.5% in MR-Accuracy, compared to Moment-DETR CM. Additionally, the difference between QD-DETR CS and CM is only 0.08% in MR-mAP@0.5, while MR-Accuracy is identical at 56.50%.

In Table 5, Moment-DETR CS scores 69.02% in MR-mAP@0.5 and 74.47% in MR-Accuracy, which is 5.06% higher in MR-mAP@0.5 and 3.46% higher in MR-Accuracy than Moment-DETR CR. Furthermore, In Table 5, among the three feature extraction methods, QD-DETR CS shows superior performance across all metrics. QD-DETR CS is higher by 4.86% in MR-mAP@0.5, by 3.46% in MR-Accuracy, and by 1.14% in HL-Min-Fair-mAP than QD-DETR CM.

In Table 8, among the three feature extraction methods, UMT and QD-DETR CS exhibit the most outstanding performance across all metrics. UMT CS scores 3.16% higher in MR-mAP@0.5, 4.52% higher in HL-Min-Fair-mAP, and 2.13% higher in MR-Accuracy than UMT CR. Furthermore, QD-DETR CS outperforms QD-DETR CR by 0.79% in MR-mAP@0.5, by 2.67% in HL-Min-Fair-mAP, and by 1.6% in MR-Accuracy.

The combination of CLIP [21] and Slowfast [16] features demonstrates an overwhelmingly synergistic effect compared to when Resnet [17] features or MIL-NCE pre-trained S3D [18] features are combined with CLIP. CLIP interprets the semantics of images in a linguistic manner, while Slowfast recognizes semantic and temporal motions. By using both methods, we can obtain an integrated representation of language, objects, and motion in videos, and we have proven that this enhances the performance of traffic accident detection.

In Table 2, among the three feature extraction methods, QD-DETR CM shows slightly higher performance across all evaluation metrics, with 78.25% in MR-mAP@0.5, 49.08% in HL-Min-Fair-mAP, and 56.50% in MR-Accuracy. Additionally, in HL-Min-Fair-mAP, QD-DETR CM scores 9.08% higher than QD-DETR CS and 2.16% higher than QD-DETR CR.

In CCHD with video lengths over 4 s, the feature extraction methods of CLIP and MIL-NCE pre-trained S3D do not exhibit superior performance across all evaluation metrics compared to other methods. The S3D network from MIL-NCE pre-trained S3D, which creates 3D convolutions from 2D convolutions, does not show high performance in datasets with relatively longer video lengths than the videos in DAD [19].

In all experiments, the feature extraction methods using CLIP and 2D Resnet generally do not show superior performance compared to other methods, and while 2D Resnet extracts image information from videos, it loses temporal information due to the two-dimensional convolution [30]. Therefore, there is a limitation in that simply combining features extracted by CLIP and 2D Resnet does not adequately consider the temporal aspect.

5.3. Comparing Self-Attention and Cross-Attention

Moment-DETR [12] and QD-DETR [14] differ in how they process video and text features. In the Moment-DETR, the features of the video and text are simply concatenated to form the query, key, and value in the transformer encoder's self-attention mechanism. In the QD-DETR, within the cross-attention transformer, the video features are defined as the query, while the text features are the key and value. The attention mechanism takes the dot-product of the query and key matrices to produce an attention score, which is then the dot-product with the value. The attention score is calculated by taking the dot-product of the query with the key, allowing only those with high similarity to prevail. In video highlight detection, the relevance between video and text is crucial. Here, it is demonstrated which of the two models' methods is more efficient in processing video and text features.

In Table 2, Moment-DETR CS exhibits the most outstanding performance, achieving 78.75% in MR-mAP@0.5 and 57.50% in MR-Accuracy. In Table 3, Moment-DETR CM

achieves 81.91% in MR-mAP@0.5 and 63.83% in MR-Accuracy, which is slightly higher than QD-DETR CM by 1.06% in MR-mAP@0.5 and by 2.13% in MR-Accuracy. In Table 4, Moment-DETR CM shows 87.14% in MR-mAP@0.5 and 74.29% in MR-Accuracy, surpassing QD-DETR CM by 10% in MR-mAP@0.5 and by 20% in MR-Accuracy.

