applied sciences

MDPI

Article

EAR-CCPM-Net: A Cross-Modal Collaborative Perception Network for Early Accident Risk Prediction

Wei Sun *, Lili Nurliyana Abdullah, Fatimah Binti Khalid and Puteri Suhaiza Binti Sulaiman

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia; liyana@upm.edu.my (L.N.A.); fatimahk@upm.edu.my (F.B.K.); psuhaiza@upm.edu.my (P.S.B.S.)

* Correspondence: gs66334@student.upm.edu.my

Abstract

Early traffic accident risk prediction in complex road environments poses significant challenges due to the heterogeneous nature and incomplete semantic alignment of multimodal data. To address this, we propose a novel Early Accident Risk Cross-modal Collaborative Perception Mechanism Network (EAR-CCPM-Net) that integrates hierarchical fusion modules and cross-modal attention mechanisms to enable semantic interaction between visual, motion, and textual modalities. The model is trained and evaluated on the newly constructed CAP-DATA dataset, incorporating advanced preprocessing techniques such as bilateral filtering and a rigorous MINI-Train-Test sampling protocol. Experimental results show that EAR-CCPM-Net achieves an AUC of 0.853, AP of 0.758, and improves the Time-to-Accident ($TTA_{0.5}$) from 3.927 s to 4.225 s, significantly outperforming baseline methods. These findings demonstrate that EAR-CCPM-Net effectively enhances early-stage semantic perception and prediction accuracy, providing an interpretable solution for real-world traffic risk anticipation.

Keywords: multimodal fusion; traffic accident prediction; early risk perception; semantic interaction

Academic Editor: Suchao Xie

Received: 26 July 2025
Revised: 21 August 2025
Accepted: 23 August 2025
Published: 24 August 2025

Citation: Sun, W.; Abdullah, L.N.; Khalid, F.B.; Suhaiza Binti Sulaiman, P. EAR-CCPM-Net: A Cross-Modal Collaborative Perception Network for Early Accident Risk Prediction. *Appl. Sci.* **2025**, *15*, 9299. https://doi.org/10.3390/app15179299

1. Introduction

Road traffic accidents represent a persistent global public health and economic burden. According to the World Health Organization (WHO), approximately 1.35 million people lose their lives annually due to road accidents, with another 20 to 50 million suffering from non-fatal injuries, many of which result in long-term disabilities [1,2]. The situation is particularly severe in low- and middle-income countries, where inadequate infrastructure, limited emergency response, and high vehicular density exacerbate road safety issues [3]. The increasing complexity of modern traffic environments marked by diverse vehicle types, dense interactions, and rapid dynamics has intensified the need for intelligent systems capable of anticipating and mitigating potential accidents before they occur.

Recent advances in deep learning and sensor technologies have enabled the use of rich, multimodal data for traffic safety research. This includes video frames, vehicle trajectories, optical flow, environmental context, and map-based semantics, offering new opportunities for early accident risk prediction [4–6]. Particularly, early risk prediction, as in Figure 1, which focuses on detecting precursors several seconds (150 frames) before the actual accident occurs, is of critical importance to intelligent transportation systems and autonomous driving platforms [7,8]. Such prediction systems, if reliable and timely, can allow for proactive interventions that prevent or reduce the severity of accidents.

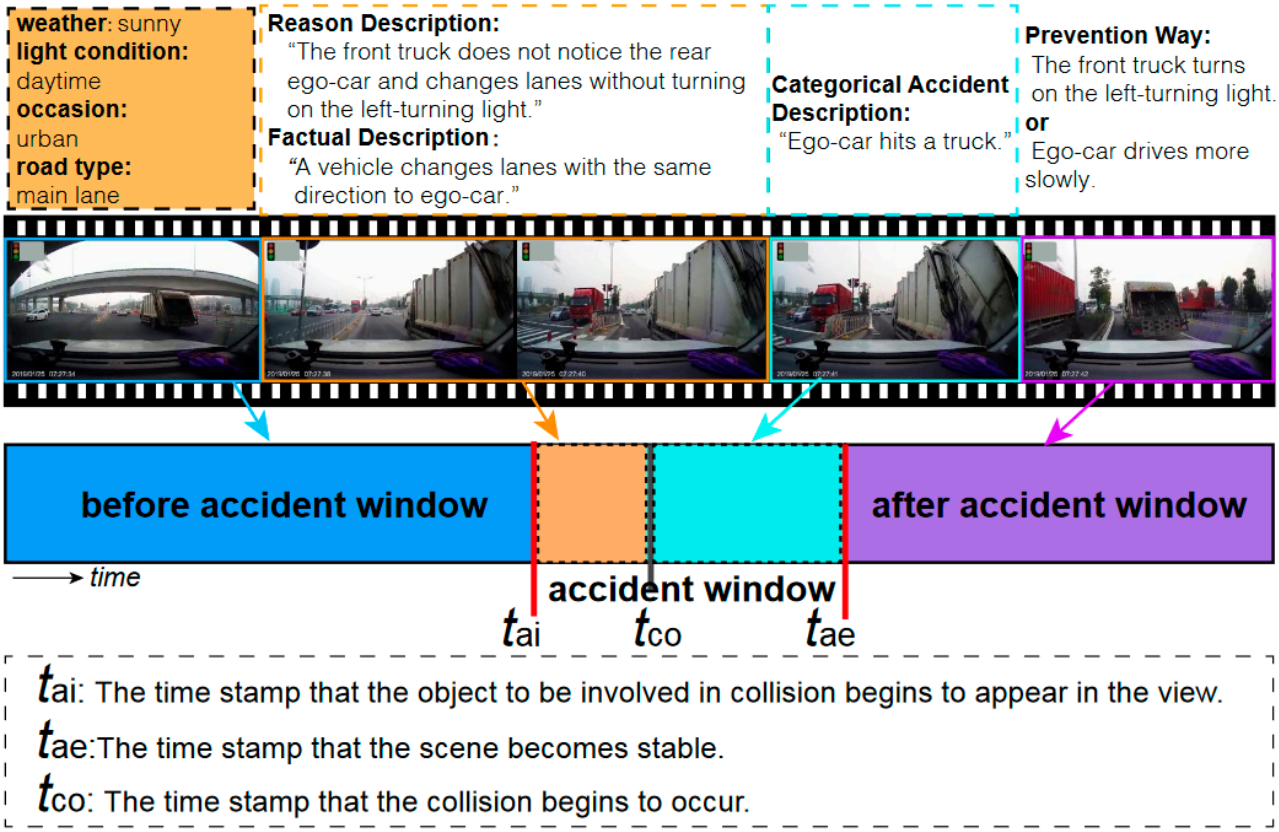

Figure 1. Illustration for different kinds of attribute annotations in CAP-DATA.

Despite this progress, a critical gap remains in how multimodal information is represented, aligned, and fused for early accident risk prediction. Existing approaches often treat each modality (e.g., visual frames, motion trajectories, semantic maps) in isolation or combine them using simple concatenation strategies. These modality-siloed designs fail to capture the complex semantic and temporal interdependencies across modalities, leading to suboptimal representations and degraded performance in early prediction tasks [9–11]. For instance, a sudden vehicle deceleration captured in trajectory data may be contextually explained by visual cues such as pedestrian emergence or lane obstructions information that can only be meaningfully interpreted through cross-modal interaction. Without collaborative perception mechanisms, such cues remain underutilized, weakening the model's understanding of causative factors and limiting its predictive accuracy in complex scenes. More importantly, few studies have explicitly addressed the early-stage prediction challenge, where precursor signals are inherently weak, sparse, and temporally distant from the accident event. The absence of interactive attention-guided fusion mechanisms further hinders the model's ability to identify and emphasize risk-relevant signals in dynamic, heterogeneous environments. This deficiency is particularly detrimental in the early risk prediction setting, where weak and sparse precursor signals must be precisely detected and temporally linked to subsequent high-risk outcomes.

To address this challenge, this study proposes a novel Early Accident Risk Cross-modal Collaborative Perception Mechanism Network (EAR-CCPM-Net). EAR-CCPM-Net introduces a hierarchical attention-guided cross-modal fusion framework that enables interactive semantic alignment between heterogeneous modalities specifically between visual and textual representations. By employing text-to-image fusion modules, the model emphasizes risk-relevant features and reinforces cross-modal consistency, thereby enhancing the model's capability to detect early risk cues and infer accident likelihood at an earlier stage.

The main contributions of this work are summarized as follows: (i) the study proposes EAR-CCPM-Net, a Cross-modal Collaborative Perception network designed for early accident risk prediction. The network incorporates hierarchical attention modules to fuse and align multimodal features dynamically; (ii) the study introduces an attention-guided text-to-image semantic fusion mechanism that enhances risk cue identification by linking temporal motion and contextual cues with visual observations in a unified feature space; (iii) the study demonstrates the effectiveness of EAR-CCPM-Net on benchmark multimodal traffic datasets, showing significant improvement in both prediction accuracy and temporal anticipation over existing unimodal and naïve multimodal fusion methods.

In the remainder of this paper, Section 2 reviews the related work on multimodal fusion and traffic risk prediction. Section 3 details the proposed EAR-CCPM-Net architecture, including the semantic alignment modules. Section 4 presents experimental settings, results, ablation studies and interpretability analyses. Finally, Section 5 concludes the paper and outlines future research directions.

2. Literature Review

With the advancement of sensing technologies and computational capabilities, deep learning has emerged as the dominant paradigm in traffic accident prediction. These approaches leverage data-driven representations to extract complex patterns from multimodal sources such as video frames, optical flow, GPS trajectories, environmental metadata, and textual descriptions [12–14]. Convolutional Neural Networks (CNNs) have been widely adopted for visual data processing, enabling the identification of traffic participants and hazardous scenarios captured by surveillance systems [15]. In parallel, Recurrent Neural Networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks have been employed to capture temporal dependencies in driving behaviors and motion sequences [16,17]. Recent studies have also explored the use of 3D-CNNs for spatiotemporal modeling of video streams and Transformer-based architecture for capturing long-range dependencies in sequential data [4]. Despite these advancements, most deep learning models remain modality-specific, lacking cross-modal integration.

Multimodal learning has emerged as a powerful paradigm in traffic accident prediction, aiming to harness complementary signals from heterogeneous sources such as image sequences, optical flow, trajectory data, environmental cues, and textual inputs [18,19]. Common fusion strategies include early fusion at the input level, late fusion at the decision level, and hybrid fusion across intermediate network layers, with the latter demonstrating superior capability in integrating modality-specific representations into a coherent semantic space [20]. Recent advancements have introduced attention-based mechanisms, Transformer architectures, and Graph Neural Networks (GNNs) to enhance dynamic feature interaction and spatial–temporal reasoning across modalities [21]. Despite their effectiveness, many existing models still process each modality independently during perception, lacking collaborative feature alignment and cross-modal guidance. Moreover, fusion designs are often shallow, limiting their ability to exploit deep semantic complementarities.

