Divide and Conquer Boosting for Enhanced Traffic Safety Description and Analysis with Large Vision Language Model

Khai Trinh Xuan[1,6,○] Khoi Nguyen Nguyen[1,6,○] Bach Hoang Ngo[2,6,⋆] Vu Dinh Xuan[3,6,⋆]
Minh-Hung An[4,†] Quang-Vinh Dinh[5,†]
[1] Ho Chi Minh City University of Technology, VNU-HCM, Vietnam
[2] University of Science, VNU-HCM, Vietnam
[3] University of Information Technology, VNU-HCM, Vietnam
[4] FPT Telecom, Vietnam
[5] AI Lab - AI VIETNAM
[6] Vietnam National University Ho Chi Minh City
[○]First authors [⋆]Core authors [†]Senior authors

Abstract

The increasing complexity of traffic dynamics has underscored the necessity for advanced traffic safety description and analysis, challenging the efficacy of current methodologies in comprehensively understanding and predicting safety conditions from transportation videos. This paper addresses these challenges by analyzing specific segments crucial for precise traffic safety descriptions. Through this examination, we introduce an innovative preprocessing method named "segment extraction", facilitating the creation of a novel segment-based training dataset. Additionally, we present a practical two-stage training framework specifically tailored for this dataset. This framework is designed to produce accurate descriptions of traffic safety by incorporating the unique attributes of our segment-based training datasets. Leveraging these contributions, our method achieved a notable 2nd rank with a score of 32.8877 in the AI City Challenge Track2 test set: Traffic Safety Description and Analysis 2024. The source code for the proposed approaches is openly accessible at https://github.com/AIVIETNAMResearch/AI-CIty-2024-Track2

1. Introduction

Securing traffic safety by captioning and analyzing data requires the complex process of observing pedestrians' focus on streets and providing warnings about nearby dangers. This study area has recently gained significant recognition, driven by the rapidly evolving demand for smart city solutions and infrastructure [27]. The combination of traffic captioning and analysis is crucial for improving road safety [17, 30]. By using data from car safety cameras, this technology helps monitor how pedestrians behave on streets, al-

[appearance] [location] [environment] [attention] [action]

Caption pedestrian: The pedestrian, a male in his 30s, was standing still on a main road in an urban area. He was wearing a gray T-shirt and black short pants. The weather was cloudy with bright visibility, and the road surface was dry and level. The pedestrian's body was oriented in the same direction as the vehicle that approached him. He was positioned diagonally to the right, in front of the vehicle. Despite being unaware of the vehicle's presence, he closely watched the parked vehicle in his line of sight. The relative distance between the pedestrian and the vehicle was far. The road had a one-way traffic flow with three lanes and sidewalks on both sides. It was a weekday, and the pedestrian seemed to have no awareness of the usual traffic volume on this road.

Caption vehicle: The vehicle was moving at a constant speed of 10km/h. It was positioned behind a pedestrian and was quite far away from them. The vehicle had a clear view of the pedestrian. It was going straight ahead without any change in direction. The environment conditions indicated that the pedestrian was a male in his 30s with a height of 160 cm. He was wearing a gray T-shirt and black short pants. The event took place in an urban area on a weekday. The weather was cloudy but the brightness was bright. The road surface was dry and level, made of asphalt. The traffic volume was usual on the main road that had one-way traffic with three lanes. Sidewalks were present on both sides of the road.

Figure 1. Segments analysis of a sample from the WTS dataset. The segments include appearance, location, environment, attention, and action. Each segment is represented by a distinct color, as shown at the top of the figure. The underlined text represents parts of sentences that include information relevant to another segment. Nonetheless, we assign the sentence to the most relevant segment to meet certain criteria outlined in section 3.1.1.

lowing for quick responses to potential dangers. This research area has become increasingly important due to the growing demand for advanced infrastructure in smart cities. As cities become more connected and technologically advanced, there is a greater need for effective traffic management solutions [24, 26]. This paper explores the task of text-video traffic analysis, aiming to provide valuable insights for improving road safety in modern cities.

In text-video traffic analysis, interpreting the multidimensional context of videos, tracking objects across frames, and generating coherent text pose significant challenges. The work in [9] underscores the necessity of developing techniques like hierarchical training based on frame rates and dual attention mechanisms to enhance the alignment between textual content and video, thereby mitigating memory and time constraints. Compared to models converting images to text, models transforming video to text confront distinctive computational challenges, a scarcity of high-quality data, and ambiguities in video captioning.