In Table 2, Moment-DETR slightly outperforms QD-DETR in the moment retrieval evaluation metrics. Additionally, it can be observed that in Tables 3 and 4, Moment-DETR performs better than QD-DETR. In a dataset characterized by the absence of audio and relatively short video lengths, a structure that simply concatenates video and text has been shown to surpass the performance of a structure that utilizes cross-attention between the video and text in the detection of traffic accident segments.

In Table 5, QD-DETR CS shows the best performance with 70.17% in MR-mAP@0.5 and Moment-DETR CS leads in MR-accuracy with 74.47%. In Table 6, Moment-DETR CM records 72.47% in MR-mAP@0.5 and 75.55% in MR-accuracy, outperforming QD-DETR CM by 6.85% in MR-mAP@0.5 and by 5.64% in MR-accuracy. Moreover, in Table 7, QD-DETR CR scores 67.38% in MR-mAP@0.5 and 71.93% in MR-accuracy, which is higher than Moment-DETR CR by 3.09% in MR-mAP@0.5 and by 3.51% in MR-accuracy.

In Table 5, Moment-DETR CS and QD-DETR CS exhibit similar performance in the moment retrieval evaluation metrics. In Tables 6 and 7, the "Cars collide" accident segment detection performance is best with Moment-DETR CM, while the "Accident" accident segment detection performance is best with QD-DETR CR. In Tables 5–7, CCHD shows that the performance differences among the models in the moment retrieval metrics are not considerable. Our study demonstrates that the performance of traffic accident detection models varies depending on the query dataset and the feature extraction method used, and we present the optimal traffic accident detection model for different types of accidents based on the feature extraction method.

5.4. Negative Loss

When comparing the HL-Min-Fair-mAP performance metrics of Moment-DETR and QD-DETR for feature extraction in Table 2, QD-DETR exhibits superior performance over Moment-DETR. Specifically, QD-DETR CR achieves 46.92% in HL-Min-Fair-mAP, which is 9.84% higher than Moment-DETR CR. Additionally, QD-DETR CM scores 49.08%, which is 4.16% higher than Moment-DETR CM, and QD-DETR CS reaches 40.00%, slightly outperforming Moment-DETR CS by 0.58%.

In Table 3, QD-DETR CR leads with 44.68% in HL-Min-Fair-mAP, which is 2.48% higher than Moment-DETR CR. In Table 4, QD-DETR CM achieves the highest score of 47.62% in HL-Min-Fair-mAP, which is 3.33% higher than Moment-DETR CM.

Figure 12 shows the predictions from Moment-DETR [12] and QD-DETR [14]. In DAD [27], videos that are not in the accident segment receive a low saliency score as they do not match the query. QD-DETR's negative loss function learns from irrelevant negative video-query pairs within DAD. This approach reinforces the lowest saliency scores for segments unrelated to accidents, aiding in the enhancement of saliency score prediction performance.

In Table 5, Moment-DETR slightly outperforms QD-DETR in the HL-Min-Fair-mAP performance metric across different feature extractions. Moment-DETR CM scores 71.30%, which is 3.45% higher than QD-DETR CM. Additionally, Moment-DETR CS achieves 70.13%, which is 1.14% higher than QD-DETR CS, and Moment-DETR CR scores 66.78%, slightly outperforming QD-DETR CR by 0.39%. In Table 6, Moment-DETR CM leads with 77.58% in HL-Min-Fair-mAP, which is 5.65% higher than QD-DETR CM. In Table 7, Moment-DETR CS is the most outstanding, scoring 45.71% in HL-Min-Fair-mAP, which is 10.11% higher than QD-DETR CS.

Figure 13 shows the predictions from Moment-DETR and QD-DETR. In CCHD, the queries also pertain to accidents, but videos without actual collisions between objects are assigned lower saliency scores. However, the number of irrelevant negative video-query pairs in this dataset is small, which may have prevented QD-DETR's negative loss from being sufficiently trained.