Recent advancements in multimodal learning for traffic accident prediction have increasingly emphasized cross-modal attention mechanisms to overcome the limitations of traditional fusion strategies. While early, late, and hybrid fusion methods offer foundational integration approaches, they often lack semantic adaptability and fail to capture dynamic interdependencies between heterogeneous inputs [22,23]. Cross-modal attention mechanisms address these limitations by allowing one modality to selectively attend to another, thereby enabling semantic enrichment and context-aware reasoning [24,25]. Unidirectional attention models such as LXMERT dynamically align auxiliary modality information such as trajectory signals to guide visual risk feature enhancement [25]. More expressive

frameworks adopt bidirectional co-attention, allowing mutual refinement of features across modalities as implemented in ViLBERT [26] and related dual-stream transformers [27]. Beyond attention, several approaches integrate modality routing, entropy-based weighting, and confidence-aware fusion for robustness under degraded or incomplete modality conditions [28–30]. Notably, models such as CAP [9] incorporate cognitive priors and gaze-inspired heatmaps to enhance semantic traceability and interpretability, although their interaction depth remains limited. Multi-Hypothesis Cross-Attention [31] further expands this line of work by encouraging semantic coherence and hypothesis diversity across multiple attention layers. Collectively, these approaches highlight a paradigm shift toward more adaptive, robust, and semantically aligned multimodal prediction systems capable of capturing complex dependencies in real-world traffic contexts.

Graph-based modeling has recently emerged as a compelling paradigm for advancing cross-modal perception in traffic accident prediction, offering structured, relational representations that go beyond the limitations of feature-level attention mechanisms. Unlike conventional attention methods that struggle to encode higher-order semantic relationships, graph-based approaches construct explicit topologies where nodes represent meaningful traffic entities and edges encode spatial, temporal, or semantic dependencies across and within modalities [32–34]. These methods are especially well-suited for modeling complex, dynamic traffic scenes characterized by heterogeneous interactions among road users, vehicle trajectories, and environmental context [35]. Spatial–temporal graph architectures such as STGAT [36] provide an adaptive framework for integrating sequential patterns with spatial layouts, combining graph attention and temporal convolution to capture evolving risks across both latent and observable structures. Recent trends have further introduced cross-modal graph variants that fuse information from visual frames, trajectory streams, and scene semantics into a unified graph for collaborative reasoning [37]. Integration with attention mechanisms such as VSGNet [38] and RA-AGN [39] has shown that graph-structured inputs enable fine-grained region-aware inference, allowing attention to be guided toward risk-relevant substructures within traffic scenes. These hybrid frameworks collectively establish graph-based interaction modeling as a powerful tool for enabling semantically aligned, structurally aware, and interpretable multimodal perception systems particularly valuable for the early detection of high-risk scenarios in dynamic environments.

To summarize, current research on multimodal traffic accident prediction has made significant progress in leveraging deep learning, cross-modal attention, and graph-based reasoning. However, several limitations remain. First, most existing models adopt relatively shallow cross-modal attention mechanisms that often treat modalities in isolation or with limited interaction, leading to underutilization of semantic complementarities. Second, fusion strategies such as early or late fusion lack flexibility and fail to capture fine-grained, dynamic relationships between modalities, particularly under complex and rapidly changing traffic conditions. Third, while graph-based methods have introduced structured reasoning, they are still in the early stages of integrating multi-level semantic alignment across visual, spatial, and temporal inputs. These issues collectively point to a persistent gap in the collaborative modeling capacity and semantic coherence of current approaches. Motivated by these limitations, this study proposes a novel deep learning framework designed with robust cross-modal collaborative perception and hierarchical semantic alignment mechanisms. By enabling deeper interactions between heterogeneous traffic modalities (e.g., visual scenes, motion trajectories, and contextual cues), the framework aims to enhance early recognition of traffic accident risks through more adaptive, interpretable, and semantically consistent multimodal reasoning.

3. Research Methodology

This section presents the methodological design of the Early Accident Risk Cross-modal Collaborative Perception Network (EAR-CCPM-Net) (Figure 2), developed to address the challenges outlined in the study. Specifically, this model targets the insufficient cross-modal semantic alignment and weak collaborative perception prevalent in current multimodal traffic risk prediction systems. By integrating vision and language modalities through a novel Cross-modal Collaborative Perception Mechanism (CCPM), EAR-CCPM-Net aims to enable more accurate, robust, and interpretable early accident risk prediction in complex traffic scenarios. To achieve this, EAR-CCPM-Net is constructed with a three-part modular design that facilitates hierarchical fusion, dynamic attention-based interaction, and context-aware temporal reasoning.

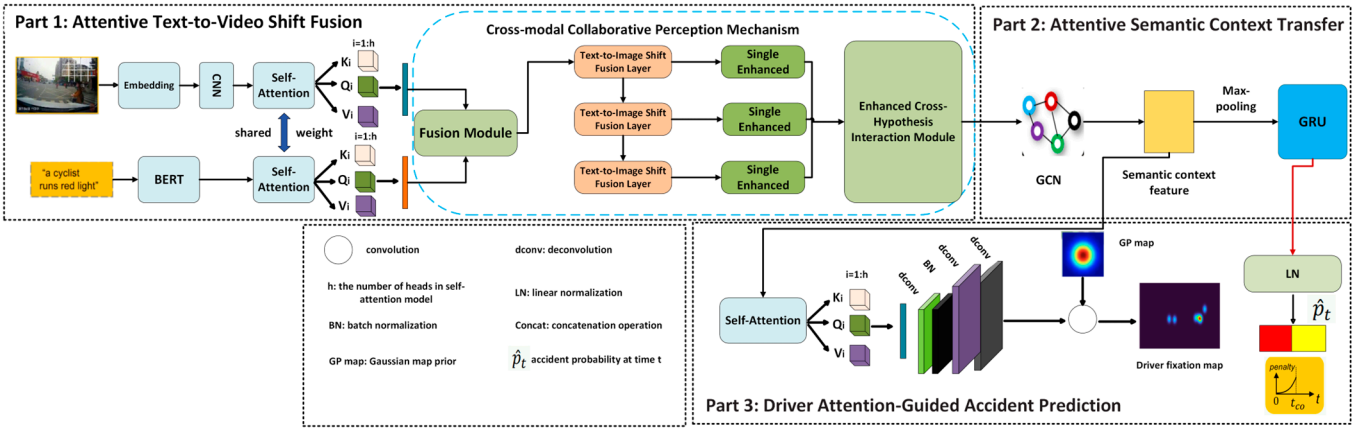

Figure 2. The architecture of early accident risk Cross-modal Collaborative Perception Mechanism Network (EAR-CCPM-Net).

Part 1: Attentive Text-to-Video Shift Fusion: This module is responsible for initial feature extraction and semantic alignment across heterogeneous input modalities, such as video frames and textual scene descriptions. Visual features are extracted using Convolutional Neural Network (CNN) followed by self-attention layers, while textual inputs are encoded using a pre-trained BERT model to capture contextual semantics. A shared-weight multi-head self-attention mechanism is employed to map both modalities into a unified semantic space. The aligned features are then progressively fused through the CCPM, which includes a Fusion unit, hierarchical Text-to-Image Shift Fusion Layers, and a dual-stage enhancement scheme (Single Enhance and Enhanced Cross-Hypothesis Interaction) to strengthen local consistency and semantic complementarity.

Part 2: Attentive Semantic Context Transfer: the semantic relationship between spatial regions is modeled through the Graph Convolutional Network (GCN), and the Gated Recurrent Unit (GRU) is used to capture the temporal evolution characteristics.

Part 3: Driver Attention-Guided Accident Prediction: With the temporal GRU, the Attentive Semantic Context Transfer module is modeled to serve the accident prediction score computation. In each frame, the driver fixation map is also reconstructed by the self-attention of the semantic context feature, which aims to fulfill a Driver Attention-Guided Accident Prediction.

Among them, the CCPM is the core innovation module of EAR-CCPM-Net, embedded in the first part. Through the text-to-image feature transfer fusion layer (Text-to-Image Shift Fusion Layer) and the enhancement module, it supports dynamic and gradual cross-modal semantic interaction and fusion, significantly improving the feature expression ability and prediction performance.

3.1. Data Collection and Preprocessing

The videos contained within CAP-DATA [9] have been sourced from a variety of accident datasets in the public domain, as well as various video stream sites, including, but not limited to YouTube, Bilibili, and Tencent. The CCD, A3D, DoTA, and DADA-2000 datasets are utilized in conjunction with additional text annotation, in conjunction with a thorough anomaly window, and with accident time stamp annotation. Furthermore, all accident videos are categorized into 58 distinct categories based on the occurrence relations of different road participants. The final collection consists of 11,727 videos with 2.19 million frames that have undergone annotation. The annotation attributes of CAP-DATA facilitate numerous valuable tasks in the field of accident analysis and research, including labeled fact–effect–reason–introspection description and temporal accident frame label. The time window with different annotations is illustrated in Figure 1.

In the context of multimodal data preprocessing, the quality of the image modality exerts a direct influence on the model's capacity to interpret the traffic scene during the feature extraction stage. This influence is particularly pronounced when the model is engaged in collaborative modeling with other semantic modalities including accident time windows, behavior labels, and participant relationships. In such scenarios, the structural integrity and edge clarity of the image modality are of paramount importance. In the context of traffic accident prediction, local texture, edge information, and contour features of traffic participants frequently serve as precursory indicators of accidents. Under data preprocessing, bilateral filtering is added to normal standard normalization and augmentation techniques applied to RGB video frames from the CAP-DATA database. The technique effectively reduces image noise while sharpening contours' edges and preserving local structure details. It thus enhances the model's semantic responsiveness to risk-related areas in subsequent collaborative fusion. The bilateral filter is an edge-preservation, non-linear denoising method with spatial closeness and pixel color similarity [40]. With the use of spatial and grayscale similarity, it effectively reduces the noise without losing edge acuteness. Unlike iterative or global filters, the bilateral filter locally filters and computationally is simple. Its greatest strength and reason come from the usual methods like Gaussian filtering in edge preservation. Although Gaussian filters are used extensively for filtering out noise, they blur high-frequency information and produce very poor edge degradation. To intuitively illustrate the visual enhancement brought by bilateral filtering, Figure 3 compares an original RGB frame with its processed counterpart. It can be observed that bilateral filtering effectively removes high-frequency noise while retaining edge contours and fine structural details such as vehicle outlines and lane boundaries. This visual improvement contributes to more accurate feature extraction in subsequent model stages.

Figure 3. Bilateral filtering enhancement on RGB video frames.