Capturing the necessary information for the model is the primary challenge in captioning dense traffic videos. The model needs information about environmental factors such as weather, traffic density, road conditions, and pedestrian-centric variables such as appearance, attention levels, and corresponding actions. Figure 1 shows a sample from the Woven Traffic Safty dataset [12], demonstrating the complexity of the captions in dense traffic videos. Producing detailed and reliable informative captions that capture this information is challenging and crucial for improving traffic safety. This paper aims to address this challenge by following a divide-and-conquer strategy. We divide the complex traffic captions into smaller pieces called "segments" and develop a model to tackle each segment individually. Our contributions can be summarized as follows:

- Firstly, we conduct an in-depth analysis of specific segments crucial for accurate traffic safety descriptions and develop an innovative preprocessing method named "segments extraction" to construct a novel segment-based training dataset.
- Secondly, we propose a practical two-stage training framework tailored for this dataset, enhancing the generation of precise traffic safety descriptions.
- We evaluate our approach on the Track 2 test set: Traffic Safety Description and Analysis 2024 of the AI City Challenge [34]. Our method achieved a notable 2nd rank with a score of 32.8877, demonstrating the effectiveness of our contributions.

The paper is structured as follows: Section 2 reviews current traffic analysis methodologies. Section 3 presents our comprehensive method, integrating advanced semantic and contextual understanding techniques. Section 4 evaluates our framework's performance, emphasizing efficiency improvements. Section 5 concludes with implications for future traffic management systems and areas for further research.

2. Related Work

2.1. Image captioning

Image Captioning The evolution of image captioning has been marked by the transition from Recurrent Neural Networks (RNNs) to transformer-based generative approaches. Initially, LSTMs, as highlighted by Vinyals et al. [33], utilized global visual features to initialize the network's hidden state for caption generation. Building on this, Xu et al. [36] introduced an attention mechanism that aligned the model's focus with relevant image features for each predicted word. Further advancing this concept, Anderson et al. [1] developed a two-layer LSTM model that refines the focus on image features through an iterative attention process across layers, enhancing the specificity and relevance of generated captions. Building on the foundation established by RNN-based techniques, the image captioning domain has embraced the Transformer decoder module, introduced in [31], by employing a cross-attention mechanism in the caption generation workflow, as demonstrated in various studies [7, 8, 23]. This exploration is further enriched by the integration of vision-language models that have been pre-trained on extensive datasets of image-text pairs [13, 14, 18, 37, 38], highlighting a pivotal shift towards leveraging large-scale pre-trained models. Notably, Zhou et al. [38] introduced a model that merges the modalities of images and text into a unified Encoder-Decoder structure, paving the way for more integrated approaches. Zhang et al. [37] dedicated their efforts to augmenting visual features through an object detection module, enhancing the model's ability to interpret and caption images accurately. Subsequently, Li et al. [13] introduced an innovative vision-language framework utilizing an asymmetrical design coupled with cross-modal skip connections to optimize computational efficiency. The BLIP model [14] then tackled the challenge of data quality in pre-training sets by proposing a method to filter out noise, thereby improving data integrity and, consequently, the model's captioning proficiency.

2.2. Multimodal Large language model

Image LLMs A widely adopted strategy for enhancing image and video captioning models involves integrating a pre-trained large language model (LLM) for text generation conditioned on visual inputs. A key challenge is designing an effective adapter to connect the LLM's text generation with visual features. BLIP-2 [15] addresses this issue by introducing the Q-Former, which efficiently aligns visual tokens with the LLM's embedding space, integrating inputs without additional fine-tuning. Following this, Liu et al.

[22] proposed a versatile visual assistant framework that excels in various tasks, including image captioning, emphasizing the importance of selecting instruction-tuning datasets for optimal model performance. Building on the theme of advanced visual processing, Bai et al. [2] introduced a multimodal Large Language Model (LLM) designed to understand bounding box annotations. This capability is essential in the context of densely populated real-world automotive footage.

Video LLMs Drawing inspiration from the groundbreaking contributions of [22], Lin et al. [19] advanced dense video captioning by integrating video data with linguistic features and enhancing feature alignment. The MoE-LLava model [20], an extension of these principles, employs a Mixture of Experts (MoE) framework to boost model efficacy while maintaining a compact model size. Despite these advancements, a notable limitation of these approaches is their underperformance in grounding-based tasks, attributed to the absence of such functions in their pre-training data. This limitation is critical for captioning footage from vehicle-mounted cameras, where precisely anchoring textual descriptions to visual elements is crucial. In light of this, leveraging image-centric LLMs with object grounding capabilities might offer better outcomes for car camera captioning tasks than video-centric models. Image-focused LLMs' inherent object recognition and localization strengths could provide a more solid foundation for generating contextually relevant and precise captions in automotive applications. Building on this premise, we utilize the work of Qwen-VL [2], a robustly pre-trained vision large language model with strong grounding capabilities.

3. Proposed Method

This section thoroughly explains our proposed pipeline, as shown in Fig. 2. We begin by describing the data preprocessing approach employed to create our innovative segment-oriented training dataset. Following this, we explain our training and inference pipelines, designed to produce an in-depth traffic safety analysis.