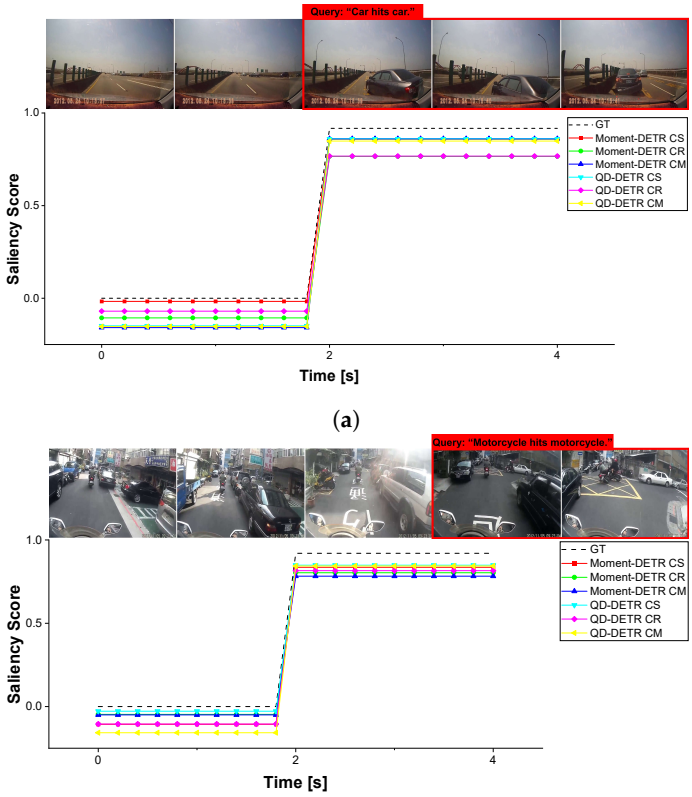


Figure 12. Visualizing the prediction results of Moment-DETR and QD-DETR using video and text features (DAD). (**a**) Query: "Car hits car"; (**b**) query: "Motorcycle hits motorcycle".

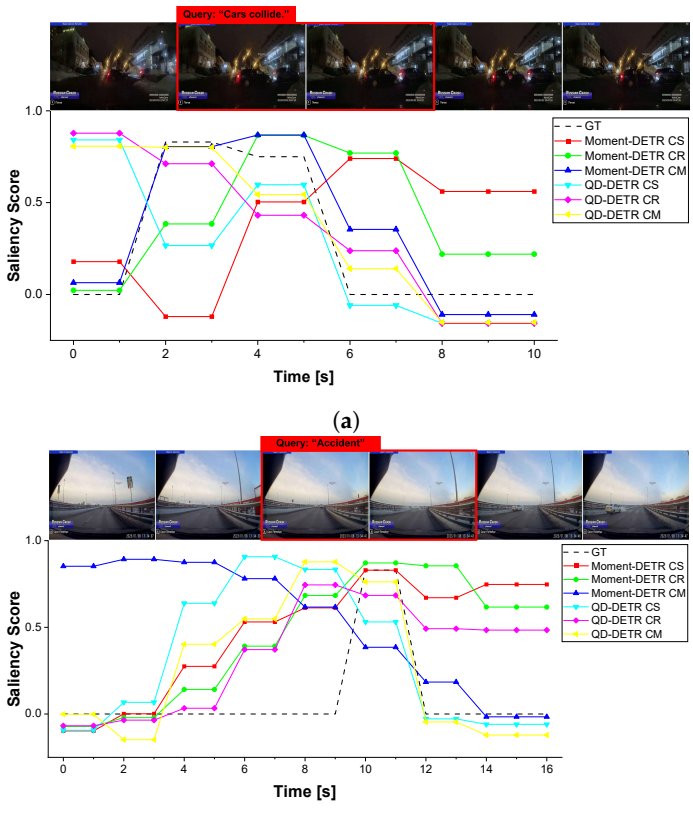


Figure 13. Visualizing the prediction results of Moment-DETR and QD-DETR using video and text features (CCHD). (**a**) Query: "Cars collide"; (**b**) query: "Accident".

5.5. Relevance and Expectation

Overall, we have reached several significant insights. Identifying types of accidents allows us to analyze patterns related to those types of accidents, which can aid in accident prevention. Additionally, data such as the frequency, timing, and location of accidents according to type can be collected and used for record-keeping and statistical analysis. We proved that the video highlight detection network [12–14] operates effectively in traffic accident detection. This could potentially facilitate the implementation of an augmented reality system capable of real-time traffic accident detection and response based on cross-modality. For this, it will be important to extend the experiment with other accident categories (e.g., head-on collision, side collision, rear-end collision, etc.) and other methods of different architectures in order to discover more detailed insights.

