Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Empowering Multimodal Road Traffic Profiling with Vision Language Models and Frequency Spectrum Fusion

Haolong Xiang[1], **Xiaolong Xu**[1*], **Guangdong Wang**[1], **Xuyun Zhang**[2], **Xiaoyong Li**[3], **Qi Zhang**[4], **Amin Beheshti**[2] and **Wei Fan**[5*]
[1]Nanjing University of Information Science and Technology
[2]Macquarie University
[3]National University of Defense Technology
[4]Tongji University
[5]University of Auckland
{hlxiang, xlxu, 202412211538}@nuist.edu.cn, {xuyun.zhang, amin.beheshti}@mq.edu.au,
sayingxmu@nudt.edu.cn, zhangqi_cs@tongji.edu.cn, wei.fan@auckland.ac.nz

Abstract

With the rapid urbanization in the modern era, smart traffic profiling based on multimodal sources of data has been playing a significant role in ensuring safe travel, reducing traffic congestion and optimizing urban mobility. Most existing methods for traffic profiling on the road level usually utilize single-modality data, i.e., they mainly focus on image processing with deep vision models or auxiliary analysis on the textual data. However, the joint modeling and multimodal fusion of the textual and visual modalities have been rarely studied in road traffic profiling, which largely hinders the accurate prediction or classification of traffic conditions. To address this issue, we propose a novel multimodal learning and fusion framework for road traffic profiling, named TraffiCFUS. Specifically, given the traffic images, our TraffiCFUS framework first introduces Vision Language Models (VLMs) to generate text and then creates tailored prompt instructions for refining this text according to the specific scene requirements of road traffic profiling. Next, we apply the discrete Fourier transform to convert multimodal data from the spatial domain to the frequency domain and perform a cross-modal spectrum transform to filter out irrelevant information for traffic profiling. Furthermore, the processed spatial multimodal data is combined to generate fusion loss and interaction loss with contrastive learning. Finally, extensive experiments on four real-world datasets illustrate superior performance compared with the state-of-the-art approaches.

1 Introduction

The rapid development of smart cities facilitates the digitalization of road traffic applications, i.e., the importance of understanding and accurately analyzing road traffic conditions cannot be overstated [Xu *et al.*, 2023b]. Efficient transportation management and profiling are essential for the smooth functioning of economies, ensuring the timely delivery of goods and services, and facilitating the daily commute of people [Chen *et al.*, 2023; Fan *et al.*, 2022]. Urban planning also heavily relies on a comprehensive understanding of traffic conditions to design sustainable and livable cities [Yan *et al.*, 2024]. Moreover, public safety is closely intertwined with traffic prediction, as accidents and congestion can pose significant risks to the well-being of individuals. Thus, road traffic profiling equips decision-makers with essential insights by means of traffic prediction, transportation management, and classification of road traffic conditions [Liu *et al.*, 2024].

Traditional methods of road traffic profiling apply basic machine-learning techniques to analyze road traffic conditions, which fails to handle complex and dynamic modern traffic systems [Sommer *et al.*, 2010]. Besides, web-sourced data with large volumes presents greater challenges to the accuracy of these traditional methods. Web sources can provide multimodal information, including real-time traffic updates, images captured by satellite or surveillance cameras, social media posts related to traffic incidents, and textual descriptions of traffic conditions [Xu *et al.*, 2023a]. How to effectively utilize these multimodal data for road traffic profiling is highly challenging.

According to task-specific learning, we can divide the road traffic profiling problems into four categories: traffic flow analysis, accident detection and analysis, fire detection and analysis, and travel time estimation [Liu *et al.*, 2023]. Deep learning techniques, like convolutional neural networks based methods [He *et al.*, 2016; Alam *et al.*, 2023], recurrent neural networks based methods [Jin *et al.*, 2017; Zheng *et al.*, 2020], graph neural networks based methods [Zhang *et al.*, 2023; Deng *et al.*, 2024], and transformer-based methods [Lin *et al.*, 2022a; Xu *et al.*, 2024], have been widely used to analyze these traffic conditions, which helps to accurately identify images or text descriptions of road traffic and make driving decisions. However, as the current road traffic scenario is complex and ever-changing with multimodal data sources, merely analyzing the information in images or text is insufficient to capture the comprehensive road traffic conditions, thereby affecting traffic decision-making and travel safety.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Large language models (LLMs) possess a strong capacity for handling multimodal data, particularly in the understanding and generalization of textual data [Khattar *et al.*, 2019; Feng *et al.*, 2024], and offer more comprehensive information understanding via the complementarity of modalities to address information loss [Chen *et al.*, 2022]. As an example, Qian et al. [Qian *et al.*, 2021] proposed a multimodal framework that jointly captures multimodal context information and the hierarchical semantics of text by BERT and ResNet, which enhances detection accuracy by fusing inter-modality and intra-modality relationships. To optimize multi-vehicle dispatching and navigation in smart cities, Chen et al.[Chen *et al.*, 2024] proposed an LLM-driven framework for efficient task allocation and an RL-based module for cooperative navigation and handling heterogeneous vehicles. Recently, Yan et al. [Yan *et al.*, 2024] utilized LLMs to enhance the textual information and adopted contrastive learning for image and text fusion, which produces multimodal representations for urban region profiling. These methods analyze multimodal data by using the powerful information extraction ability of LLMs, but the task-specific applications hinder the exploration of road traffic profiling. Currently, it is still a challenging task to analyze complex road traffic conditions with web-sourced images and a powerful complement of textual data.