For the MINI-Train-Test, with the same setting of CAP-DATA, the study takes 512 and 168 raw video sequences for training and testing. Figure 4 shows the sampling strategy of positive samples and negative samples in the training dataset of MINI-Train. For MINI-Train, because of the limited training samples, the study randomly samples the temporal window with the end time of *t_co* (with the length of 150 frames) in raw accident videos, and the positive and negative samples are defined by checking whether the collision frame at time *t_ai* is contained in the sampled window or not. Due to the limited number of raw videos and overlapping sliding windows, many of the positive and negative samples contain partially shared frame sequences. After sampling, a total of 3312 training samples (1778 positive, 1534 negative) and 815 testing samples (608 positive, 207 negative) are generated. Consequently, the positive and negative samples in MINI-Train-Test have many overlapping frames. The sampling strategy is the same as in [9]. The total number of training and testing samples in MINI-Train-Test is shown in Table 1.

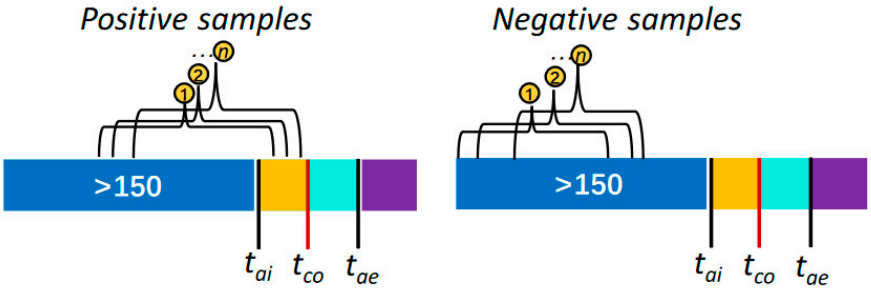

Figure 4. The sampling strategy for the positive and negative samples.

Table 1. The number of training and testing samples in MINI-Train-Test evaluation.

Eval	MINI-Train-Test			
	Raw Videos		Positive	Negative
Train	512	sampling	1778	1534
Test	168	sampling	608	207

3.2. Architecture Design and Rationale of the Proposed CCPM

To address the challenges of heterogeneous modality fusion in early accident risk prediction, we propose the EAR-CCPM-Net with a structured Cross-modal Collaborative Perception Mechanism (CCPM). These mechanism models feature alignment, enhancement, and interaction across visual and textual modalities.

Unlike traditional fusion strategies that rely on static or shallow integration, CCPM adopts a multi-stage dynamic interaction design. This design progressively refines cross-modal semantic representations to capture richer contextual information.

As illustrated in Figure 5, CCPM is organized into four core modules: (i) Fusion for initial feature embedding and semantic unification; (ii) Text-to-Image Shift Fusion Layers (T2I-SFLayers) [9] for progressive cross-modal semantic alignment; (iii) Single Enhanced Module for intra-modal feature strengthening and noise suppression; (iv) Enhanced Cross-Hypothesis Interaction module for deep cross-modal collaborative enhancement. Each module addresses a specific stage of the cross-modal integration process, ultimately contributing to a robust, adaptive, and semantically consistent multimodal representation.

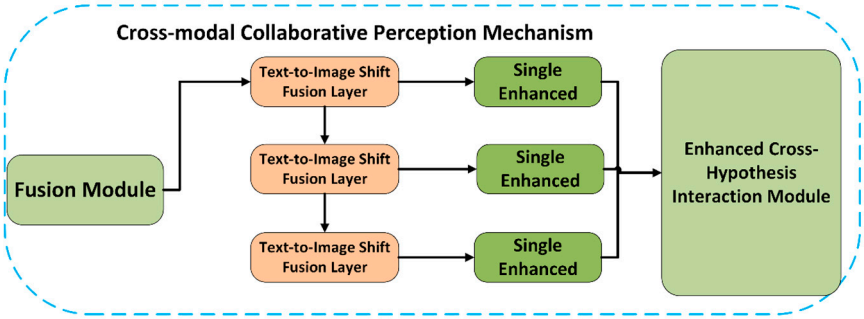

Figure 5. The architecture of Cross-modal Collaborative Perception Mechanism (CCPM).

3.2.1. Fusion Module

The Fusion Module constitutes the building block of the Cross-modal Collaborative Perception Mechanism (CCPM) that is implemented to construct a unified semantic space across heterogeneous modalities. Its key intention is to enable early interaction and alignment of visual and textual features prior to further cross-modal integration. As depicted in Figure 6, the module has a hierarchical structure with three successive subcomponents: (i) a Multi-Head Self-Attention (MHSA) layer on visual features [41], (ii) a Multi-Head Cross-modal Attention (MHCA) mechanism to facilitate inter-modal interactions [42], and (iii) a Cross-Gated Feedforward Network (CGFN) for adaptive refinement of features. All these components make incremental contributions to modality-specific representations as well as semantic convergence, thereby establishing sufficient ground for the subsequent stages of cross-modal collaborative perception.

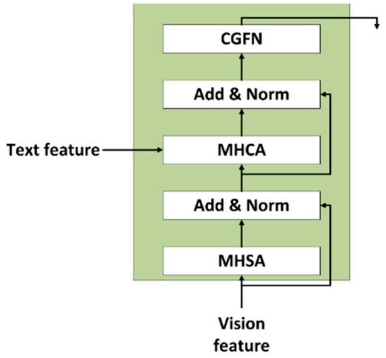

Figure 6. The design of Fusion Module.

Visual Features Encoding: The input image is first segmented into non-overlapping patches through a Patch Embedding (PE) module. The patches are subsequently passed to CNN to achieve preliminary visual features. Next, a Multi-Head Self-Attention (MHSA) mechanism is used to capture long-range dependencies so that the model can incorporate global contextual information and enhance the quality of visual representations. This process effectively preserves key spatial semantics relevant to accident detection.

Textual Features Encoding: Textual annotations or descriptions related to traffic scenes (e.g., "a cyclist runs red light") are pre-trained with BERT to extract contextual semantic vectors, and then the self-attention mechanism is connected to further enhance the global semantic modeling capability. Like visual modality processing, textual features are also refined through MHSA layers, which capture semantic interdependencies and contextual nuances present in the textual data.

Given the video frame $T_t \in R^{224\times224\times3}$ and the text description γ_t, they are encoded by the Patch Embedding and pre-trained BERT model. Patch Embedding is fulfilled by partitioning the video frame into $N \times N$ patches, and each patch is encoded

by a 2D convolution with the kernel size of 16×16 with the stride of 16. Therefore, the patch width is $N = \frac{224}{16} = 14$, and encoded video frame embedding is denoted as $M_t \in R^{14\times14\times768}(16 \times 16 \times 3 = 768)$. To reduce the computational burden, M_t^I is down sampled to $7 \times 7 \times m$ with a 1×1 convolution, where m is a hyperparameter that denotes the dimension after down sampling convolution and is set as $m = 120$ for all experiments. Text description $\gamma_{1:t}$ is encoded by a pre-trained BERT model, and the output of the text embedding in this work is denoted as $M_t^T \in R^{nT\times m}$, where n_T is also a hyperparameter which is determined by the max length of the input text and is set as 15 in this work.

After completing modality-specific embedding and encoding, the study obtains the visual embedding $M_t^I \in R^{H\times W\times D}$ of the video frame and the semantic embedding $M_t^T \in R^{L\times D}$ of the text description, where $H \times W$ is the spatial dimension of the visual feature (the number of image patches), L is the number of words of the text feature, and D is the embedding dimension. Before entering the fusion stage, the study first flattens the video embedding M_t^I into a 49×120 matrix, denoted as $X_{version} \in R^{(H\times W)\times D}$. Correspondingly, the text features are denoted as $X_{text} \in R^{L\times D}$.

The Fusion Module, as illustrated in Figure 6, is composed of three primary sub-modules: a Multi-Head Self-Attention (MHSA) mechanism applied to visual features, a Multi-Head Cross-modal Attention (MHCA) mechanism that integrates textual features with visual context, and a Cross-Gated Feedforward Network (CGFN) for feature refinement. The detailed computational process is defined as follows. Multi-Head Self-Attention (MHSA): The visual features are first passed through *MHSA* to model the global spatial dependencies between image feature blocks. The specific calculation formula is

$$MHSA(X_{version}) = Concat(head_1, \ldots, head_h)W^O \tag{1}$$

The calculation of each attention head is defined as:

$$head_i = Attention\left(X_{version}W_i^Q, X_{version}W_i^K, X_{version}W_i^V\right) \tag{2}$$

where

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where W_i^Q,W_i^K,W_i^V are the corresponding learnable parameter matrices, d_k is the dimension of each head, and h is the number of attention heads. In this study, the number of attention heads is set to 8. Consequently, the number of tokens in vision and text pathways are 49 and 15. The visual feature output after *MHSA* is recorded as

$$O_t^I = LayerNorm(X_{vision} + MHSA(X_{vision})) \tag{4}$$

Multi-Head Cross-modal Attention (MHCA): The text features are directly input into the *MHCA*, and the visual features are used as the key and value of the cross-modal attention mechanism. In this way, the interactive fusion and semantic alignment of text features and visual features are achieved. The specific calculation formula is as follows:

$$MHCA\left(X_{text}, O_t^I\right) = ConcatConcat\left(head_1^{'}, \ldots, head_h^{'}\right)W^{O^{'}} \tag{5}$$

Among them, each cross-modal attention head $head_i'$ is defined as

$$head_i^{'} = Attention\left(X_{text}W_i^{Q^{'}}, O_t^I W_i^{K^{'}}, O_t^I W_i^{V^{'}}\right) \tag{6}$$

Through this step, the spatial information contained in the visual features is integrated into the text features to achieve cross-modal semantic alignment. Then through Residual Connection and Layer Normalization,

$$O_t^{IT} = LayerNorm\left(X_{text} + MHCA\left(X_{text}, O_t^I\right)\right) \tag{7}$$

Cross-Gated Feedforward Network (CGFN): Following the cross-modal attention operation, the fused semantic features are further refined using a Cross-Gated Feedforward Network (CGFN) as shown in Figure 7, which is specifically designed to enhance salient information and suppress irrelevant or noisy features through a dual-path gating mechanism. The CGFN adopts a structured multi-step processing pipeline, consisting of linear projection, gated convolutional enhancement, and final fusion. The detailed computational process is as follows.

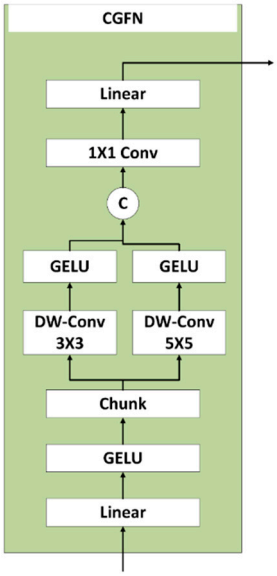

Figure 7. The design of CGFN module.