6. Conclusions

We are the first to apply existing video highlight detection techniques to traffic accident classification and analyze their performance. Additionally, we have generated supplementary information necessary for video highlight detection within the traffic accident dataset. Our paper analyzes the impact of performance when utilizing heterogeneous information from video or a combination of video and audio. The performance of traffic accident detection across models is compared and analyzed based on three different video feature extraction methods, and the optimal video feature extraction method for traffic accident classification is presented for each model. We also examine the differences between self-attention and cross-attention when combining video and text information, as well as consider the effect of the existing negative loss on the performance of traffic accident detection.

In the future, it will be possible to implement a real-time traffic accident detection and response system based on cross-modality augmented reality. Cameras mounted on vehicles, along with query generation modules, will be able to show drivers augmented reality in real-time. Additionally, the augmented reality system will be able to accurately detect accidents in the real world by utilizing the interaction between video and audio data. Once an accident is detected, the system can immediately display the time of the accident, the type of accident, and the saliency to the driver to help prevent secondary accidents.

In the field of traffic accident detection, efficiently combining video, audio, and text features is crucial. Moving forward, we will research designing an efficient model structure that accommodates these three modalities. We plan to expand our models by applying video with LLM and video with LAM to create a comprehensive system. Through the use of a large language model, we aim to develop a system that can accurately recognizes and interprets text information within videos, such as road signs, traffic lights, and vehicle license plates, to gain a detailed understanding of accident scenarios. Additionally, we intend to enhance the accuracy of accident classification by analyzing various auditory signals from the road, such as the patterns of screeching brakes and collision sounds, through a large auditory model.

Author Contributions: C.O. conceived, designed, and performed the experiments; C.O. and Y.B. analyzed the data and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2022R1A5A8026986 and No. 2022R1F1A1073745), the Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0020536, HRD Program for Industrial Innovation), and Chungbuk National University BK21 program (2021).

Data Availability Statement: Data is contained within the article. The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CCHD	Car Crash Highlights Dataset
MIL-NCE	Multiple Instance Learning and Noise Contrastive Estimation
S3D	Separable 3D CNN
QID	query identification
VID	video identification
MR	moment retrieval
HL	highlight
mAP	mean average precision
CS	CLIP and Slowfast
CR	CLIP and 2D Resnet
CM	CLIP and Mil-NCE pre-trained S3D
LLM	Large Language Model
LAM	Large Auditory Model