To address the above issues, we propose the first-ever multimodal fusion framework (TraffiCFUS) in road traffic profiling, which refines the text with the prompt design of Vision Language Models (VLMs) and optimizes the multimodal representations with frequency spectrum learning. The main contributions of our work are three-fold:

- TraffiCFUS is a very early attempt for multimodal learning and fusion framework for road traffic profiling, which introduces VLMs to generate text and then creates tailored prompt instructions for refining this text.
- We convert the multimodal data from the spatial domain to frequency domain and perform a cross-modal spectrum transform to filter out irrelevant information. Four losses with different aspects are designed to construct the objective loss for downstream task optimization.
- Extensive experiments on four public traffic datasets illustrate the effectiveness of our framework. We also conduct a series of ablation studies to show the influence of different components in our framework. Our source code is available at https://anonymous.4open.science/r/TraffiCFUS-87EF.

2 Related Work

2.1 Deep Learning for Road Traffic Profiling

Deep learning techniques have been widely applied in traffic and driver profiling, demonstrating significant advancements in understanding and predicting complex behaviors within these domains [Dui *et al.*, 2024]. For example, Cura et al. [Cura *et al.*, 2020] utilized data from the CAN Bus system to train LSTM and 1D-CNN models, finding that CNNs are particularly effective at distinguishing aggressive driving styles. This research underscores the importance of deep learning in driver profiling, with applications extending to fields like insurance and fleet management. Additionally, Abdelrahman et al. [Abdelrahman *et al.*, 2020] presented a robust data-driven framework leveraging supervised machine learning to evaluate drivers' risk profiles. Using the SHRP2 dataset, they identified key risk factors through Random Forest and Deep Neural Network models, enabling real-time risk assessment via cloud-based applications. Sekula et al. [Sekuła *et al.*, 2018] introduced a machine-learning framework trained on vehicle probe data for traffic volume estimation. Conducted in Maryland, this study showed that neural networks integrated with profiling methods can improve volume estimations by 24% over traditional techniques. These approaches demonstrate the power of deep learning models in addressing the challenges of road traffic profiling, but they fail to utilize multimodal data of text and image simultaneously to train the real-world applicable models.

2.2 LLMs for Road Traffic Profiling

Recent advances in traffic analysis have leveraged Large language models (LLMs) for road traffic profiling [Cao *et al.*, 2024], particularly in traffic condition classification where LLMs, such as GPT-2 and GPT-3, are exploited. Chen et al. [Chen *et al.*, 2024] proposed LiMeDa, which combined an LLM for efficient task allocation and an RL-based module for cooperative navigation and handling heterogeneous vehicles. However, LiMeDa has not fully addressed the extraction and integration of multi-modal information, which is crucial for effectively handling complex tasks across both visual and textual data. Hu et al. [Hu *et al.*, 2024] extended a multimodal LLM (BLIVA) that combined visual and textual information, by leveraging learned query embedding and encoded patch embedding, to address complex visual question-answering (VQA) tasks. Yang et al. [Yang *et al.*, 2024] introduced EMMA, which trained VLM agents through cross-modal imitation learning to adapt better to visual world dynamics. Besides, Li et al. [Li *et al.*, 2024] designed OmniActions, which introduced a novel pipeline to process multimodal sensory inputs and form explicit reasoning. OmniActions made it possible to predict context-aware digital actions based on an integrated design space.

However, these approaches did not fully combine the multimodal data to generate powerful representations for road traffic profiling. In this paper, we aim to leverage LLMs to fuse multimodal data, which enables a more comprehensive understanding of road traffic conditions and provides a multimodal framework for accurate road traffic profiling.

3 Problem Definition

Definition 1 (Traffic Area)**.** *According to the previous studies, we can divide the whole transportation of a city into M traffic scene areas for road traffic profiling.*

Definition 2 (Traffic Image)**.** *By using cameras or other sensor devices, traffic scene images are captured to depict various aspects of traffic conditions in a traffic area $\mathbb{A}$, including images of vehicles, pedestrians, traffic signs, signals, and the overall road/weather environment. Each input traffic image can be represented as $g_{\mathbb{A}}$, which satisfies the condition:*

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

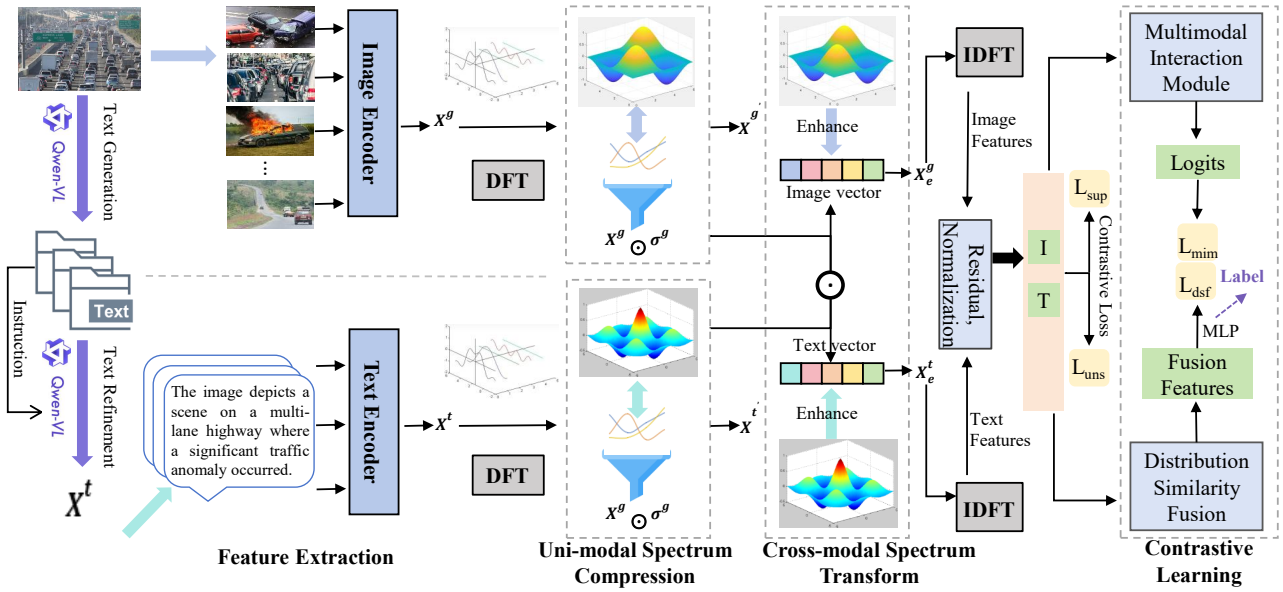

Figure 1: The framework of proposed TraffiCFUS. Our method consists of four main steps: text refinement and feature extraction, uni-modal spectrum compression, cross-modal spectrum transform, and contrastive learning.