3.1. Segment Extraction

The Woven Traffic Safety (WTS) dataset provides an in-depth spatial and temporal analysis dataset of pedestrian-centric traffic footage. The dataset requires captioning for pedestrian-centric and vehicle-centric features. As shown in Fig. 1, pedestrian-centric captioning requires generating text about specific segments from the segment set μ, including: 1) appearance: the pedestrian's appearance, age, and height, 2) environment: the surrounding environment around the pedestrian and the vehicle, 3) location: the pedestrian's location relative to the vehicle and 4) attention: the visibility and behavior of the pedestrian concerning the vehicle. In addition to appearance, environment, and location segments, vehicle-centric captioning requires segment action, describing the vehicle's movements. Based on the observations above, we present a novel data preprocessing methodology for precise segment extraction. This method consists of two main phases: pre-segment and dynamic prompt segment phase.

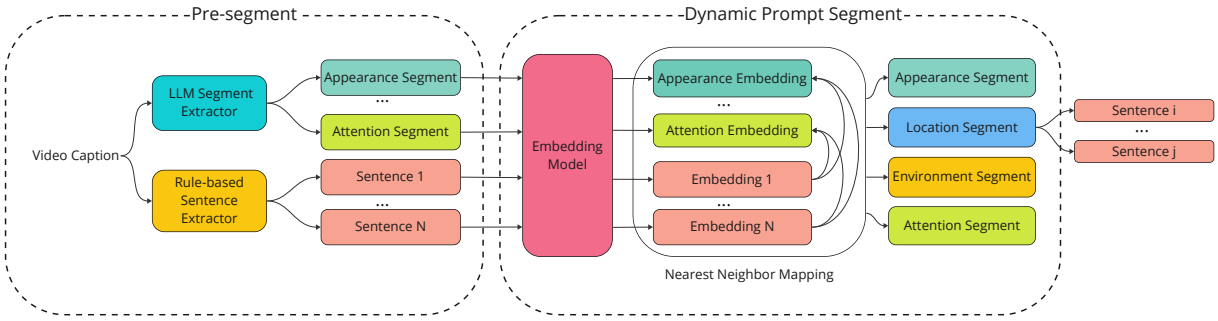
3.1.1 Phase 1: Pre-Segment

In the first phase, we utilize LLM to divide the description into specific segments. However, according to [4], LLM fails on complex tasks in zero-shot settings. Therefore, we employ the technique of few-shot prompting to guide LLM toward accurate segment division. The design of few-shot prompts is based on the methodology presented in [5]. From our observation, each sentence in the description may belong to many segments, combining many attributes. As shown in Fig. 1, our human-labeled segmentation strategy assigns sentences belonging to multiple segments to the most dominant segment. In this approach, we accept the trade-off between precision in segment extraction and the preservation of sentence structure. However, employing our human-annotated process directly for the design of few-shot prompts can lead to ambiguity for LLM when they encounter complex samples characterized by multiple segments.

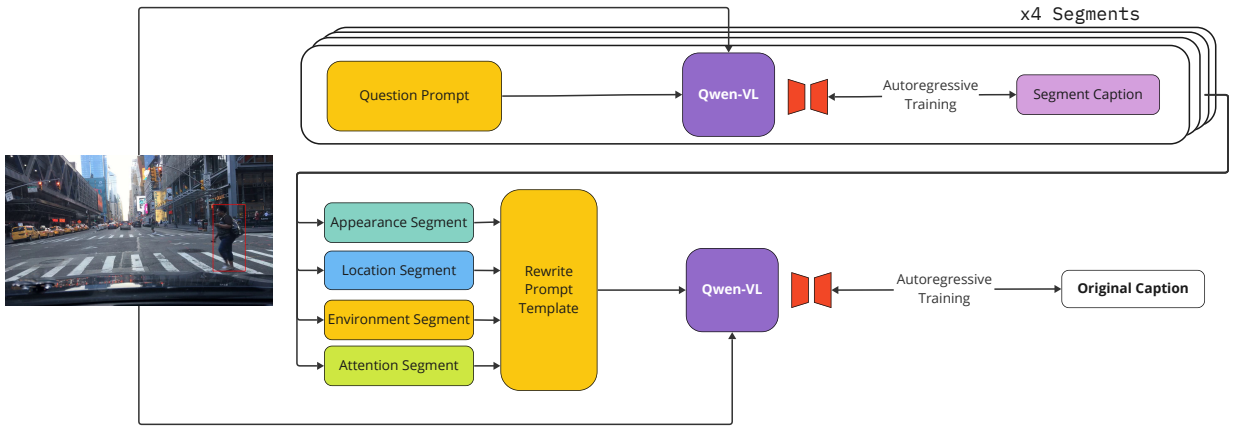
Consequently, in this phase, to guarantee accurate extraction of segments, we design our few-shot prompts to direct the LLM on retaining elements relevant to the required segment only. Our few-shot prompt technique is shown in Fig. 3, which ensures that the LLM receives clear guidance on adequately structuring its output and producing precise segment extraction. Figure 2a shows a complete overview of our pre-segment phase.