Step 1: Initial Linear Transformation and Activation: the input feature $X \in R^{N \times D}$ first undergoes a linear transformation followed by a *GELU* activation:

$$X' = GELU(W_1 X + b_1) \tag{8}$$

where $W_1 \in R^{D \times D}$ and $b_1 \in R^D$ are learnable parameters.

Step 2: Channel-wise Splitting and Depth-wise Convolutions: the activated feature X' is then split equally along the channel dimension into two sub-features $X_1', X_2' \in R^{N \times (D/2)}$:

$$X', X_1', X_2' = Split(X') \tag{9}$$

Each sub-feature is subsequently processed by a separate depth-wise convolution to capture local spatial structures at different receptive field sizes:

$$Y_1 = GELU\left(DW - Conv_{3\times3}\left(X_1'\right)\right) \tag{10}$$

$$Y_2 = GELU\left(DW - Conv_{5\times5}\left(X_2'\right)\right) \tag{11}$$

where $DW - Conv_{3\times3}$ and $DW - Conv_{5\times5}$ denote depth-wise convolutions with 3×3 and 5×5 kernels, respectively, allowing the model to adaptively capture fine-grained and broader spatial contexts.

Step 3: Feature Concatenation and Channel Mixing: the outputs Y_1 and Y_2 are concatenated along the channel dimension:

$$Y = Concat(Y_1, Y_2) \tag{12}$$

and passed through a 1×1 pointwise convolution to achieve cross-channel mixing:

$$Y' = Conv_{1\times1}(Y) \tag{13}$$

This operation fuses multi-scale spatial information while maintaining the original token-wise structure.

Step 4: Final Linear Projection and Residual Connection: finally, a second linear projection is applied to the fused feature, followed by a residual connection with the original input X, along with Layer Normalization:

$$O = LayerNorm\left(X + W_2Y' + b_2\right) \tag{14}$$

where $W_2 \in R^{D\times D}$ and $b_2 \in R^D$ are learnable parameters.

The output $O \in R^{N\times D}$ serves as the refined feature representation, selectively preserving critical semantic cues while suppressing noise and redundant information, thereby significantly improving the expressiveness and robustness of the fused features.

3.2.2. Single Enhanced Module

The Single Enhanced Module is designed to further strengthen intra-modality feature representations and stabilize the dynamic fusion process across stages. Following each T2I-SFLayer, a dedicated Single Enhanced Module is applied to improve internal feature consistency and suppress redundant noise.

As shown in Figure 8, the Single Enhanced Module comprises three primary operations: Layer Normalization (LN), Multi-Head Self-Attention (MHSA), and a Cross-Gated Feedforward Network (CGFN). Through this structured processing pipeline, Single Enhanced Module progressively refines the semantic coherence and expressive power of the shifted features.

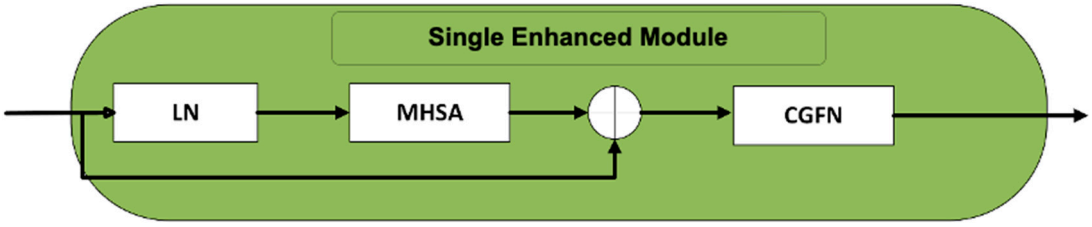

Figure 8. The design of Single Enhanced Module.

Input Normalization: given the shifted fusion features from the previous stage, denoted as $X_{shift}^{(l+1)}$, the Single Enhanced Module first applies *Layer Normalization (LN)* to stabilize the feature distributions across different modalities:

$$X_{norm}^{(l+1)} = \mathrm{LN}\left(X_{shift}^{(l+1)}\right) \tag{15}$$

where the *LN* operation is defined as

$$LN(X) = \gamma\frac{X-\mu}{\sigma+\epsilon} + \beta \tag{16}$$

Here, μ and σ represent the mean and standard designation computed along the feature dimension, ϵ is a small constant for numerical stability, and γ and β are learnable affine transformation parameters.

Multi-Head Self-Attention (MHSA): The normalized features $X_{norm}^{(l+1)}$ are then processed by a Multi-Head Self-Attention (MHSA) mechanism to enhance internal dependencies and contextual relationships among feature tokens. The *MHSA* operation is given by

$$MHSA\left(X_{norm}^{(l+1)}\right) = Concat(head_1, \ldots, head_h)W^O \tag{17}$$

where each attention head is computed as

$$head_i = Attention\left(X_{norm}^{(l+1)}W_i^Q, X_{norm}^{(l+1)}W_i^K, X_{norm}^{(l+1)}W_i^V\right) \tag{18}$$

where W_i^Q, W_i^K, W_i^V are the query, key, and value projection matrices, d_k is the dimension of each head, and h is the number of attention heads.

After *MHSA*, a residual connection is applied to maintain the stability of feature representation:

$$X_{MHSA}^{(l+1)} = X_{shift}^{(l+1)} + MHSA\left(X_{norm}^{(l+1)}\right) \tag{19}$$

Cross-Gated Feedforward Network (CGFN): To further refine the feature representations, the output $X_{MHSA}^{(l+1)}$ is passed through a Cross-Gated Feedforward Network (CGFN). *CGFN* selectively amplifies informative features and suppresses irrelevant or noisy components through a multi-branch gating mechanism.

The *CGFN* operation is defined as

$$X_{CGFN}^{(l+1)} = CGFN\left(X_{MHSA}^{(l+1)}\right) \tag{20}$$

where the *CGFN* structure includes Figure 7. Finally, the output of the Single Enhanced Module is

$$X_{enhance}^{(l+1)} = X_{CGFN}^{(l+1)} \tag{21}$$

Through the combination of normalization, self-attention, and dynamic feature gating, the Single Enhanced Module effectively reinforces intra-modal feature structures, enhances semantic consistency, and suppresses noise accumulation across fusion stages. This progressive refinement is critical for maintaining robust and expressive feature representations throughout the multi-stage CCPM architecture.

3.2.3. Enhanced Cross-Hypothesis Interaction Module

The Enhanced Cross-Hypothesis Interaction Module serves as a crucial component in the CCPM, aiming to further refine the semantic consistency and multi-level feature expressiveness across different abstraction layers within the multimodal fusion process. By modeling intra-level and cross-level feature interactions, the Enhanced Cross-Hypothesis Interaction Module strengthens the dynamic alignment between visual and textual modalities, thus contributing directly to the research objective regarding CCPM for early accident risk prediction.

As illustrated in Figure 9, the Enhanced Cross-Hypothesis Interaction Module comprises three sequential stages: (i) intra-level semantic refinement using Layer Normalization (LN) and Multi-Head Self-Attention (MHSA); (ii) cross-level collaborative enhancement using Multi-Head Cross Attention (MCA); (iii) final feature consolidation through Layer Normalization and Cross-Gated Feedforward Network (CGFN) enhancement.

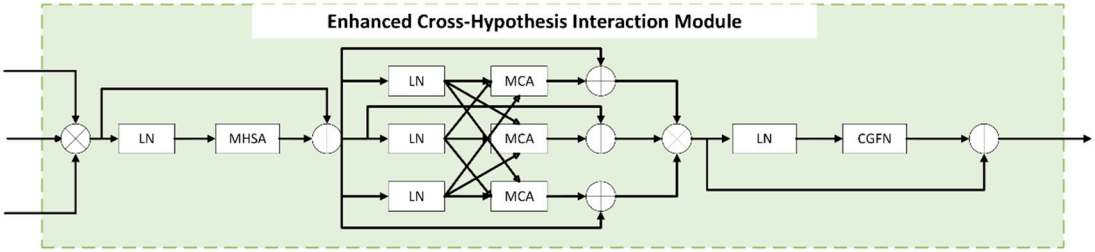

Figure 9. The design of Enhanced Cross-Hypothesis Interaction Module.

Intra-Level Self-Enhancement: Given the fused semantic representations from different abstraction levels, denoted as $Z^{(1)}, Z^{(2)}, Z^{(3)} \in R^{T\times D}$, where T is the token sequence length and D is the feature dimension, the module first applies intra-level refinement independently for each level.

Each input $Z^{(l)}$ is normalized and processed by a Multi-Head Self-Attention mechanism to capture internal contextual relationships:

$$\hat{Z}^{(l)} = LN\left(Z^{(l)}\right) \tag{22}$$

The *MHSA* operation for each level is defined as

$$MHSA\left(\hat{Z}^{(l)}\right) = Concat(head_1, \ldots, head_h)W^O \tag{23}$$

where each attention head is computed by

$$head_i = Attention\left(\hat{Z}^{(l)}W_Q^{(l)}, \hat{Z}^{(l)}W_K^{(l)}, \hat{Z}^{(l)}W_V^{(l)}\right) \tag{24}$$

and the attention function is

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{25}$$

The self-attention output is added back to the original input using residual connection:

$$Z_{MHSA}^{(l)} = Z^{(l)} + MHSA\left(\hat{Z}^{(l)}\right) \tag{26}$$

This intra-level enhancement enables each semantic abstraction layer to independently refine and stabilize its internal feature distributions.

Cross-Level Collaborative Enhancement: to capture complementary semantic cues across different abstraction levels, the Enhanced Cross-Hypothesis Interaction Module introduces a Cross-Level Multi-Head Cross Attention (MCA) mechanism.

For each level l, the query is projected from its own self-enhanced feature:

$$Q^{(l)} = Z_{MHSA}^{(l)}W_Q^{(l)} \tag{27}$$

while the keys and values are concatenated from the other levels:

$$K^{(-l)} = Concat_{l'\neq l}\left(Z_{MHSA}^{(l')}W_K^{(l')}\right) \tag{28}$$

$$V^{(-l)} = Concat_{l'\neq l}\left(Z_{MHSA}^{(l')}W_V^{(l')}\right) \tag{29}$$

The cross-level attention is then computed as

$$MCA^{(l)} = Softmax\left(\frac{Q^{(l)}\left(K^{(-l)}\right)^{T}}{\sqrt{d}}\right)V^{(-l)} \tag{30}$$

Following the *MCA* operation, a residual connection integrates the attended cross-level features:

$$\hat{Z}^{(l)} = Z^{(l)}_{mhsa} + MCA^{(l)} \tag{31}$$

This step allows each semantic level to incorporate supportive and complementary information from the other abstraction layers, enhancing multi-level semantic coherence and representation diversity.