$g_{\mathbb{A}} \in \mathbb{R}^{L \cdot H \cdot 3}$, *$L$ and H are length and width.*

Definition 3 (Traffic Description)**.** *We use text description $t_{\mathbb{A}}$ to show the traffic condition in a traffic scene area $\mathbb{A}$. These text descriptions can be extracted from traffic scene images by image captioning tools, i.e., the well-trained VLM models, such as VL-Plus, Llama-adaptor and InternVL2.*

Definition 4 (Traffic Category)**.** *Different traffic scene Categories represent different traffic conditions. Each category provides distinct insights into the specific characteristics and challenges of that particular traffic condition, which is crucial for effective traffic profiling. In this paper, we use four categories $Cat = \{C_f, C_a, C_c, C_l\}$ to represent fire, accident, congestion and light traffic, respectively.*

The processing of traffic conditions is defined as a multi-modal classification problem, where multimodal u contains two parts of image g and text t. Given a traffic dataset $D = \{d_1, d_2, ..., d_n\}$, each data instance is denoted as (x_i, y_i) and $x_i = \{x^g, x^t\}$, where x^g represents the image element and x^t represents the text element. Besides, $y_i = \{1, 2, ..., m\} (m \geq 2)$ and m represent the number of types of traffic conditions, i.e., y_i is the label of data instances. When $m = 2$, the traffic scene prediction is a binary classification problem, otherwise, it is a multi-class classification problem. Then, a mapping function $\mathcal{F}$ is designed to map all the data, including traffic image, text and other environmental data (e.g., road, pedestrian), to a vector $\boldsymbol{v}_{\mathbb{A}} = \mathcal{F}(g_{\mathbb{A}}, t_{\mathbb{A}})$. Finally, this vector can be used to predict the traffic condition by $\boldsymbol{v}_{\mathbb{A}} \rightarrow y_i$.

4 Methodology

We propose a text-refined road traffic profiling framework (TraffiCFUS) with multimodal interaction and fusion, which can effectively identify multimodal traffic conditions from the web. As shown in Figure 1, our method consists of four parts: text refinement and feature extraction, uni-modal spectrum compression, cross-modal spectrum transform, and contrastive learning. The first step is the feature extraction of images and text, which applies VLMs to refine text and generate the initial representations. Then, these features are transformed into spectral features by discrete Fourier transform (DFT). The second step is to compress the uni-modal spectrum and use a filter bank to filter out useless information. The third step is enhancing the informative spectrum and suppressing the irrelevant spectrum by the cross-modal transform. Finally, the spectral features are converted back into the spatial domain by inverse discrete Fourier transform (IDFT) and the output features are used for loss calculation with contrastive learning.

4.1 Text Refinement and Feature Extraction

The first step of TraffiCFUS is to generate texts from images and enhance these texts by prompt design with VLMs. As shown in Figure 2, an image-to-text foundation model, Qwen-VL plus version [Yue *et al.*, 2024], is applied to analyze images and generate descriptive text. With its ability to accurately describe images, Qwen-VL can enhance the way we interact with traffic scenes and improve the accessibility of information. But the text directly generated by VLMs is rough and contains a lot of useless information, which will affect the analysis and classification of traffic conditions. As demonstrated in Figure 2, an example is presented to illustrate that the descriptions of road traffic conditions are highly redundant and contain irrelevant events.

The process for the above text is to use the Qwen-VL and design an instruction-based prompt to optimize the generated text. We have designed three instructions to optimize the description of traffic conditions. One is to get a detailed description of all traffic conditions, including vehicles, pedestrians,

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

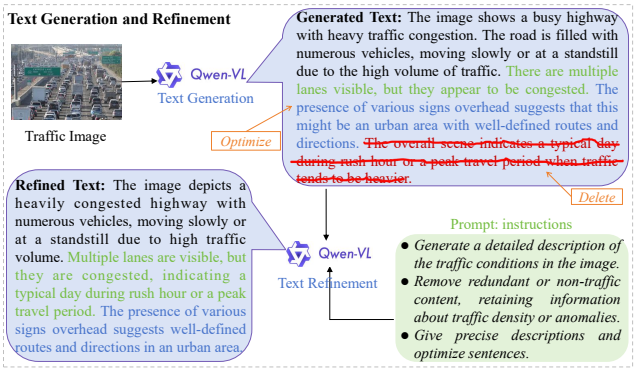

Figure 2: The process of text generation and refinement. Prompt instructions are designed to refine the text by a VLM.

traffic signs, signals, etc. Another is to remove redundant or non-traffic content. The other is to simplify the text description and retain the most crucial information. As a result, the text is enhanced with a concise and high-quality summary for traffic condition judgment, shown in Figure 2.

Given a traffic area A with its satellite image g_A, we divide it into a sequence of patches g_p with dimensions $h \times w$ (height $\times$ width). Then, a patch embedding is used to embed the image into a dense vector as:

$$\boldsymbol{e_p^g} = \boldsymbol{W_p} g_p^T + b_p, \tag{1}$$

where $\boldsymbol{W_p}$ and b_p are random variables, which need to be trained by models. The final embedding of a image consists of a patch embedding and a positional embedding $\boldsymbol{e_{po}^g}$. Then, the final embedding of an image can be calculated by $\boldsymbol{e_p^{g'}} = \boldsymbol{e_p^g} + \boldsymbol{e_{po}^g}$. To produce a meaningful representation, a pre-trained CNN, such as VGGNet, ResNet, or Inception, is typically used to calculate the patch features:

$$\boldsymbol{Pt} = CNN(\boldsymbol{e_p^{g'}}). \tag{2}$$

An image consists of several patches with the features represented as $x^g = \{\boldsymbol{Pt_1}, \boldsymbol{Pt_2}, ..., \boldsymbol{Pt_n}\}$.