3.1.2 Phase 2: Dynamic Prompt Segment

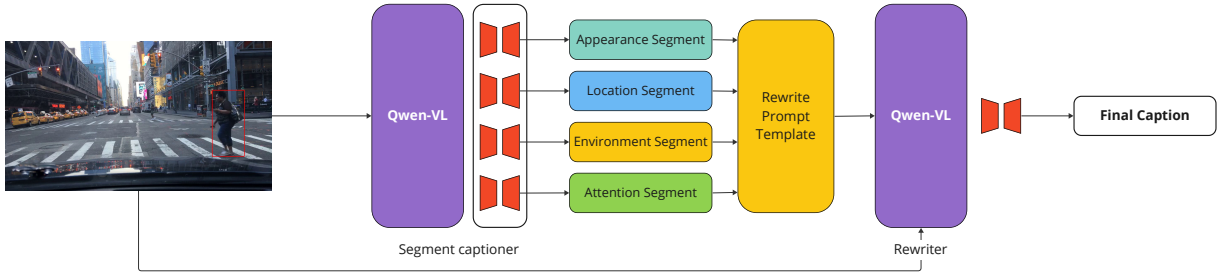
During the pre-segment phase, the output may combine multiple sentences from the original caption and exclude irrelevant information for precise segment extraction, as shown in Fig. 4. In addition, we notice some paraphrasing in the source sentences; the output segments occasionally lose some of the required elements. Furthermore, we also encounter some misclassification compared with our human-labeled segment. As a result, using this output directly for training could significantly reduce model performance, especially given the wide range of stylistic descriptions in the WTS dataset. This variability implies that training with such rephrased and incomplete segment-based data may considerably hinder the model's ability to produce various caption styles and its accuracy in capturing detailed information. Therefore, to overcome such a problem, we introduce a novel post-processing approach. This module functions as an automated voting system, classifying each



(a) Segment extraction pipeline. This preprocessing stage is divided into pre-segment and dynamic prompt segment phases. In the first phase, the video caption is divided into sentences and segments using the sentence and the LLM segment extractor. Leveraging embedding similarity, the nearest neighbor mapping function associates sentence i^{th} obtained from the first phase with the segment most closely related in the next phase. Here, we illustrate an example of sentence i being classified into a location segment.



(b) Training pipeline. Firstly, for each segment generated from the segment extraction phase of a training sample, a video frame and a question prompt designed specifically for that segment are used as input to the Qwen-VL model. The objective is to train the model to generate that specified segment accurately. Secondly, given a rewrite prompt, we train the Qwen-VL model to combine those generated segments to obtain the original description of the video scenario.



(c) Inference pipeline. Each segment is extracted from the video frame using the corresponding Qwen-VL's LoRA adapter for that segment. Then, all information is synthetic, using the rewrite Qwen-VL module to provide the final caption.

Figure 2. Overview of our proposed method for training a multi-model system to generate safety descriptions from traffic videos involving data preprocessing pipelines, training procedures, and inference mechanisms.

sentence from the original description into the segment that most closely aligns with it, thereby simulating our human-labeled segment procedure. Specifically, this process involves dividing the video description into a set of sentences α using the period ('.') as a delimiter with $\alpha = \{\chi_i, \forall i \in N\}$ where χ_i is the i^{th} sentence and N is the number of sentences in the caption. Subsequently, an embedding model E is employed to derive embedding vectors for each sentence in the original caption $\gamma_i = E(\chi_i), \forall i \in N$ where γ_i is the embedding vector of sentence i^{th} in the caption. Similarly, for each segment information extracted from the origin caption by LLM: $\delta_j = E(\epsilon_j), \forall j \in \mu$ where δ_j is

Fixed Few Shot Prompt

User:
Paragraph: The pedestrian, a female in her 40s, stands diagonally to the right and in front of the vehicle, with her body oriented perpendicular to the vehicle and to the left. She is closely watching the passing vehicle, unaware of its presence. As she prepares to cross the road, she wears a purple T-shirt on her upper body and black slacks on her lower body. Standing on the sidewalk of an urban area, this situation occurs on a bright and clear weekday, with dry and level asphalt as the road surface. The road is a main road with one-way traffic and one lane. Despite the usual traffic volume, the pedestrian remains unaware of the vehicle due to her line of sight being occupied by the passing vehicle.

Answer:

Assistant:
- appearance: The pedestrian, a female in her 40s, she wears a purple T-shirt on her upper body and black slacks on her lower body.
- location: She stands diagonally to the right and in front of the vehicle, with her body oriented perpendicular to the vehicle and to the left. Standing on the sidewalk of an urban area.
- environment: The road is a main road with one-way traffic and one lane. This situation occurs on a bright and clear weekday, with dry and level asphalt as the road surface
- attention: She is closely watching the passing vehicle, unaware of its presence. Despite the usual traffic volume, the pedestrian remains unaware of the vehicle due to her line of sight being occupied by the passing vehicle.