References

1. Traffic Accident Analysis System. OECD Countries Traffic Accident Incidence. 2023. Available online: https://taas.koroad.or.kr/sta/acs/gus/selectOecdTfcacd.do?menuId=WEB_KMP_OVT_MVT_TAC_OAO (accessed on 1 February 2023).
2. Tian, D.; Zhang, C.; Duan, X.; Wang, X. An automatic car accident detection method based on cooperative vehicle infrastructure systems. *IEEE Access* **2019**, *7*, 127453–127463. [CrossRef]
3. Razzaq, S.; Dar, A.R.; Shah, M.A.; Khattak, H.A.; Ahmed, E.; El-Sherbeeny, A.M.; Lee, S.M.; Alkhaledi, K.; Rauf, H.T. Multi-factor rear-end collision avoidance in connected autonomous vehicles. *Appl. Sci.* **2022**, *12*, 1049. [CrossRef]
4. Zhang, Y.; Sung, Y. Traffic Accident Detection Using Background Subtraction and CNN Encoder–Transformer Decoder in Video Frames. *Mathematics* **2023**, *11*, 2884. [CrossRef]
5. Alkhaiwani, A.H.; Alsamani, B.S. A Framework and IoT-Based Accident Detection System to Securely Report an Accident and the Driver's Private Information. *Sustainability* **2023**, *15*, 8314. [CrossRef]
6. Hozhabr Pour, H.; Li, F.; Wegmeth, L.; Trense, C.; Doniec, R.; Grzegorzek, M.; Wismüller, R. A machine learning framework for automated accident detection based on multimodal sensors in cars. *Sensors* **2022**, *22*, 3634. [CrossRef]
7. Yao, Y.; Xu, M.; Wang, Y.; Crandall, D.J.; Atkins, E.M. Unsupervised traffic accident detection in first-person videos. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Macau, China, 3–8 November 2019; pp. 273–280.
8. Basheer Ahmed, M.I.; Zaghdoud, R.; Ahmed, M.S.; Sendi, R.; Alsharif, S.; Alabdulkarim, J.; Albin Saad, B.A.; Alsabt, R.; Rahman, A.; Krishnasamy, G. A real-time computer vision based approach to detection and classification of traffic incidents. *Big Data Cogn. Comput.* **2023**, *7*, 22. [CrossRef]
9. Robles-Serrano, S.; Sanchez-Torres, G.; Branch-Bedoya, J. Automatic detection of traffic accidents from video using deep learning techniques. *Computers* **2021**, *10*, 148. [CrossRef]
10. Khan, S.W.; Hafeez, Q.; Khalid, M.I.; Alroobaea, R.; Hussain, S.; Iqbal, J.; Almotiri, J.; Ullah, S.S. Anomaly detection in traffic surveillance videos using deep learning. *Sensors* **2022**, *22*, 6563. [CrossRef] [PubMed]
11. Pradana, H. An end-to-end online traffic-risk incident prediction in first-person dash camera videos. *Big Data Cogn. Comput.* **2023**, *7*, 129. [CrossRef]
12. Lei, J.; Berg, T.L.; Bansal, M. Detecting moments and highlights in videos via natural language queries. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 11846–11858.
13. Liu, Y.; Li, S.; Wu, Y.; Chen, C.W.; Shan, Y.; Qie, X. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3042–3051.
14. Moon, W.; Hyun, S.; Park, S.; Park, D.; Heo, J.P. Query-dependent video representation for moment retrieval and highlight detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 23023–23033.
15. Li, L.; Chen, Y.C.; Cheng, Y.; Gan, Z.; Yu, L.; Liu, J. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv* **2020**, arXiv:2005.00200.
16. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
18. Miech, A.; Alayrac, J.B.; Smaira, L.; Laptev, I.; Sivic, J.; Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9879–9889.

19. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.
20. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]
21. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
22. Hirasawa, K.; Maeda, K.; Ogawa, T.; Haseyama, M. Detection of Important Scenes in Baseball Videos via a Time-Lag-Aware Multimodal Variational Autoencoder. *Sensors* **2021**, *21*, 2045. [CrossRef] [PubMed]
23. Nergård Rongved, O.A.; Stige, M.; Hicks, S.A.; Thambawita, V.L.; Midoglu, C.; Zouganeli, E.; Johansen, D.; Riegler, M.A.; Halvorsen, P. Automated event detection and classification in soccer: The potential of using multiple modalities. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 1030–1054. [CrossRef]
24. Tseng, S.M.; Yeh, Z.T.; Wu, C.Y.; Chang, J.B.; Norouzi, M. Video Scene Detection Using Transformer Encoding Linker Network (TELNet). *Sensors* **2023**, *23*, 7050. [CrossRef] [PubMed]
25. Park, J.H.; Mahmoud, M.; Kang, H.S. Conv3D-based video violence detection network using optical flow and RGB data. *Sensors* **2024**, *24*, 317. [CrossRef] [PubMed]
26. Garcia del Molino, A.; Gygli, M. Phd-gifs: Personalized highlight detection for automatic gif creation. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 26 October 2018; pp. 600–608.
27. Chan, F.H.; Chen, Y.T.; Xiang, Y.; Sun, M. Anticipating accidents in dashcam videos. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part IV 13; Springer: Berlin/Heidelberg, Germany, 2017; pp. 136–153.
28. Fang, J.; Yan, D.; Qiao, J.; Xue, J.; Yu, H. DADA: Driver attention prediction in driving accident scenarios. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 4959–4971. [CrossRef]
29. Hong, F.T.; Huang, X.; Li, W.H.; Zheng, W.S. Mini-net: Multiple instance ranking network for video highlight detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XIII 16, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 345–360.
30. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.