For textual feature extraction, LLMs like BERT (Bidirectional Encoder Representations from Transformers) is used to generate a latent textual representation. BERT takes a sequence of words as input and outputs a context-aware embedding for each word. To obtain a text embedding from BERT, one can take the average or sum of the output embeddings for all the words in the text, or use the embedding of a special token (like [CLS]) that represents the entire text. Given a text sequence $x^t = \{\boldsymbol{Se_1}, \boldsymbol{Se_2}, ..., \boldsymbol{Se_m}\}$ (m is the number of words), we can obtain the word embedding $\boldsymbol{e_w^t}$ and the positional embedding $\boldsymbol{e_{po}^t}$ to encode each word. Then, the feature representation of one word can be calculated by:

$$\boldsymbol{Se_1} = \boldsymbol{e_w^t} + \boldsymbol{e_{po}^t}. \tag{3}$$

A text consists of a sequence of words that has m feature representations.

4.2 Uni-modal Spectrum Compression

The spectrum of an image can reveal the distribution of different frequency spectrum components in the traffic image. The low-frequency part usually corresponds to the overall outline of the image and slowly changing areas, such as the background of a large area of traffic roads, vehicle conditions, pedestrian conditions, etc. The high-frequency part represents the details, edges, and texture information in the image. By analyzing the spectrum, features within a specific frequency range can be extracted in a targeted manner and used for tasks such as the analysis of traffic images and the classification of target categories. As shown in Figure 1, Discrete Fourier transform (DFT) is applied to transform the spatial features into spectrum features. Given the spatial signal x^g, the Fourier transform of the image embedding is defined as:

$$X^g[k] = \sum_{i=0}^{n-1} x^g[i] e^{-j\frac{2\pi}{n}ki}, \tag{4}$$

where $X^g \in \mathbb{C}^{n \times d}$ is a complex tensor, k is the frequency index, and j is the imaginary unit. Thus, we can get the image spectrum $X^g[k]$ at the frequency $\frac{2\pi k}{n}$.

Besides, we can obtain the Fourier transform of the text embedding in a similar way, defined as:

$$X^t[k] = \sum_{i=0}^{m-1} x^t[i] e^{-j\frac{2\pi}{m}ki}, \tag{5}$$

where $X^t \in \mathbb{C}^{m \times d}$ is a complex tensor. Different text in an image has distinct spectral characteristics. For example, smooth text has less high-frequency content than rough text. Analyzing the above spectrum can help in road traffic profiling and image classification.

Then, we use uni-modal spectrum compression to process spectral characteristics, achieving a greater degree of feature compression while retaining more important information. We introduce a filter bank for each modality X^a ($a \in \{g, t\}$), which can divide the input frequency spectrum into multiple sub-bands. For uni-modal spectrum compression, the filter bank can be designed to target the specific frequency range where the uni-modal spectrum is located. For an FIR filter bank, the impulse response $H^a = [\boldsymbol{h_1^a}, \boldsymbol{h_2^a}, ..., \boldsymbol{h_k^a}]$ of k filter can be designed using windowing techniques. For example, we can use a Hamming window to obtain the following filter:

$$\boldsymbol{h_i^a} = w[i] \boldsymbol{h_{ideal}^a}, \tag{6}$$

where $w[i]$ is the Hamming window function and $\boldsymbol{h_i deal^a}$ is the ideal impulse response for the filter. The Hamming window function is given by: $w[i] = 0.54 - 0.46cos(\frac{2\pi i}{k-1})$, and k is the length of the filter. Through the uni-modal spectrum compression, we can get the spectrum as:

$$X^{a'} = \sum_{i=1}^{k} \frac{1}{l} |X^a|^2 \odot \boldsymbol{h_i^a}, \tag{7}$$

where $a \in \{g, t\}$, l represents the length of each modality, and $\odot$ represents the element-wise multiplication. the design of $|X^a|^2$ smooths the spectrum and highlights the main components, which helps with the subsequent learning of uni-modal compression. Besides, the Hamming window function helps to aggregate the main information in the traffic features, realizing efficient frequency domain feature compression.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

4.3 Cross-modal Spectrum Transform

Fixed spectrum compression has limited contributions in processing traffic images and cannot remove the higher noise in the frequency spectrum. Therefore, we introduce the average pooling scheme to enhance informative components. Average pooling smoothes the spectrum by reducing high-frequency noise and random fluctuations. By preserving overall trends and making the spectrum more regular, it enhances the efficiency of compression algorithms and identifies traffic image features more effectively.

Then, we also make a cross-modal spectrum transform for the uni-modal spectrum of image and text. This process offers an enhanced representation by capturing more complex relationships and features, improves understanding of the connections between visual and semantic aspects, enables multimodal analysis for complex road traffic, and provides robustness by being less affected by noise and variations in either modality. The enhanced spectrum X_e^a with cross-modal transform can be calculated by:

$$X_e^g = X^{g'} \odot Avg(X^{t'} \odot \Phi^{t'}), \tag{8}$$

$$X_e^t = X^{t'} \odot Avg(X^{g'} \odot \Phi^{g'}), \tag{9}$$

where $\odot$ represents the element-wise multiplication, $Ave(\cdot)$ represents the average pooling function, and $\Phi^{a'}$ represents a matrix that has the same dimension with $X^{a'}$.