Figure 3. An example of our predefined few-shot prompt to direct the LLM for segment extraction.

the embedding of segment ϵ_j. Finally, we measure the similarity between the attribute extracted in the previous phase and each split sentence from the origin caption and categorize the split sentences into the most relevant segment. Let ν_i be the segment for the sentence i^{th} in the caption, ν_i is computed as: $\nu_i = \arg\max_j(\{S(\delta_j, \gamma_i)|\forall j \in \mu\})$, where S represents the cosine similarity function. The overall pipeline for our dynamic prompt segment phase is shown in Fig. 2a.

3.2. Training Pipeline

We can divide these segments into spatial and temporal attributes based on the segment set defined in the preceding section. Segments such as appearance, location, and environment are spatial attributes, implying that they do not require temporal information from consecutive video frames for accurate captioning. On the other hand, attention and action segments are temporal attributes, necessitating temporal information from successive frames. In the literature, high-performance video models that can perform temporal captioning, such as Video-LLaVA [19] and VideoChat2 [16], have limited capabilities in providing detailed captions for specific instances in videos. Conversely, recent advancements have showcased the impressive ability of image-based models to offer detailed instance captioning. As a result, we chose Qwen-VL [2], a robust vision-language model noted for its ability to understand and identify particular instances, as the foundation of our training pipeline. Figure 2b shows the division of our training pipeline into two steps. We first perform training Qwen-VL on each segment individually with question prompts designed specifically for each segment. However, piecing together all attributes in a predetermined order to form the final caption may not yield the best results. Certain segments could appear as several sentences spread throughout the caption rather than in one continuous section. Therefore, in the following training phase, Qwen-VL is trained to combine and refine the segments generated from the initial phase, resulting in a detailed and precise video description. We follow Qwen-VL's guidelines, which incorporate instruction-tuning techniques for fine-tuning and employing cross entropy as the loss function. This loss function is computed for one sample as follows:

$$\mathcal{L} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp(\hat{y}_{m,c})}{\sum_{j=1}^{C}\exp(\hat{y}_{m,j})}\{y_{m,c} \neq \textit{ignore}\}, \quad (1)$$

where C is the total number of classes in the model's corpus, M is the sentence length, $y_{m,c}$ is the m token labeled with class c, $\hat{y}_{m,j}$ is the corresponding model output logit for the m token of class j and *ignore* indicate the ignore token.

For the inference process, as shown in Fig. 2c, we first generate a set of segments for each video sample. Following this, we utilize the rewrite module to combine these segments into a cohesive description.

4. Experimentals

4.1. Track 2 Datasets

Our proposed model is trained and tested using the WTS dataset, which includes two subsets of data: internal and external. The dataset is divided into training, testing, and validation sets. The internal subset contains over 1,200 video events spanning over 130 unique traffic scenarios, captured from various viewpoints, including vehicle ego and fixed overhead cameras. The external subset encompasses around 4,800 publicly available pedestrian-centric traffic videos sourced from the BDD100K dataset. Each video sample has five phases: pre-recognition, recognition, judgment, action, and avoidance. For each phase, detailed captions are provided from the perspective of both pedestrians and cars, referred to as pedestrian captions and vehicle captions, respectively. In our current approach, we consider each stage a distinct sample and examine each perspective separately. We leave the exploration of the interconnections between stages and perspectives for future research.

4.2. Evaluation Metrics

To measure the accuracy of the candidate caption compared to the origin caption, the evaluation system for track 2 uses four metrics, including BLEU [25], Rouge-L [21], Meteor [3], CIDEr [32] then compute the average score as follows:

Original Caption			
The pedestrian, a male in his 20s with a height of 180 cm, was standing still on a main road in an urban area. He was dressed in a navy blue jacket and black slacks. It was a clear and bright day, with the road surface dry and the visibility good. The pedestrian's orientation was perpendicular to the vehicle and to the left, while his position was diagonally to the right in front of the vehicle. From his line of sight, it appeared that he was closely watching the passing vehicle. However, the pedestrian was unaware of the vehicle's presence. Despite being far away from the vehicle, he seemed to have a clear view of it. The road had two lanes and sidewalks on both sides. Overall, it was a normal weekday with usual traffic volume. The event captured the pedestrian's static state on a main road, unaware of the vehicle passing by.			
Segments	**Human-labeled**	**Phase 1: Pre-Segment**	**Phase 2: Dynamic Prompt Segment**
appearance	- The pedestrian a male in his 20s with a height of 180 cm, was standing still on a main road in an urban area. - He was dressed in a navy blue jacket and black slacks.	- The pedestrian, a male in his 20s with a height of 180 cm, **was dressed in a navy blue jacket and black slacks.**	- The pedestrian a male in his 20s with a height of 180 cm, **was standing still on a main road in an urban area.** **- He was dressed in a navy blue jacket and black slacks.**
environment	- It was a clear and bright day, with the road surface dry and the visibility good. - The road had two lanes and sidewalks on both sides. - Overall, it was a normal weekday with usual traffic volume.	- It was a clear and bright day, with the road surface dry and the visibility good. - The road had two lanes and sidewalks on both sides. ***- Information loss.***	- It was a clear and bright day, with the road surface dry and the visibility good. - The road had two lanes and sidewalks on both sides. **- Overall, it was a normal weekday with usual traffic volume.**
location	- The pedestrian's orientation was perpendicular to the vehicle and to the left, while his position was diagonally to the right in front of the vehicle.	**- His position was diagonally to the right in front of the vehicle, with an orientation perpendicular to the left of the vehicle.** **- He was standing still on a main road in an urban area.**	**- The pedestrian's orientation was perpendicular to the vehicle and to the left, while his position was diagonally to the right in front of the vehicle.**
attention	- From his line of sight, it appeared that he was closely watching the passing vehicle. - However, the pedestrian was unaware of the vehicle's presence. - Despite being far away from the vehicle, he seemed to have a clear view of it. - The event captured the pedestrian's static state on a main road, unaware of the vehicle passing by.	- From his line of sight, it appeared that he was closely watching the passing vehicle. **- Despite being far away from the vehicle and having a clear view of it, the pedestrian was unaware of its presence.** ***- Information loss.***	- From his line of sight, it appeared that he was closely watching the passing vehicle. **- However, the pedestrian was unaware of the vehicle's presence.** **- Despite being far away from the vehicle, he seemed to have a clear view of it.** **- The event captured the pedestrian's static state on a main road, unaware of the vehicle passing by.**

Figure 4. A case study for our segment extraction pipeline. In the diagram, red highlights identify text rephrased compared to human-labeled annotations. The term *information loss* marks positions where certain segments do not contain the required information. Blue highlights indicate sentences that have been misclassified into an incorrect segment. The use of green highlights showcases the role of phase 2 in structure-preserving and information loss filtering on the result from phase 1.

$$\text{Score} = \frac{\text{BLEU} + \text{Rouge-L} + \text{Meteor} + \text{CIDEr}}{4}. \quad (2)$$

4.2.1 BLEU

BLEU provides an automatic and quantitative estimate of the quality of the candidate sentence compared to the target sentence. The formula to calculate the BLEU score:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N_B} w_n \cdot \log\left(\frac{\text{clipped\_count}_n}{\text{count}_n}\right)\right), \quad (3)$$

where BP represents the brevity penalty, and it is calculated as follows:

$$\text{BP} = \begin{cases} 1 & \text{if } c_l > r_l \\ e^{(1-\frac{r_l}{c_l})} & \text{if } c_l \leq r_l \end{cases}, \quad (4)$$

where N_B is the maximum n-gram order, w_n is the weight associated with the n-gram order n, clipped_count_n is the total count of n-grams of the candidate sentence that appears in the target sentence but not exceeding the total count of that n-grams in the target sentence, count_n is the total count of n-grams in the candidate sentences, and c_l and r_l is the length of candidate sentence and the reference sentence. In the competition, the evaluation benchmark uses $N_B = 4$ and $w_n = 1/N_B$.

4.2.2 Rouge-L

The Rouge-L metric primarily evaluates the longest common subsequence (LCS) between the generated and reference description. The Rouge-L metric is formulated as:

$$\text{Rouge-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}, \quad (5)$$

where

$$\text{P}_{lcs} = \frac{LCS(X,Y)}{c_l}, \quad (6)$$

and

$$\text{R}_{lcs} = \frac{LCS(X,Y)}{r_l}, \quad (7)$$

where $LCS(X,Y)$ represents the length of the longest common subsequence between the reference string X of length r_l and the candidate string Y of length c_l. In the competition benchmark, β is set to 1.

4.2.3 Meteor

The Meteor metric employs a comprehensive approach that extends beyond simple unigram matching, considering various linguistic aspects such as unigrams and word order to

assess the quality of machine-generated translations compared to human-produced reference translations. The Meteor metric is calculated as follows:

$$\text{Meteor} = F_{\text{mean}} \cdot (1 - Penalty), \tag{8}$$

where

$$Penalty = 0.5(\frac{chunks}{u_m})^3, \tag{9}$$

$$F_{\text{mean}} = \frac{10PR}{R + 9P}, \tag{10}$$

$$P = \frac{u_m}{c_m}, \tag{11}$$

and

$$R = \frac{u_m}{r_m}, \tag{12}$$

where $chunks$ is the number of matching chunks (a chunk is a set of unigrams positioned next to each other in the reference and the candidate), u_m is the number of mapped unigrams between the reference and the candidate sentence, c_m and r_m are the unigram counts in the candidate and the reference sentence, respectively.