Feature Consolidation and Output: After the intra- and cross-level enhancements, the outputs from all semantic levels are concatenated:

$$Z_{concat} = Concat\left(\tilde{Z}^{(1)}, \tilde{Z}^{(2)}, \tilde{Z}^{(3)}\right) \tag{32}$$

The concatenated representation undergoes *LN* to stabilize feature distributions:

$$Z_{norm} = LN(Z_{Concat}) \tag{33}$$

Subsequently, the normalized feature is passed through a *CGFN* for dynamic feature selection and local context enhancement:

$$Z_{CGFN} = CGFN(Z_{norm}) \tag{34}$$

Finally, a residual connection combines the *CGFN* output with the normalized feature to produce the final output of the Enhanced Cross-Hypothesis Interaction Module:

$$Z_{corssenhance} = Z_{norm} + Z_{CGFN} \tag{35}$$

The resulting robust, semantically consistent multimodal representation is subsequently fed into the Graph Convolutional Network (GCN) module for higher-order structural reasoning and final risk prediction tasks.

Through progressive intra-level self-enhancement and cross-level collaborative feature refinement, the Enhanced Cross-Hypothesis Interaction Module significantly improves the dynamic alignment and semantic coherence of fused multimodal features across different abstraction layers.

3.3. Experiment Setup

To verify the effect of the proposed CCPM on the early risk prediction ability, this section explains the specific settings of this experiment from three aspects.

3.3.1. Input Modality and Preprocessing

To simulate multimodal collaborative scenarios, the model input includes the following two modalities:

Video frame: extracted from the video at 5 fps used to capture the appearance features of the scene with an input resolution of 224 × 224.

Video-level factual text description: generated from driving video descriptions, mainly including lane status, weather, traffic rules prompt, etc., obtained through BERT encoding.

All modal inputs were normalized and dimensionally aligned before training to ensure that cross-modal features can be fused.

3.3.2. Training and Testing Settings

In the preliminary stage of model training, we independently trained the EAR-CCPM-Net that contains CCPM to validate its fusion capacity and semantic alignment capabilities under supervised settings. The training and testing were conducted on a server equipped with two NVIDIA Tesla V100 32 GB GPUs, using the Adam optimizer, an initial learning rate of 0.0001, and a batch size of 16. The model was trained for 100 epochs, and the weighted binary cross-entropy was selected as the loss function. To prevent overfitting and enhance generalization, the study applied an early stopping strategy, where training was monitored by using AP, AUC, mTTA, and $TTA_{0.5}$ performance on the validation set. The learning rate scheduling was performed dynamically to ensure stable convergence.

3.3.3. Performance Measurement

To comprehensively evaluate the effectiveness of the proposed framework in multi-modal traffic accident prediction, this study establishes a structured and objective evaluation index system aligned with the three core research objectives. To evaluate the performance of the proposed framework in multimodal traffic accident prediction, this study utilizes Average Precision (AP), Time-To-Accident (TTA), which includes the variants of $TTA_{0.5}$ and Mean Time-To-Accident (mTTA), and Area Under the ROC Curve (AUC).

AP: AP serves as a measure to evaluate the model's ability to accurately detect the occurrence of traffic accidents within videos, especially in scenarios where there is an imbalance between positive and negative samples. In binary classification tasks, assuming that TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively, we can calculate the model's recall $R = \frac{TP}{TP+FN}$ and precision $P = \frac{TP}{TP+FP}$. Recall is the fraction of actual positive instances correctly predicted by the model, and precision is the fraction of positive predictions which are actually positive. Precision–recall curve is derived from these two values, and *Average Precision (AP)* is the area under this curve which quantifies the overall precision–recall performance of the model. In practice, the area under the curve is approximated using discrete summation:

$$AP = \int P(R)dR = \sum_{k=0}^{m} P(k)\Delta R(k) \tag{36}$$

TTA: *Time-To-Accident (TTA)* quantifies the model's ability to accurately predict an impending accident in advance, reflecting its effectiveness in early risk perception.

$$TTA = t_{accident} - t_{predict} \tag{37}$$

Among them, $t_{accident}$ is the time point when the accident occurred, $t_{predict}$ is the time point when the model first predicted an accident.

mTTA: *Mean Time-To-Accident (mTTA)* represents the average TTA across varying prediction thresholds, providing a comprehensive measure of the model's early warning capability under different operational conditions.

$$mTTA = \frac{1}{N}\sum_{i=1}^{N} TTA_i \tag{38}$$

$TTA_{0.5}$: $TTA0.5$ is calculated by setting the accident score threshold at 0.5 for each video, providing a measure of how early a positive prediction is made. $mTTA$ represents the mean TTA, obtained by averaging the TTA values across a range of accident score

thresholds from 0 to 1, thereby reflecting the model's overall early warning performance under varying sensitivity settings.

$$TTA_{0.5} = t_{accident} - t_{predict}^{(score>\theta_k)} \tag{39}$$

AUC: This curve measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). The *AUC* represents a probability measure for the classifier's ability to distinguish between the positive and negative classes. The higher the *AUC*, the better the model is at predicting 0 s as 0 s and 1 s as 1 s. An *AUC* of 0.5 suggests no discrimination ability (equivalent to random guessing), and an *AUC* of 1.0 represents a perfect model that makes no mistakes in classification. *AUC* is a metric for evaluating the detection or classification, which is adopted for checking the prediction performance with the output of the occurrence probability of future accidents.

$$AUC = \int_0^1 TPR\Big(FPR^{-1}(x)\Big)dx \tag{40}$$

To evaluate the performance of the proposed framework in multimodal traffic accident prediction, this study utilizes KL divergence (KLdiv), Correlation Coefficient (CC), Similarity Metric (SIM), and Center-biased Area under ROC Curve (s-AUC). For the classification of traffic accident prediction, this study utilizes ACC (accuracy), REC (recall), PRE (precision), and F1 (F1-score).

KLdiv [43]: *KL* divergence measures the degree of difference between two probability distributions. In this study, *KLdiv* is used to evaluate the difference between the distribution of attention heatmaps generated by the model and the distribution of human attention regions.

$$KL\ (P||Q) = \sum_i P(i) log(\frac{P(i)}{Q(i)}) \tag{41}$$

where $P(i)$ represents the probability distribution of human attention areas, and $Q(i)$ represents the probability distribution of attention heatmaps generated by the model.

CC [44]: The correlation coefficient measures the linear correlation between two variables. In this study, *CC* is used to evaluate the linear correlation between the attention heatmap generated by the model and the human attention area.

$$CC = \frac{\sum_i \left(P(i) - \overline{P}\right)\left(Q(i) - \overline{Q}\right)}{\sqrt{\sum_i \left(P(i) - \overline{P}\right)^2}\sqrt{\sum_i \left(Q(i) - \overline{Q}\right)^2}} \tag{42}$$

where $\overline{P}$ and $\overline{Q}$ represent the average values of human attention regions and model attention heatmaps, respectively.

SIM [45]: *SIM* measures the degree of similarity between two distributions. In this study, *SIM* is used to evaluate the overlap between the attention heatmaps generated by the model and the human attention regions.

$$SIM = \sum_i min(P(i),\ Q(i)) \tag{43}$$

where $P(i)$ and $Q(i)$ represent the probability distribution of human attention area and model attention heatmap, respectively.

s-AUC [46]: *s-AUC* measures the discrimination ability between the attention heatmaps generated by the model and the human attention regions while controlling for the effect of center bias.

$$s - AUC = AUC_{model} - :: contentReference[oaicite\ :64]index = 64 \tag{44}$$

4. Results and Discussion

To verify the effectiveness of the proposed CCPM within EAR-CCPM-Net in enhancing early accident risk perception capabilities, this section presents detailed quantitative results from two major sets of comparative experiments: (i) Training Results; (ii) Baseline model comparisons; (iii) EAR-CCPM-Net ablation and structural variant evaluations; (iv) EAR-CCPM-Net performance comparison.

4.1. Training Results

The comprehensive training curve analysis of EAR-CCPM-Net demonstrates its strong convergence behavior, semantic alignment ability, and multimodal robustness. The training loss drops rapidly within the first 10 epochs and gradually converges to 0.26 by epoch 100, while the validation loss decreases from near 1.0 to below 0.15, indicating both effective feature learning and strong generalization on unseen data as shown in Figure 10. In terms of classification performance, the precision increases steadily to reach approximately 0.83, and the recall rises to 0.82, suggesting that the model effectively reduces false positives and captures most accident-relevant samples as shown in Figure 11. Moreover, the object detection evaluation metrics further support the model's interpretability and localization capabilities: the mAP@0.50 score climbs to 0.88, reflecting high sensitivity in coarse-level accident region detection, while the mAP@0.95 steadily improves to 0.62, confirming that the network is also capable of achieving accurate fine-grained spatial localization as shown in Figure 12. These quantitative results collectively validate that the proposed EAR-CCPM-Net achieves a strong balance between early risk detection accuracy, multimodal semantic consistency, and spatial interpretability, supported by its hierarchical attention fusion and cross-modal collaborative perception design.

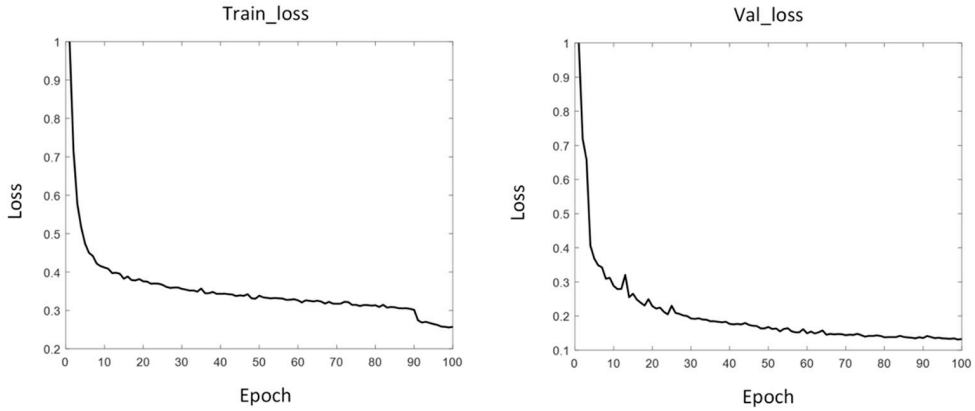

Figure 10. Training and validation loss curves of EAR-CCPM-Net.

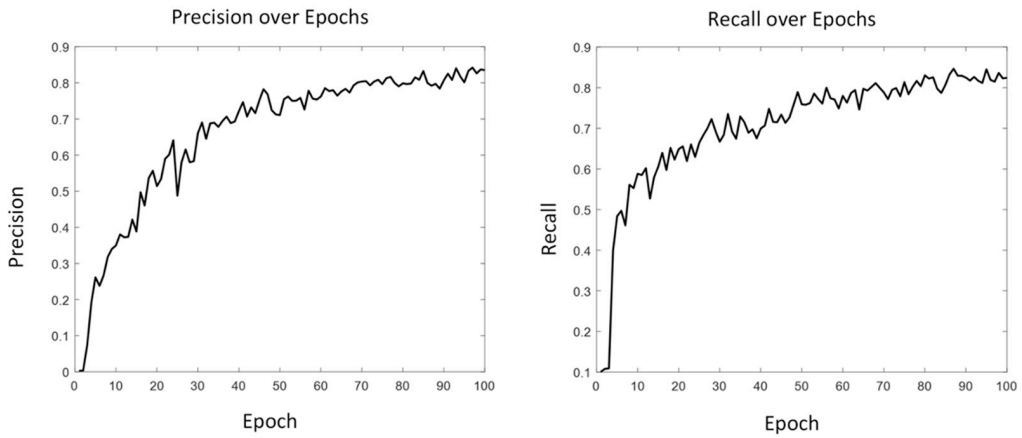

Figure 11. Precision and recall curves of EAR-CCPM-Net.