After obtaining the enhanced frequency spectrum, we employ the inverse discrete Fourier transform (IDFT) to convert them back to the spatial representations for loss construction. The calculation of IDFT in image and text is:

$$x^g[i] = \frac{1}{n}\sum_{i=0}^{n-1} X^g[k] e^{j\frac{2\pi}{n}ki}, \tag{10}$$

$$x^t[i] = \frac{1}{m}\sum_{i=0}^{m-1} X^t[k] e^{j\frac{2\pi}{m}ki}. \tag{11}$$

Through the processing of the DFT and cross-model transform, we can fuse the multimodal features of images and texts and filter out unimportant information, thus providing more reliable features for road traffic profiling.

4.4 Multimodal Interaction and Fusion with Contrastive Learning

Contrastive Loss. To train the parameters of the proposed multimodal method, we apply a dual contrastive learning scheme. The first is a supervised loss $\mathcal{L}_{sup}$ that optimizes the model by maximizing the similarity of samples within the same class and minimizing the similarity of samples from different classes. Given a mini-batch $\mathcal{M}$ that contains $|\mathcal{M}|$ data instances, we can divide samples into l types $\{L_1, L_2, ...L_l\}$ according to the traffic types of datasets. For an instance $s_j \in L_i$, we can produce a pairwise (s_j, s_p) to calculate the loss, where $s_p \in L_i, p \neq j$. Then, we can define the pairwise objective function with each instance and different types of samples $\mathcal{L}_i(\boldsymbol{x}^a, \boldsymbol{x}^a), a \in (g, t)$ [Lin *et al.*, 2022b]. The whole supervised contrastive learning loss is calculated by:

$$\mathcal{L}_{sup} = \sum_{\mathcal{M}}\sum_{\mathcal{N}}\Big(\sum_{s_j \in L_i}\frac{1}{|L_i|}\sum_{s_p \in L_i, p\neq j}\mathcal{L}_i(\boldsymbol{x}_j^a, \boldsymbol{x}_p^a)\Big), \tag{12}$$

where $\mathcal{N} = \{L_1, L_2, ...L_l\}$, $|L_i|$ represents the number of instances in L_i.

The second is an unsupervised loss that optimizes the model by maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs. Specifically, InfoNCE loss [He *et al.*, 2020] is applied in our scheme by a pairwise contrastive loss $\mathcal{L}_s(\boldsymbol{x}^g, \boldsymbol{x}^t)$ and $\mathcal{L}_s(\boldsymbol{x}^t, \boldsymbol{x}^g)$. The whole unsupervised contrastive loss consisted of two InfoNCE losses, defined as follows:

$$\mathcal{L}_{uns} = \frac{1}{2|\mathcal{M}|}\sum_{i=1}^{\mathcal{M}}[\mathcal{L}_s(\boldsymbol{x}_i^g, \boldsymbol{x}_i^t) + \mathcal{L}_s(\boldsymbol{x}_i^t, \boldsymbol{x}_i^g)], \tag{13}$$

where $|\mathcal{M}|$ represents the number of instances in a selected batch.

Multimodal Interaction Loss. Previous multimodal interaction is very shallow and fails to learn the multimodal representations well. Thus, we design a multimodal interaction loss based on the work [Yu *et al.*, 2020] and a multimodal fusion loss with distribution similarity learning. The decoder architecture based on the Transformer is utilized to merge unimodal visual and textual representations into multimodal ones. Specifically, the multimodal cross-attention uses image modality as the query and text modality as the keys and values. After the interaction, we can generate the textual description for the road traffic profiling by optimising the language modelling loss $\mathcal{L}_{mim}$. This interaction loss can be learned through minimizing the conditional likelihood of the text description $t_{\mathbb{A}}$ in traffic area $\mathbb{A}$, which can be calculated by:

$$\mathcal{L}_{mim} = \sum_{i=1}^{\mathbb{A}} logP_\theta(t_i | t_{<l}, g). \tag{14}$$

Multimodal Fusion Loss. Except for the multimodal interaction, we also design a multimodal fusion scheme by measuring the distribution similarity between image and text representations. Specifically, Jensen-Shannon (JS) divergence between two types of representations is measured to calculate the distribution similarity and then outputs the classification of traffic conditions. To generate the posterior probability of the training instances, we need to produce an approximation of sample distribution $\mathbb{I}$. Then the posterior probability of the image and text can be represented as $\mathbb{I}(\boldsymbol{u}^g|\boldsymbol{x}^g)$ and $\mathbb{I}(\boldsymbol{u}^t|\boldsymbol{x}^t)$, separately. The JS divergence of fusion modal in $\boldsymbol{x}^a$ can be calculated as:

$$\zeta = JS(\mathbb{I}(\boldsymbol{u}^g|\boldsymbol{x}^g)||\mathbb{I}(\boldsymbol{u}^t|\boldsymbol{x}^t)), \tag{15}$$

where ζ represents the similarity measured by JS divergence, and $JS(\cdot)$ represents the JS divergence function. Then, we can use the distribution similarity to calculate the representations after multimodal fusion as follows:

$$\boldsymbol{r}^a = (1-\zeta)(\boldsymbol{W}^g\boldsymbol{x}^g + \boldsymbol{W}^t\boldsymbol{x}^t) + \zeta\boldsymbol{x}^g + \zeta\boldsymbol{x}^t, \tag{16}$$

where $\boldsymbol{W}^g$ and $\boldsymbol{W}^t$ represents the training parameters of image and text. Through a fully connected layer, we can obtain the labels of data instances $\hat{y}$.