4.2.4 CIDEr

The CIDEr (Consensus-based Image Description Evaluation) metric quantifies the coherence between a generated caption and human-written reference captions for the same image. The definition of the CIDEr metric between candidate sentence c_i and a set of image descriptions $S_i = \{s_{i1}, ..., s_{im}\}$ is calculated as follows:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^{N_C} w_n \text{CIDEr}_n(c_i, S_i), \tag{13}$$

where

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}, \tag{14}$$

and

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right). \tag{15}$$

Each sentence is represented using a set of n-grams. ω_k is an n-gram of a set of one or more ordered words. The number of time an n-gram ω_k occurs in a reference sentences s_{ij} is denoted by $h_k(s_{ij})$ or $h_k(c_i)$ for the candidate sentence c_i. $g_k(s_{ij})$ is the TF-IDF [29] weighting for each n-gram w_k by Eq. 15. Ω is the vocabulary of all n-grams, and $|I|$ is the number of images in the set of all images in the dataset. $g^n(c_i)$ is a vector formed by $g_k(c_i)$ corresponding to all n-grams of length n and $\|g^n(c_i)\|$ is the magnitude of the vector $g^n(c_i)$. Similarly for $g^n(s_{ij})$. In the competition, uniform weight w_n for each n-gram is set to $\frac{1}{N_C}$, $N_C = 4$, and the number of references description per candidate sentences $m = 1$.

Segment Extraction Technique	Score
Pre Segment	57.5951
+ Dynamic Prompt Segment	**91.3527**

Table 1. Evaluation score between combined segments with original description on the training set of the external subset.

4.3. Implementation Details

Our study utilizes the Mistral 7B model [11] as the LLM segment extractor and incorporates the MiniLM [35] model for embedding extraction. For the external subset, our training parameters comprise a batch size of 6 with gradient accumulation set to 8. The initial learning rate is set at 1e-4, and the training procedure takes three epochs. For the internal subset, the checkpoint trained on an external dataset served as the initializing point for training. While retaining the batch size and gradient accumulation configuration from the external subset training, we reduced the learning rate to 1e-5 and extended the training duration to 5 epochs. Furthermore, we adopt different batch sizes to train the rewrite module: 4 for the internal subset and 6 for the external subset. In addition, we use the AdamW optimizer and Cosine Annealing learning rate scheduling technique in all of our studies, with a warm-up ratio of 0.01. During training, we also leverage Deepspeed ZeRO [28], Gradient Checkpointing [6], and LoRA [10] techniques to achieve parameter-efficient fine-tuning of large-scale pre-trained vision-language models. All of our experiments are conducted on a single A6000 GPU.

4.4. Ablation Studies

4.4.1 Impacts of each segment extraction process

This section explores how the segment-based data, derived from each stage of our preprocessing pipeline, impacts the evaluation metric. For notational clarity, we denote environment as "E", appearance as "A", attention as "AT", action as "AC", and location as "L". Figure 4 shows the outcomes of applying each segment extraction process on a training sample. It is observed that the segments from the initial phase are often the paraphrased versions of human-labeled segments, with notable information loss. However, after the second phase, the output is balanced between segment extraction accuracy and sentence structure preservation. Some information missing in the first phase is added during the second phase, resulting in a complete segment-based dataset. For a comprehensive evaluation, Table 1

Training method	Score
Single LoRA	34.4306
+ Per Segment Pedestrian LoRA	34.5050
+ Per Segment Vehicle LoRA	**34.7804**

Table 2. Evaluation score for each training method. The evaluation score for using a single adapter across all segments is presented in the first row. The second and third rows highlight the enhancement achieved by switching from a single adapter for all segments to individual adapters for each segment, applied sequentially to pedestrian and vehicle captions.

is constructed to benchmark the paraphrasability resulting from each stage of our preprocessing phase when applied to the external training data of the WTS dataset compared with the original caption. We use the metric defined in Eq. 2 to evaluate our experiments from this section onward. For evaluation purposes, we concatenate all segments in the order {A, E, L, AT} and {AC, L, A, E} for pedestrians and vehicles, respectively. Noticeably, after adopting the dynamic prompt segment phase, the score increases significantly by 60% in the mean score, indicating that a decrease in information loss and an improvement in the preservation of sentence structure heavily influenced the metric's outcome.

4.4.2 Single vs Multiple adapter

In this part, we examine how the training method affects the performance of our proposed framework. Initially, we fine-tuned the Qwen-VL model on our segment-based training dataset by creating dialogue data following [2] for instruction fine-tuning. This involves constructing dialogues that consist of a question-and-answer pair for each segment, one by one. However, based on our observations, each segment can be treated individually as its information can be inferred independently; thus, we attempt to train an adapter expert for each segment. Following the same combination sequence as the previous section, Table 2 reveals that training an individual adapter for each segment improves performance over training a single adapter for all segments by 0.3498 mean score, increasing from 34.4306 to 34.7804.