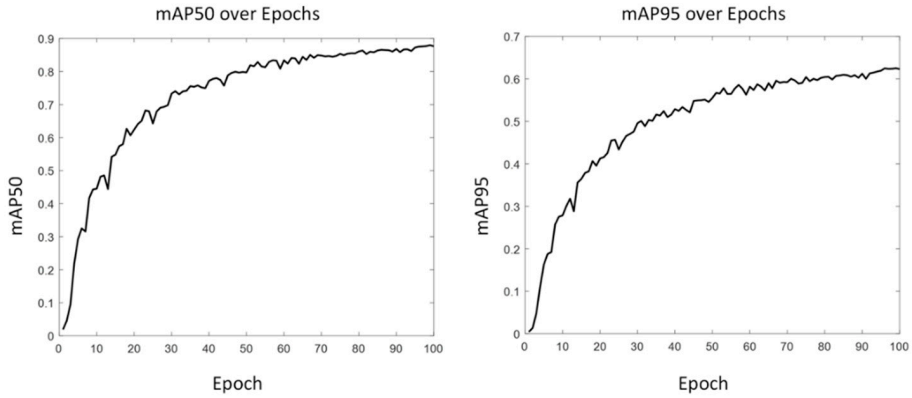

Figure 12. mAP50 and mAP95 curves of EAR-CCPM-Net.

4.2. Baseline Performance Comparison

In terms of quantitative performance evaluation, this paper selects several representative benchmark models in the current field of traffic accident prediction for comparative experiments, including traditional sequence modeling methods (such as DSA-RNN), time series adaptive learning strategies (such as AdaLEA, UncertaintyTA), visual attention-driven models (such as DRIVE), and current mainstream multimodal fusion frameworks (such as CAP). All comparisons are conducted on the MINI-Test test partition, using a unified evaluation index system, including the area under the curve (AUC), average precision (AP), the time point when the risk confidence level of 0.5 is first reached ($TTA_{0.5}$), and the average prediction lead time (mTTA), to comprehensively measure the performance of the model in terms of accuracy, stability, and foresight.

As can be seen from Table 2, the framework has achieved the best results in all four indicators. In terms of classification indicators, its AUC value reached 0.85 and AP was 0.75, both exceeding the current best-performing CAP model (0.81/0.73) and DRIVE model (0.69/0.72), demonstrating stronger classification capabilities and risk identification accuracy. In terms of time series prediction, the framework's $TTA_{0.5}$ is 4.225 s and mTTA is 4.672 s, significantly better than DRIVE (3.657/4.295) and CAP (3.927/4.209), indicating that the framework can capture potential high-risk events earlier and has better response speed and stability in time series warning. This performance improvement can be attributed to the synergy between its multi-layer attention fusion strategy and robust time series modeling modules.

Table 2. Performance comparison of traffic accident prediction of our framework.

Baselines	AUC ↑	AP ↑	$TTA_{0.5}$ (s) ↑	mTTA (s) ↑
DSA-RNN [47]	0.47	-	3.095	-
AdaLEA [48]	0.55	-	3.890	-
UncertaintyTA [49]	0.60	-	3.849	-
DRIVE [50]	0.69	0.72	3.657	4.295
CAP [9]	0.81	0.73	3.927	4.209
EAR-CCPM-Net (Proposed)	0.85	0.75	4.225	4.672

Note: arrows (↑) indicate that higher values are better.

4.3. Ablation Variants of EAR-CCPM-Net

To explore deeper the structural roles played by the proposed EAR-CCPM, this section conducts a series of ablation tests in a progressive way, ablating one large component after another and comparing to the entire EAR-CCPM-Net. To be tested are the original Fusion Module, the Text-to-Image Shift Fusion Layers (T2I-SFLayers), the Single Enhanced Module,

and the Enhanced Cross-Hypothesis Interaction Module. Quantitative results are listed in Table 3. The baseline performance of the overall EAR-CCPM-Net achieves the optimal performance in all the metrics with an AUC of 0.853, AP of 0.758, $TTA_{0.5}$ of 4.225 s, and mTTA of 4.672 s, demonstrating strong early prediction capability and semantic stability.

Table 3. Ablation and variants of EAR-CCPM-Net.

Model Name	Fusion Module	T2I-SFLayers	Single Enhanced Module	Enhanced CHIM	AUC ↑	AP ↑	$TTA_{0.5}$ (s) ↑	mTTA (s) ↑
EAR-CCPM-Net w/o Fusion	×	√	√	√	0.813	0.734	3.927	4.209
EAR-CCPM-Net w/o T2I-SFLayer	√	×	√	√	0.826	0.740	3.997	4.371
EAR-CCPM-Net w/o Single Enhanced	√	√	×	√	0.837	0.749	4.089	4.482
EAR-CCPM-Net w/o Enhanced CHIM	√	√	√	×	0.842	0.752	4.121	4.536
EAR-CCPM-Net (Full)	√	√	√	√	0.853	0.758	4.225	4.672

Note: arrows (↑) indicate that higher values are better.

When the Fusion Module is removed, the performance drops most significantly among all ablation variants, with AUC decreasing to 0.813 and AP to 0.734, while $TTA_{0.5}$ reduces to 3.927 s. This highlights the critical role of early semantic alignment across modalities. Without fusion, the model fails to establish a shared feature space, resulting in misaligned representations and weakened cross-modal attention effectiveness.

Removing the T2I-SFLayers also leads to notable degradation, with AUC falling to 0.826 and AP to 0.740. The absence of progressive cross-modal interaction across layers prevents effective hierarchical semantic blending, causing a delay in early risk signal recognition ($TTA_{0.5}$ = 3.997 s). This demonstrates that cross-modal interactions must be staged and recurrent, not single-step.

The performance gain owes to four excellent aspects of CCPM: (i) the first-ever Fusion Module, with the very first modality alignment; (ii) progressive shift fusion in T2I-SFLayers that allows deep semantic fusion at stages; (iii) hierarchical feature augmentation by Single Enhanced Module and Enhanced Cross-Hypothesis Interaction Module that contributes intra- as well as inter-level semantics richness; and (iv) using Cross-Gated Feedforward Networks (CGFNs), selectively amplifying informative features and eliminating redundancy.

To further determine the optimal configuration of cross-modal fusion depth, we extend the ablation study by incrementally varying the number of T2I-SFLayers from 1 to 4. As reported in Table 4, the model performance improves progressively from one to three layers. Specifically, a single-layer setting yields an AUC of 0.833 and AP of 0.743, while the two-layer configuration achieves 0.842 and 0.748, indicating that deeper semantic interaction facilitates stronger alignment across modalities. Notably, the three-layer configuration outperforms all others, reaching the highest AUC (0.853), AP (0.758), and earliest detection time ($TTA_{0.5}$ = 4.225 s), establishing it as the optimal choice. However, when a fourth T2I-SFLayer is added, performance slightly declines (AUC drops to 0.839, AP to 0.745, $TTA_{0.5}$ = 4.087 s), suggesting diminishing returns and potential overfitting due to over-stacking. The non-monotonic trend confirms that while multi-stage interaction is essential, excessively deep fusion introduces redundancy and noise propagation that may interfere with early semantic clarity. These results empirically validate our architectural decision to employ three progressive T2I-SFLayers as a balanced configuration deep

enough to enable hierarchical cross-modal alignment yet compact enough to maintain generalization and efficiency.

Table 4. T2I-SFLayer number ablation study.

Variant Name	Number of T2I-SFLayer	AUC ↑	AP ↑	$TTA_{0.5}$ (s) ↑	mTTA (s) ↑
EAR-CCPM-Net w/o T2I-SFLayer	0	0.826	0.740	3.997	4.371
EAR-CCPM-Net (1 T2I-SFLayer)	1	0.833	0.743	4.032	4.412
EAR-CCPM-Net (2 T2I-SFLayer)	2	0.842	0.748	4.107	4.545
EAR-CCPM-Net (3 T2I-SFLayer)	3	0.853	0.758	4.225	4.672
EAR-CCPM-Net (4 T2I-SFLayer)	4	0.839	0.745	4.087	4.491

Note: arrows (↑) indicate that higher values are better.

In the variant without Single Enhanced the model still retains T2I-SFLayers but lacks local feature refinement at each stage. This results in an AUC of 0.837 and AP of 0.749, indicating a measurable loss in performance. The mTTA also drops by nearly 0.2 s. The results confirm that Single Enhanced Modules help stabilize intermediate representations, enabling better feature propagation across layers.

To further investigate the optimal configuration of the Single Enhanced Module number, the study conducted layer-wise ablation by varying the number of Single Enhanced Modules from 1 to 4. As shown in Table 5, model performance steadily improves from 1 to 3 enhancement stages. With one single enhance layer, the model achieves an AUC of 0.846 and AP of 0.752, while the two-layer setting increases to 0.848 and 0.754, respectively, suggesting that deeper intra-modal refinement contributes to more stable and expressive intermediate representations. The configuration with three Single Enhanced Modules reaches the peak performance across all metrics: AUC (0.853), AP (0.758), and $TTA_{0.5}$ (4.225 s), validating it as the most effective design. The additional fourth layer shows no further gain in AP (0.756) and results in slightly reduced AUC (0.851) and earlier $TTA_{0.5}$ (4.122 s), indicating marginal overfitting and oversmoothing of representations due to excessive self-attention refinement. These findings confirm that while local enhancement is necessary for robust temporal feature propagation and semantic stabilization, an overly deep enhancement structure may degrade the diversity of feature responses. Therefore, the three-layer configuration provides an optimal balance between expressive power and generalization capacity in multi-stage fusion architectures such as EAR-CCPM-Net. The Enhanced Cross-Hypothesis Interaction Module shows the least performance drop when removed, yet the impact is still evident: AUC drops to 0.842 and AP to 0.752, while early detection ($TTA_{0.5}$ = 4.121 s) is marginally reduced. This suggests that although each stage produces aligned features, the absence of semantic reinforcement across levels weakens the model's global consistency and robustness, particularly in ambiguous or noisy traffic scenarios.