In road traffic profiling, the classification of traffic conditions is an important application, which helps to predict future congestion on roads and ensure safe travel. Because traffic

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Tasks	Sparse (1) & Accident (0)								Sparse (1) & Fire (0)							
Methods	Acc	F1	class 1			class 0			Acc	F1	class 1			class 0		
			Pre	Rec	F1	Pre	Rec	F1			Pre	Rec	F1	Pre	Rec	F1
Resnet-18	0.851	0.850	0.936	0.764	0.842	0.791	0.942	0.860	0.921	0.921	0.943	0.889	0.915	0.900	0.951	0.924
att-RNN	0.873	0.871	0.901	0.891	0.896	0.882	0.915	0.898	0.873	0.871	0.920	0.891	0.905	0.882	0.935	0.908
EANN	0.892	0.892	0.933	0.951	0.942	0.924	0.935	0.929	0.906	0.908	0.944	0.925	0.934	0.914	0.907	0.910
MAVE	0.881	0.885	0.922	0.954	0.938	0.872	0.928	0.899	0.931	0.931	0.903	0.862	0.882	0.931	0.956	0.943
HMCAN	0.926	0.924	0.913	0.905	0.909	0.947	0.936	0.941	0.943	0.942	0.951	0.957	0.954	0.945	0.957	0.951
CAFE	0.927	0.926	0.940	0.942	0.941	0.914	0.895	0.904	0.913	0.910	0.943	0.914	0.928	0.951	0.922	0.936
LogicDM	0.931	0.934	0.946	0.911	0.928	0.936	0.947	0.941	0.934	0.933	0.966	0.927	0.946	0.955	0.944	0.949
UrbanCLIP	0.844	0.843	0.896	0.790	0.840	0.799	0.901	0.847	0.943	0.943	0.919	**0.973**	0.945	**0.970**	0.912	0.940
TraffiCFUS	**0.973**	**0.973**	**0.970**	**0.978**	**0.974**	**0.976**	**0.967**	**0.972**	**0.968**	**0.968**	**0.973**	0.965	**0.969**	0.963	**0.972**	**0.967**

Table 1: Comparison of different methods on binary classification tasks of dataset Traffic-Net_2. Our proposed framework TraffiCFUS has better accuracy performance and higher robustness compared with other baselines over all datasets. "Acc" represents the accuracy performance, "Pre" represents the precision performance, and "Rec" represents the recall performance.

classification is a multi-classification problem, we apply the multi-class cross-entropy loss as the final fusion loss:

$$\mathcal{L}_{dsf} = \frac{1}{l}\sum_{i}\mathcal{L}_i = -\mathbb{E}_{y\sim\hat{Y}}\sum_{c=1}^{l} y_c log(\hat{y}_c), \qquad (17)$$

where l represents the number of labels of traffic conditions.

Finally, the objective loss can be calculated by two contrastive losses, multimodal interaction loss, and multimodal fusion Loss, which are denoted as:

$$\mathcal{L} = \alpha\mathcal{L}_{sup} + \beta\mathcal{L}_{uns} + \gamma\mathcal{L}_{mim} + \mathcal{L}_{dsf}, \qquad (18)$$

where α, β, and γ are three hyperparameters to balance the influence of different losses.

5 Experiments

Comprehensive experiments are conducted on four real-world road traffic datasets to address the research questions (RQ) outlined below.

- **RQ1:** Does TraffiCFUS outperform existing approaches in road traffic classification tasks?
- **RQ2:** What is the impact of different components (e.g., enhanced text, multimodal interaction) in TraffiCFUS?
- **RQ3:** How effective is the feature visualization of the proposed TraffiCFUS framework?

5.1 Experimental Setting

Baselines. Our framework is compared with eight state-of-the-art baselines: Resnet-18 [He *et al.*, 2016], att-RNN [Jin *et al.*, 2017], EANN [Wang *et al.*, 2018], MVAE [Khattar *et al.*, 2019], HMCAN [Qian *et al.*, 2021], CAFE [Chen *et al.*, 2022], LogicDM [Liu *et al.*, 2023], and UbanCLIP [Yan *et al.*, 2024]. We evaluate all methods on four real-world datasets, including "Traffic-Net_2", "Traffic-Net_4", "DAWN", and "TCN". We refer the reader to Appendix [1] for more details of the baselines and datasets.

Metrics. The performance is evaluated by Accuracy, Precision, Recall, and F1 score metrics[Liu *et al.*, 2023; Yan *et al.*, 2024]. To ensure a fair comparison, we follow the optimal parameter settings of the baselines. Please see the Appendix for more details about the parameter settings. All experiments are run 15 times and averaged results are reported.

[1]Please refer to the version of this paper with Appendix in arXiv.

Datasets	TrafficNet_4		DAWN		TCN	
Methods	Acc	F1	Acc	F1	Acc	F1
Resnet-18	0.793	0.792	0.736	0.735	0.787	0.787
att-RNN	0.773	0.769	0.725	0.724	0.755	0.756
EANN	0.823	0.822	0.767	0.763	0.751	0.751
MAVE	0.829	0.829	0.771	0.774	0.767	0.762
HMCAN	0.872	0.871	0.776	0.778	0.779	0.780
CAFE	0.862	0.864	0.781	0.779	0.785	0.788
LogicDM	0.846	0.847	0.764	0.764	0.793	0.795
UrbanCLIP	0.842	0.842	0.781	0.780	0.751	0.750
TraffiCFUS	**0.894**	**0.893**	**0.810**	**0.808**	**0.816**	**0.814**

Table 2: Comparison of different methods on complex real-world datasets. TraffiCFUS outperforms other baselines on both accuracy and F1 score results.

5.2 RQ1: Performance Comparison

We conduct comparative experiments on four datasets of road traffic classification, which helps to evaluate the accuracy performance of TraffiCFUS and the baseline methods. Table 1 shows the experimental results for the binary classification tasks in sparse vs. accident and sparse vs. fire. Our framework outperforms other baselines on almost all metrics, except for the results of recall and precision on the task "Sparse and Fire". But the f1 score of our framework ranks the top one on this task. Besides, LogicDM performs well on task "Sparse and Accident". HMCAN performs well on task "Sparse and Fire". Although UrbanCLIP ranks the first result on the recall and precision of the task "Sparse and Fire", this method's performance is less stable across different datasets. The classical methods, like att-RNN, EANN and MAVE did not work well on both tasks.