4.4.3 Rewrite Module

In this section, we first explore the impact of the combined sequence order on the evaluation score and conclude the need for our rewriting module. We perform experiments on several combination sequences of pedestrian caption as illustrated in Table 3. The findings reveal a significant impact of sequence order on the scores. Specifically, the sequence {A, AT, L, E}, outperformed the sequence {L, AT, E, A} by 0.059 in the external subset. However, it fails to perform as well as {L, AT, E, A} in the internal subset, 30.4965, compared to 30.7010, highlighting the unexpected influence of sequence arrangement on evaluation results.

Combined Method	External	Internal	Score
E + A + AT + L	34.2668	29.9008	32.0838
A + E + L + AT	34.7804	30.3989	32.5896
A + AT + L + E	**35.0753**	30.4965	32.7859
L + AT + E + A	35.0163	**30.7010**	**32.8586**
Rewrite	**36.0801**	**32.1105**	**34.0953**

Table 3. Performance scores for various sequence combinations when merging pedestrian and fixed vehicle segments to {AC, L, A, E}. The second and third columns show the scores for external and internal validation subsets, respectively, while the final column presents the average score across these subsets. The last row evaluates the performance of our rewrite module.

Rank	Team Name	Score
1	AliOpenTrek	33.4308
2	**AIO_ISC(Ours)**	**32.8877**
3	Lighthouse	32.3006
4	VAI	32.2778
5	Santa Claude	29.7838
6	UCF-SST-NLP	29.0084
7	Monitor	28.7485
8	X	27.7771
9	HCMUS_AGAIN	22.7371

Table 4. Leaderboard on the full test set of Track 2 in the AI City Challenge 2024. Our proposed method achieved 2nd rank with a score of 32.8877.

Subsequently, we introduce a rewrite module designed to function as a dynamic combiner. This module accepts a set of segments, determines the optimal sequence for combination, and dynamically refines the resulting sentences for each video sample. As indicated in Table 3, our rewrite module significantly surpasses the performance of the primary fixed combination approach by 1.2367 mean score, showcasing its effectiveness in enhancing the coherence of the combined caption.

Our proposed method achieved 2nd rank on the whole test set with a score of 32.8877, as shown in Table 4.

5. Conclusion

This paper has addressed the challenge posed in Track 2 of the AICITY Challenge 2024, which involves generating safety descriptions for pedestrians depicted in video footage. We introduce a robust methodology for tackling this problem by leveraging the capabilities of pre-trained vision language models and our novel data preprocessing approach. Additionally, we define the fundamental motivation driving our solution, describe its inherent limits, and outline potential areas for further research and advancement.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2017. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3, 5, 8

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[5] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171*, 2023. 3

[6] Jianwei Feng and Dong Huang. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11442, 2021. 7

[7] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[8] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[9] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 7

[11] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 7

[12] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, David C. Anastasiu, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. 2024. 2

[13] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 2

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 2

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023. 2

[16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023. 5

[17] Wei Li, Zhao wei Qu, Haiyu Song, Pengjie Wang, and Bo Xue. The traffic scene understanding and prediction based on image captioning. *IEEE Access*, 9:1420–1427, 2021. 1

[18] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 2020. 2

[19] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 3, 5

[20] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*, 2024. 3

[21] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[23] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2286–2293, 2021. 2

[24] Bach Hoang Ngo, Dat Thanh Nguyen, Nhat-Tuong Do-Tran, Phuc Pham Huy Thien, Minh-Hung An, Tuan-Ngoc Nguyen, Loi Nguyen Hoang, Vinh Dinh Nguyen, and Vinh Dinh. Comprehensive visual features and pseudo labeling for robust natural language-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5409–5418, 2023. 2

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 5

[26] Ehsan Qasemi and Alessandro Oltramari. Intelligent traffic monitoring with hybrid ai. *ArXiv*, abs/2209.00448, 2022. 2

[27] Ehsan Qasemi, Jonathan M Francis, and Alessandro Oltramari. Traffic-domain video question answering with automatic captioning. *ArXiv*, abs/2307.09636, 2023. 1

[28] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 7

[29] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 60(5):503–520, 2004. 7

[30] Parniya Seifi and Abdolah Chalechale. Traffic captioning: Deep learning-based method to understand and describe traffic images. *2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–6, 2022. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2014. 2

[34] Shuo Wang, David C. Anastasiu, Zheng Tang, Ming-Ching Chang, Yue Yao, Liang Zheng, Mohammed Shaiqur Rahman, Meenakshi S. Arya, Anuj Sharma, Pranamesh Chakraborty, Sanjita Prajapati, Quan Kong, Norimasa Kobori, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Ganzorig Batnasan, Fady Alnajjar, Ping-Yang Chen, Jun-Wei Hsieh, Xunlei Wu, Sameer Satish Pusegaonkar, Yizhou Wang, Sujit Biswas, and Rama Chellappa. The 8th AI City Challenge. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[35] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 7

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057, Lille, France, 2015. PMLR. 2

[37] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584, 2021. 2

[38] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13041–13049, 2020. 2