Empirical results confirm that CCPM significantly improves the model's temporal sensitivity and semantic coherence compared to conventional approaches such as CAP, which lack progressive multimodal enhancement. EAR-CCPM-Net consistently outperforms the baseline across AUC, AP, and early detection metrics ($TTA_{0.5}$, mTTA), demonstrating that cross-modal alignment enables the model to detect early risk cues otherwise inaccessible to late- or uni-modal fusion strategies.

Further ablation experiments highlight the indispensable role of each CCPM subcomponent. The Fusion Module provides foundational semantic alignment, while the T2I-SFLayers and Single Enhanced Module enable progressive feature refinement. The Enhanced Cross-Hypothesis Interaction Module reinforces inter-level consistency. The removal of any single component causes notable degradation, with early fusion proving most critical. Depth-wise analysis reveals that a three-layer configuration of T2I-SFLayers and the Single Enhanced Module achieves an optimal balance between representational number and generalization.

Table 5. Single Enhanced Module number ablation study.

Variant Name	Number of T2I-SFLayer	AUC ↑	AP ↑	$TTA_{0.5}$ (s) ↑	mTTA (s) ↑
EAR-CCPM-Net w/o Single Enhanced	0	0.837	0.749	4.089	4.482
EAR-CCPM-Net (1 Single Enhanced)	1	0.846	0.752	4.050	4.500
EAR-CCPM-Net (2 Single Enhanced)	2	0.848	0.754	4.090	4.540
EAR-CCPM-Net (3 Single Enhanced)	3	0.853	0.758	4.225	4.672
EAR-CCPM-Net (4 Single Enhanced)	4	0.851	0.756	4.122	4.635

Note: arrows (↑) indicate that higher values are better.

4.4. EAR-CCPM-Net Performance Comparison

To evaluate the interpretability of the multi-layer attention mechanism in EAR-CCPM-Net in multimodal semantic perception, this section introduces the mainstream attention map alignment evaluation index to quantitatively analyze the attention distribution of the model in the risk prediction process. This analysis is not intended to evaluate the performance of the model in terms of prediction accuracy but focuses on whether its internal attention maps are semantically consistent with the human cognitive path, thereby verifying its interpretability and cognitive transparency.

Although EAR-CCPM-Net is not designed for the task of "driver gaze prediction", the gaze heatmap of real drivers in traffic scenes is widely regarded as a "reference standard for cognitive rationality" in interpretability research. Recent studies [9,50] have shown that if the attention heatmap generated by the deep model in the accident prediction process is closer to the driver's gaze area in space, its perception path and reasoning method are closer to the human cognitive process, thus it has higher interpretability. Therefore, this section uses the real driver's gaze map as an evaluation benchmark and uses four common saliency alignment indicators to compare the attention alignment performance of the framework with multiple existing methods.

Table 6 shows the degree of alignment between the attention maps generated by different models and the real driver's gaze area under the MINI-Test data partition. The four indicators used are Kullback–Leibler divergence (KLdiv, the lower the value, the better), Pearson correlation coefficient (CC, the higher the better), heatmap distribution similarity index (SIM, the higher the better), and shuffled-AUC (s-AUC, the higher it is, the closer to the real gaze point). These indicators together reflect the spatial focusing accuracy and cognitive consistency of the model's attention area in risk scenarios. Among all evaluated models, the framework achieves the best interpretability scores across most metrics, with the lowest KLdiv (1.72) and the highest s-AUC (0.87), outperforming both early attention models (e.g., BDDA: KLdiv = 3.32, s-AUC = 0.63) and more recent multimodal architectures

(e.g., CAP: KLdiv = 1.89, s-AUC = 0.85). The improvements in s-AUC and CC indicate that the framework's attention heatmaps exhibit stronger spatial consistency with ground-truth driver gaze regions. These gains can be attributed to the model's multi-level attention design, which enables progressive semantic refinement and modal interaction, resulting in attention distributions that are not only spatially precise but also cognitively aligned with human visual reasoning under driving conditions.

Table 6. The HASI Module performance comparison.

Baselines	KLdiv ↓	CC ↑	SIM ↑	s-AUC ↑
BDDA [51]	3.32	0.33	0.25	0.63
DR (eye)VE [52]	2.27	0.45	0.32	0.64
ACLNet [53]	2.51	0.35	0.35	0.64
DADA [54]	2.19	0.50	0.37	0.66
DRIVE [50]	2.65	0.33	0.19	0.66
CAP [9]	1.89	0.38	0.28	0.85
(EAR-CCPM-Net) Proposed	1.72	0.40	0.29	0.87

Note: arrows (↑) indicate that higher values are better; arrow (↓) indicate that lower values are better.

To further substantiate the quantitative findings presented in Table 6, Figure 13 provides a visual comparison of attention heatmaps across three models: CAP, framework, and the human gaze ground-truth. The comparison spans two representative frames (#085 and #145) from a real-world urban driving scenario. Each column, respectively, illustrates the human attention distribution, CAP's model output, and framework 's attention map. In frame #085, both the ground-truth and framework clearly focus on the leading vehicle, which is a potential source of risk due to its proximity and possible intention to decelerate or change lanes. Compared to CAP, whose heatmap covers a broader but less precise region including irrelevant parts of the road, the HASI framework maintains a focused and centered attention distribution directly on the high-risk object. In frame #145, this pattern continues: while the human gaze remains tightly centered on the vehicle directly ahead and its motion trajectory, the HASI framework 's attention heatmap shows a similarly compact and aligned focus. In contrast, CAP still demonstrates spatial divergence, with attention slightly drifting to less critical areas such as the sidewalk or background vehicles.

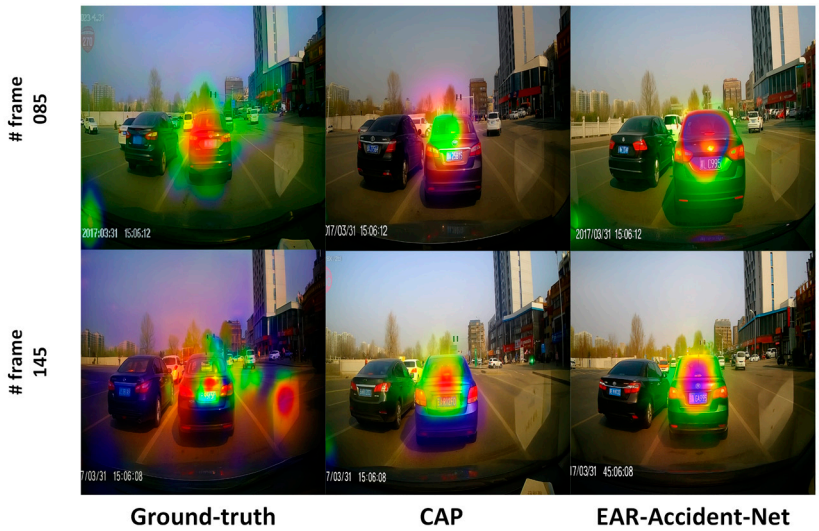

Figure 13. Visual comparison of attention heatmaps across ground-truth, CAP, and EAR-Accident-Net on representative risk frames.

These visual comparisons reinforce the saliency alignment metrics previously reported, where the framework achieved the highest s-AUC (0.87) and lowest KLdiv (1.72). The attention maps generated by the framework not only achieve finer spatial resolution but also follow human-like cognitive trajectories in risk anticipation. This is largely attributed to its hierarchical attention mechanism, which gradually refines spatial semantics from coarse contextual awareness to entity-specific risk focus across visual and temporal layers.

5. Conclusions

This study successfully addresses research questions by proposing the EAR-CCPM-Net, a novel early accident risk prediction framework built upon the Cross-modal Collaborative Perception Mechanism (CCPM). The CCPM integrates a multi-stage design that includes Text-to-Image Shift Fusion Layers (T2I-SFLayers), the Single Enhanced Module, and the Enhanced Cross-Hypothesis Interaction Module, allowing for progressive semantic alignment and mutual reinforcement across visual, motion, and textual modalities. Quantitative evaluations demonstrate the superior predictive performance of the model: EAR-CCPM-Net achieves an AUC of 0.85 and AP of 0.75, representing respective gains of +4% and +2% over the CAP baseline. Moreover, the model extends the average Time-to-Accident ($TTA_{0.5}$) from 3.831 s to 4.225 s, confirming its enhanced capability for early warning. Ablation studies further validate the contribution of each CCPM subcomponent to semantic coherence and temporal awareness. These results collectively affirm that EAR-CCPM-Net not only improves early-stage accident cue capture but also achieves robust cross-modal semantic alignment, thereby fulfilling the design goals. While EAR-CCPM-Net exhibits notable advances in multimodal fusion and early risk forecasting, several limitations warrant further exploration. The current fusion paradigm primarily operates under static cross-modal alignments, lacking sensitivity to the dynamic relevance of modalities across varying traffic contexts.

Despite these contributions, several limitations remain. First, the current fusion strategy operates under static cross-modal alignment and lacks context-adaptive mechanisms to dynamically modulate the relevance of different modalities. Future work should explore context-aware fusion architectures capable of adaptively weighting modality contributions based on evolving traffic scene complexity and uncertainty. Second, the semantic textual modality used in this study is derived from manually annotated descriptions, which limit the model's deployability in real-world, real-time systems. To mitigate this dependency, future extensions will consider replacing textual input with automatically extractable features, such as scene captions generated by vision–language models, object trajectory graphs, or metadata from smart traffic infrastructure. Third, the adopted MINI-Train-Test sampling strategy—consistent with prior work [9]—involves sliding window sampling within the same set of accident videos, which may result in partial overlap between training and testing clips. Although this design ensures fair comparison with existing baselines, it may inadvertently introduce data leakage and overestimate model generalization. To address this, we plan to implement stricter video-level separation protocols in future experiments. Lastly, while the CCPM modules are supported by formal mathematical formulations, the distinction between our proposed components and related mechanisms (e.g., ViLBERT-style cross-attention, original CGFN) requires further theoretical clarification and architectural comparison. We aim to deepen the analysis of the unique inductive biases introduced by our CCPM modules, particularly in the context of temporal anticipation and modality-specific interaction.

Author Contributions: W.S.: Conceptualization, Methodology, Resources, Writing—original draft, Writing—review and editing, Visualization. L.N.A., F.B.K. and P.S.B.S.: Software, Supervision, Project administration, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/asefjamilajwad/car-crash-dataset-ccd (accessed on 31 May 2020).

Acknowledgments: Finally, I am very grateful to my school, Universiti Putra Malaysia, for opening the methodology for us, so that we can have a better understanding of research. I hope that in the future we can continue to expand other influencing factors in this research and put forward better suggestions for solving traffic accidents.