Table 2 details the experimental results of complex real-world datasets on different methods. We can conclude that TraffiCFUS performs the best on all datasets compared with other baselines. On dataset "TrafficNet_4", the following model is HMCAN, which is about 2% worse than our proposed framework. The second ranking method UrbanCLIP is about 3% worse than our proposed framework on dataset "DAWN" and LogicDM is about 2% worse than our proposed framework on dataset "TCN". Besides, Resnet-18 and att-RNN have poor performance on multi-class classification, indicating the limitations in multimodal data processing. The above experimental results and analysis confirm the superi-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Tasks	Sparse (1) & Accident (0)								Sparse (1) & Fire (0)							
Sets	Acc	F1	class 1			class 0			Acc	F1	class 1			class 0		
			Pre	Rec	F1	Pre	Rec	F1			Pre	Rec	F1	Pre	Rec	F1
TraffiCFUS	**0.973**	**0.973**	**0.970**	**0.978**	**0.974**	**0.976**	**0.967**	**0.972**	**0.968**	**0.968**	**0.973**	**0.965**	**0.969**	**0.963**	**0.972**	**0.967**
-w/o TR	0.953	0.953	0.942	0.966	0.954	0.965	0.941	0.953	0.919	0.919	0.935	0.901	0.917	0.904	0.938	0.921
-w/o MIM	0.951	0.951	0.939	0.964	0.951	0.964	0.939	0.951	0.957	0.957	0.959	0.956	0.957	0.957	0.957	0.957
-w/o FUS	0.951	0.951	0.937	0.966	0.951	0.965	0.936	0.950	0.953	0.953	0.951	0.957	0.954	0.955	0.949	0.952
-w/o CL	0.936	0.936	0.921	0.953	0.937	0.953	0.918	0.935	0.942	0.941	0.951	0.931	0.941	0.934	0.951	0.942

Table 3: Influence of different modules in the TraffiCFUS.

(a) TraffiCFUS (b) -w/o TR (c) -w/o MIM (d) -w/o FUS (e) -w/o CL

Figure 3: T-SNE visualization of the influence of different modules on the feature generation.

ority of our framework, with a detailed feasibility analysis provided in the Appendix.

5.3 RQ2: Ablation Studies

To investigate the influence of different modules in the proposed framework, we conduct a series of ablation experiments on dataset "TrafficNet_2". Specifically, we remove the text refinement module (denoted as -w/o TR), the multimodal interaction module (denoted as -w/o MIM), the fusion module (denoted as -w/o FUS), and the contrastive loss (denoted as -w/o CL) from the TraffiCFUS framework to show the performance on accuracy and F1 score. The detailed experimental results of these components' influence are shown in Table 3, which illustrates that the lack of any component will reduce the accuracy performance of our proposed framework.

Firstly, the absence of text refinement results in the generation of text containing irrelevant information and significant redundancy, which interferes with information extraction and leads to lower model accuracy. Without the multimodal interaction module, the multimodal interaction loss can not be obtained, leading to insufficient integration of visual and textual representations and inadequate multimodal representation learning. The absence of the fusion module meant that the model directly uses the inverse Fourier-transformed text and image features for prediction, hindering effective multimodal information learning. Without the contrastive loss module, the model's performance in classification accuracy, feature representation separation, and multimodal feature fusion declines, which impairs the overall effectiveness of traffic profiling. In summary, the model achieves optimal performance only when all these modules are included.

5.4 RQ3: Feature Visualization

To further investigate the impact of various modules on the features generated by the model, we utilize T-SNE [van der Maaten and Hinton, 2008] dimensionality reduction to visualize the features before they enter the final linear layer, as shown in Figure 3. The dataset used for this analysis is the sparse traffic & accident sub-dataset. In the visualization, green points represent samples with sparse traffic, while grey points denote samples with accidents.

The visualization reveals that TraffiCFUS and its various ablated versions exhibit different degrees of overlap in feature representation. Notably, the full TraffiCFUS model shows the clearest boundaries between the two classes, indicating more distinct and well-separated feature representations. The ablated models (-w/o TR, -w/o MIM, -w/o FUS, and -w/o CL) demonstrate varying levels of feature overlap, which correlates with their reduced performance in classification tasks.

6 Conclusion and Future Work

This paper investigates different road traffic scenarios, including sparse traffic, congested traffic, accidents, fires, and weather conditions, to make a full profiling of traffic conditions. To analyze the multimodal data on traffic, we propose a first-ever multimodal fusion framework (TraffiCFUS) for road traffic profiling. Powered by VLMs, TraffiCFUS generates enhanced textual information from web-sourced images to assist in road traffic profiling. Moreover, TraffiCFUS filters the irrelevant information and keeps the informative representations by converting the spatial features to spectrum features. The design of multimodal feature interaction and fusion further enhances the extracted representations. Finally, extensive experiments demonstrate the superiority of our proposed framework. Ablation studies show how the components influence the performance of TraffiCFUS. In the future, our studies aim to design dynamic text-refine schemes for different textual applications and explore a wider range of application possibilities in different realistic scenarios.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 92267104, and Jiangsu Provincial Major Project on Basic Research of Cutting-edge and Leading Technologies, under grant no. BK20232032.

References

[Abdelrahman *et al.*, 2020] Abdalla Ebrahim Abdelrahman, Hossam S Hassanein, and Najah Abu-Ali. Robust data-driven framework for driver behavior profiling using supervised machine learning. *IEEE transactions on intelligent transportation systems*, 23(4):3336–3350, 2020.

[Alam *et al.*, 2023] Md Golam Rabiul Alam, Mahmudul Haque, Md Rafiul Hassan, Shamsul Huda, Mohammad Mehedi Hassan, Fred L Strickland, and Salman A AlQahtani. Feature cloning and feature fusion based transportation mode detection using convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):4671–4681, 2023.

[Cao *et al.*, 2024] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 21819–21830, 2024.

[Chen *et al.*, 2022] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022 (WWW)*, pages 2897–2905, 2022.

[Chen *et al.*, 2023] Jing Chen, Mengqi Xu, Wenqiang Xu, Daping Li, Weimin Peng, and Haitao Xu. A flow feedback traffic prediction based on visual quantified features. *IEEE Transactions on Intelligent Transportation Systems*, 24(9):10067–10075, 2023.

[Chen *et al.*, 2024] Ruiqing Chen, Wenbin Song, Weiqin Zu, ZiXin Dong, Ze Guo, Fanglei Sun, Zheng Tian, and Jun Wang. An llm-driven framework for multiple-vehicle dispatching and navigation in smart city landscapes. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2147–2153. IEEE, 2024.

[Cura *et al.*, 2020] Aslıhan Cura, Haluk Küçük, Erdem Ergen, and İsmail Burak Öksüzoğlu. Driver profiling using long short term memory (lstm) and convolutional neural network (cnn) methods. *IEEE Transactions on Intelligent Transportation Systems*, 22(10):6572–6582, 2020.

[Deng *et al.*, 2024] Xianwen Deng, Yijun Wang, and Zhi Xue. An-net: an anti-noise network for anonymous traffic classification. In *Proceedings of the ACM on Web Conference 2024 (WWW)*, pages 4417–4428, 2024.

[Dui *et al.*, 2024] Hongyan Dui, Songru Zhang, Meng Liu, Xinghui Dong, and Guanghan Bai. Iot-enabled real-time traffic monitoring and control management for intelligent transportation systems. *IEEE Internet of Things Journal*, 2024.

[Fan *et al.*, 2022] Wei Fan, Shun Zheng, Xiaohan Yi, Wei Cao, Yanjie Fu, Jiang Bian, and Tie Yan Liu. Depts: Deep expansion learning for periodic time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2022.

[Feng *et al.*, 2024] Yu Feng, Zhen Tian, Yifan Zhu, Zongfu Han, Haoran Luo, Guangwei Zhang, and Meina Song. Cp-prompt: Composition-based cross-modal prompting for domain-incremental continual learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2729–2738, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9729–9738, 2020.

[Hu *et al.*, 2024] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, volume 38, pages 2256–2264, 2024.

[Jin *et al.*, 2017] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia (ACM MM)*, pages 795–816, 2017.

[Khattar *et al.*, 2019] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference (WWW)*, pages 2915–2921, 2019.

[Li *et al.*, 2024] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. Omniactions: Predicting digital actions in response to real-world multimodal sensory inputs with llms. In *Proceedings of the CHI Conference on Human Factors in Computing Systems(CHI)*, pages 1–22, 2024.

[Lin *et al.*, 2022a] Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, and Jing Yu. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In *Proceedings of the ACM Web Conference 2022 (WWW)*, pages 633–642, 2022.

[Lin *et al.*, 2022b] Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. Modeling intra-and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics (ACL)*, volume 29, pages 7124–7135, 2022.

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

[Liu *et al.*, 2023] Hui Liu, Wenya Wang, and Haoliang Li. Interpretable multimodal misinformation detection with logic reasoning. *Findings of the Association for Computational Linguistics (ACL)*, 2023.

[Liu *et al.*, 2024] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

[Qian *et al.*, 2021] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multimodal contextual attention network for fake news detection. In *ACM SIGIR conference on research and development in information retrieval (ACM SIGIR)*, pages 153–162, 2021.

[Sekuła *et al.*, 2018] Przemysław Sekuła, Nikola Marković, Zachary Vander Laan, and Kaveh Farokhi Sadabadi. Estimating historical hourly traffic volumes via machine learning and vehicle probe data: A maryland case study. *Transportation Research Part C: Emerging Technologies*, 97:147–158, 2018.

[Sommer *et al.*, 2010] Christoph Sommer, Reinhard German, and Falko Dressler. Bidirectionally coupled network and road traffic simulation for improved ivc analysis. *IEEE Transactions on mobile computing*, 10(1):3–15, 2010.

[van der Maaten and Hinton, 2008] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Wang *et al.*, 2018] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multimodal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining (SIGKDD)*, pages 849–857, 2018.

[Xu *et al.*, 2023a] Haowen Xu, Andy Berres, Srikanth B Yoginath, Harry Sorensen, Phil J Nugent, Joseph Severino, Sarah A Tennille, Alex Moore, Wesley Jones, and Jibonananda Sanyal. Smart mobility in the cloud: Enabling real-time situational awareness and cyber-physical control through a digital twin for traffic. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3145–3156, 2023.

[Xu *et al.*, 2023b] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13712–13722, 2023.

[Xu *et al.*, 2024] Xiaolong Xu, Chenbin Li, Haolong Xiang, Lianyong Qi, Xuyun Zhang, and Wanchun Dou. Attention based document-level relation extraction with none class ranking loss. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Yan *et al.*, 2024] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference (WWW)*, pages 4006–4017, 2024.

[Yang *et al.*, 2024] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 26275–26285, 2024.

[Yu *et al.*, 2020] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2020.

[Yue *et al.*, 2024] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024.

[Zhang *et al.*, 2023] Haozhen Zhang, Le Yu, Xi Xiao, Qing Li, Francesco Mercaldo, Xiapu Luo, and Qixu Liu. Tfe-gnn: A temporal fusion encoder using graph neural networks for fine-grained encrypted traffic classification. In *Proceedings of the ACM Web Conference 2023 (WWW)*, pages 2066–2075, 2023.

[Zheng *et al.*, 2020] Wenbo Zheng, Chao Gou, Lan Yan, and Shaocong Mo. Learning to classify: A flow-based relation network for encrypted traffic classification. In *Proceedings of The Web Conference 2020 (WWW)*, pages 13–22, 2020.