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

References

1. World Health Organization. Save Lives: A Road Safety Technical Package. 2017. Available online: https://apps.who.int/iris/bitstream/handle/10665/255199/9789241511704-eng.pdf (accessed on 6 May 2018).
2. World Health Organization. *Global Status Report on Road Safety 2018*; World Health Organization: Geneva, Switzerland, 2019.
3. Heydari, S.; Hickford, A.; McIlroy, R.; Turner, J.; Bachani, A.M. Road Safety in Low-Income Countries: State of Knowledge and Future Directions. *Sustainability* **2019**, *11*, 6249. [CrossRef]
4. Anik, B.T.H.; Islam, Z.; Abdel-Aty, M. A Time-Embedded Attention-based Transformer for Crash Likelihood Prediction at Intersections Using Connected Vehicle Data. *Transp. Res. Part C Emerg. Technol.* **2024**, *169*, 104831. [CrossRef]
5. Siddique, I. Advanced Analytics for Predicting Traffic Collision Severity Assessment. *World J. Adv. Res. Rev.* **2024**, *21*, 30574. [CrossRef]
6. Chen, J.; Tao, W.; Jing, Z.; Wang, P.; Jin, Y. Traffic Accident Duration Prediction Using Multi-mode Data and Ensemble Deep Learning. *Heliyon* **2024**, *10*, e25957. [CrossRef] [PubMed]
7. Elsayed, H.A.G.; Syed, L. An Automatic Early Risk Classification of Hard Coronary Heart Diseases Using Framingham Scoring Model. In Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing, Cambridge, UK, 22–23 March 2017; pp. 1–8.
8. Ma, J.; Jia, C.; Yang, X.; Cheng, X.; Li, W.; Zhang, C. A Data-driven Approach for Collision Risk Early Warning in Vessel Encounter Situations Using Attention-BiLSTM. *IEEE Access* **2020**, *8*, 188771–188783. [CrossRef]
9. Fang, J.; Li, L.L.; Yang, K.; Zheng, Z.; Xue, J.; Chua, T.S. Cognitive Accident Prediction in Driving Scenes: A Multimodality Benchmark. *arXiv* **2022**, arXiv:2212.09381.
10. Fang, S.; Liu, J.; Ding, M.; Cui, Y.; Lv, C.; Hang, P.; Sun, J. Towards Interactive and Learnable Cooperative Driving Automation: A Large Language Model-Driven Decision-Making Framework. *arXiv* **2024**, arXiv:2409.12812. [CrossRef]
11. Jamshidi, H.; Jazani, R.K.; Khani Jeihooni, A.; Alibabaei, A.; Alamdari, S.; Kalyani, M.N. Facilitators and Barriers to Collaboration Between Pre-hospital Emergency and Emergency Department in Traffic Accidents: A qualitative study. *BMC Emerg. Med.* **2023**, *23*, 58. [CrossRef]
12. Adewopo, V.A.; Elsayed, N.; ElSayed, Z.; Ozer, M.; Abdelgawad, A.; Bayoumi, M. A Review on Action Recognition for Accident Detection in Smart City Transportation Systems. *J. Electr. Syst. Inf. Technol.* **2023**, *10*, 57. [CrossRef]
13. Liu, C.; Xiao, Z.; Long, W.; Li, T.; Jiang, H.; Li, K. Vehicle trajectory data processing, analytics, and applications: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–36. [CrossRef]
14. Wang, L.; Xiao, M.; Lv, J.; Liu, J. Analysis of Influencing Factors of Traffic Accidents on Urban Ring Road based on the SVM Model Optimized by Bayesian Method. *PLoS ONE* **2024**, *19*, e0310044. [CrossRef]
15. 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]
16. Hasan, F.; Huang, H. MALS-Net: A Multi-head Attention-based LSTM Sequence-to-sequence Network for Socio-temporal Interaction Modelling and Trajectory Prediction. *Sensors* **2023**, *23*, 530. [CrossRef] [PubMed]
17. Leon, F.; Gavrilescu, M. A Review of Tracking and Trajectory Prediction Methods for Autonomous Driving. *Mathematics* **2021**, *9*, 660. [CrossRef]

18. Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; et al. Aide: A Vision-driven Multi-view, Multimodal, Multi-tasking Dataset for Assistive Driving Perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 20459–20470.
19. Tang, Q.; Liang, J.; Zhu, F. A Comparative Review on Multimodal Sensors Fusion Based on Deep Learning. *Signal Process.* **2023**, *213*, 109165. [CrossRef]
20. Zheng, Y.; Xu, Z.; Wang, X. The fusion of deep learning and fuzzy systems: A state-of-the-art survey. *IEEE Trans. Fuzzy Syst.* **2021**, *30*, 2783–2799. [CrossRef]
21. Marchetti, F.; Mordan, T.; Becattini, F.; Seidenari, L.; Bimbo, A.D.; Alahi, A. CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting. *IEEE Trans. Intell. Veh.* 2024; 1–10, *early access*. [CrossRef]
22. He, R.; Zhang, C.; Xiao, Y.; Lu, X.; Zhang, S.; Liu, Y. Deep Spatio-temporal 3D Dilated Dense Neural Network for Traffic Flow Prediction. *Expert Syst. Appl.* **2024**, *237*, 121394. [CrossRef]
23. Liu, Z.; Yang, N.; Wang, Y.; Li, Y.; Zhao, X.; Wang, F.Y. Enhancing Traffic Object Detection in Variable Illumination with RGB-Event Fusion. *arXiv* **2023**, arXiv:2311.00436. [CrossRef]
24. Chitta, K.; Prakash, A.; Jaeger, B.; Yu, Z.; Renz, K.; Geiger, A. Transfuser: Imitation with Transformer-based Sensor Fusion for Autonomous Driving. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 12878–12895. [CrossRef]
25. Tan, H.; Bansal, M. Lxmert: Learning Cross-modality Encoder Representations from Transformers. *arXiv* **2019**, arXiv:1908.07490. [CrossRef]
26. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv* **2019**, arXiv:1908.02265.
27. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal Image-text Representation Learning. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 104–120.
28. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Volume 1: Long Papers. pp. 2236–2246.
29. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the Conference. Association for Computational Linguistics. Meeting, Florence, Italy, 28 July–2 August 2019; Volume 2019, p. 6558.
30. Rahman, W.; Hasan, M.K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.P.; Hoque, E. Integrating Multimodal Information in Large Pretrained Transformers. In Proceedings of the Conference. Association for Computational Linguistics. Meeting, Online, 5–10 July 2020; Volume 2020, p. 2359.
31. Li, W.; Liu, H.; Tang, H.; Wang, P.; Van Gool, L. Mhformer: Multi-hypothesis Transformer for 3d Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13147–13156.
32. Dhanasekaran, S.; Gopal, D.; Logeshwaran, J.; Ramya, N.; Salau, A.O. Multi-model Traffic Prediction in Smart Cities Using Graph Neural Networks and Transformer-Based Multi-Source Visual Fusion for Intelligent Transportation Management. *Int. J. Intell. Transp. Syst. Res.* **2024**, *22*, 518–541.
33. Li, L.; Dou, Y.; Zhou, J. Traffic Accident Detection based on Multimodal Knowledge Graphs. In Proceedings of the 2023 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence (RICAI), Hangzhou, China, 1–3 December 2023; pp. 644–647.
34. Zhang, Y.; Tiwari, P.; Zheng, Q.; El Saddik, A.; Hossain, M.S. A Multimodal Coupled Graph Attention Network for Joint Traffic Event Detection and Sentiment Classification. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 8542–8554. [CrossRef]
35. Ektefaie, Y.; Dasoulas, G.; Noori, A.; Farhat, M.; Zitnik, M. Multimodal Learning with Graphs. *Nat. Mach. Intell.* **2023**, *5*, 340–350. [CrossRef]
36. Kong, X.; Xing, W.; Wei, X.; Bao, P.; Zhang, J.; Lu, W. STGAT: Spatial-temporal Graph Attention Networks for Traffic Flow Prediction. *IEEE Access* **2020**, *8*, 134363–134372. [CrossRef]
37. Zhang, Y.; Dong, X.; Shang, L.; Zhang, D.; Wang, D. A Multimodal Graph Neural Network Approach to Traffic Risk Prediction in Smart Urban Sensing. In Proceedings of the 2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Como, Italy, 22–25 June 2020; pp. 1–9.
38. Ulutan, O.; Iftekhar, A.S.M.; Manjunath, B.S. Vsgnet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13617–13626.
39. Liu, A.A.; Tian, H.; Xu, N.; Nie, W.; Zhang, Y.; Kankanhalli, M. Toward Region-aware Attention Learning for Scene Graph Generation. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 7655–7666. [CrossRef] [PubMed]

40. Sultana, S.; Ahmed, B. Robust Nighttime Road Lane Line Detection Using Bilateral Filter and SAGC Under Challenging Conditions. In Proceedings of the 2021 IEEE 13th International Conference on Computer Research and Development (ICCRD), Beijing, China, 5–7 January 2021; pp. 137–143.
41. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
42. Yao, H.; Wang, L.; Cai, C.; Wang, W.; Zhang, Z.; Shang, X. Language Conditioned Multi-Scale Visual Attention Networks for Visual Grounding. *Image Vis. Comput.* **2024**, *150*, 105242. [CrossRef]
43. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
44. Pearson, K. VII. *Note on Regression and Inheritance in the Case of Two Parents. Proc. R. Soc. Lond.* **1895**, *58*, 240–242.
45. Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to Predict Where Humans Look. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 2106–2113.
46. Borji, A.; Sihite, D.N.; Itti, L. Quantitative Analysis of Human-model Agreement in Visual Saliency Modeling: A Comparative Study. *IEEE Trans. Image Process.* **2012**, *22*, 55–69. [CrossRef]
47. 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; Springer International Publishing: Berlin/Heidelberg, Germany, 2017. Revised Selected Papers, Part IV 13. pp. 136–153.
48. Suzuki, T.; Kataoka, H.; Aoki, Y.; Satoh, Y. Anticipating Traffic Accidents with Adaptive Loss and Large-scale Incident db. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3521–3529.
49. Bao, W.; Yu, Q.; Kong, Y. Uncertainty-based Traffic Accident Anticipation with Spatio-temporal Relational Learning. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2682–2690.
50. Bao, W.; Yu, Q.; Kong, Y. Drive: Deep Reinforced Accident Anticipation with Visual Explanation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7619–7628.
51. Xia, Y.; Zhang, D.; Kim, J.; Nakayama, K.; Zipser, K.; Whitney, D. Predicting Driver Attention in Critical Situations. In Proceedings of the Computer vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2019. Revised Selected Papers, Part v 14. pp. 658–674.
52. Palazzi, A.; Abati, D.; Solera, F.; Cucchiara, R. Predicting the Driver's Focus of Attention: The DR (eye) VE Project. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1720–1733. [CrossRef] [PubMed]
53. Wang, W.; Shen, J.; Xie, J.; Cheng, M.M.; Ling, H.; Borji, A. Revisiting Video Saliency Prediction in the Deep Learning Era. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 220–237. [CrossRef] [PubMed]
54. 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]