

# Supplementary: VRU-Accident: A Vision-Language Benchmark for Video Question Answering and Dense Captioning for Accident Scene Understanding

## A. Supplementary Organization

We organize the supplementary material as follows:

- Section B: VRU-Accident Details
- Section C: Qualitative Examples
- Section D: Prompts for VRU-Accident Curation
- Section E: Reproduction of Experiment
- Section F: Acknowledgements

## B. VRU-Accident Details

### B.1. Task-Specific Prompts

We present the input prompts used for MLLMs for each of the two tasks in VRU-Accident: VQA and Dense Captioning. The prompts are designed to elicit informative and structured responses from the models when paired with video inputs.

**Prompts for VQA task** For each question category in the VQA task, we provide a specific prompt. The prompts used are:

- **Weather and Light Condition:** What’s the weather and lighting? Choose the correct option (A, B, C, or D) without any explanations.
- **Traffic Environment:** Where did the accident happen? Choose the correct option (A, B, C, or D) without any explanations.
- **Road Configuration:** What type of road is shown? Choose the correct option (A, B, C, or D) without any explanations.
- **Accident Type:** What kind of accident occurred? Choose the correct option (A, B, C, or D) without any explanations.
- **Accident Cause:** Why did the accident happen? Choose the correct option (A, B, C, or D) without any explanations.
- **Accident Prevention Measure:** How could this accident be prevented? Choose the correct option (A, B, C, or D) without any explanations.

**Prompts for Dense Captioning task** To generate textual descriptions that match the level of detail and narrative style

of the ground truth dense captions in VRU-Accident, we used the following prompt for all MLLMs:

- Provide a detailed description of this accident video. Use clear and complete sentences with appropriate traffic and accident-related terminology. Include descriptions of weather conditions, road type, and vehicle or pedestrian appearance (such as clothing and posture). Mention vehicle speed, trajectory, and movements, as well as any changes in the pedestrian’s behavior. Focus on the dynamics of the collision, including vehicle approach, pedestrian movement, and final impact.

### B.2. Evaluation Metrics

In this section, we supplement explanation of evaluation metrics used for both VQA and Dense Captioning tasks in the main manuscript.

**VQA Evaluation Metrics** We evaluate the performance of models on the VQA task using standard classification accuracy per category. For each category  $j$ , we define the accuracy  $\text{Acc}_j$  as  $\text{Acc}_j = N_j^{\text{correct}} / N_j^{\text{total}}$  where  $N_j^{\text{correct}}$  denotes the number of correctly predicted samples, and  $N_j^{\text{total}}$  is the total number of samples for category  $j$ . *Note that*, unlike conventional classification tasks where a model selects from a fixed set of global class labels, each question  $q_j$  in our benchmark is associated with a dynamically constructed candidate set  $\hat{Y}_j$  that includes one correct answer and three counterfactual distractors. This makes the task more challenging, as models must perform fine-grained reasoning to distinguish the correct answer from contextually plausible but incorrect alternatives.

**Dense Captioning Evaluation Metrics** We select well-accepted and reliable metrics, including SPICE [1], ME-TEOR [6], COMET [11], and ROUGE scores [9], to quantitatively evaluate the generated descriptions from the VLMs.

First of all, the SPICE metric evaluates caption quality by comparing semantic content in the form of tuples extracted from scene graphs. Each caption is parsed into a set of tuples representing objects, attributes, and relations. The SPICE F1 score is computed as the harmonic mean of tuple-level precision and recall:

$$\text{SPICE}_{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (1)$$

where Precision and Recall are defined as:

$$\text{Precision} = \frac{|\mathcal{T}_{\text{gen}} \cap \mathcal{T}_{\text{ref}}|}{|\mathcal{T}_{\text{gen}}|}, \quad (2)$$

$$\text{Recall} = \frac{|\mathcal{T}_{\text{gen}} \cap \mathcal{T}_{\text{ref}}|}{|\mathcal{T}_{\text{ref}}|}, \quad (3)$$

with  $\mathcal{T}_{\text{gen}}$  and  $\mathcal{T}_{\text{ref}}$  denoting the sets of tuples from the generated and reference captions, respectively.

Secondly, METEOR captures both precision and recall of matching words, balancing linguistic precision and semantic recall. The METEOR score is calculated using the following formula:

$$\text{METEOR} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \alpha \times \text{Recall} + (1 - \alpha)} \quad (4)$$

Precision represents the proportion of words in the generated text that match the reference text, while Recall indicates the proportion of words in the reference text that are captured in the generated text. The parameter  $\alpha$  functions as a weighting factor, balancing linguistic precision and semantic recall to provide an adaptive evaluation of the generated text’s fidelity and coverage compared to the reference.

Thirdly, COMET is a neural-based metric that leverages pre-trained multilingual language models to predict human judgment scores for machine-generated text. Unlike surface-level metrics that rely solely on n-gram overlaps, COMET evaluates the semantic adequacy and fluency of the generated captions by comparing them against reference texts using contextual embeddings. Formally, COMET operates by encoding the source ( $S$ ), reference ( $R$ ), and hypothesis ( $H$ ) using a pre-trained model and passing them through a regression head to produce a quality score:

$$\text{COMET}(S, R, H) = f_{\theta}(\text{Enc}(S), \text{Enc}(R), \text{Enc}(H)) \quad (5)$$

where  $\text{Enc}(\cdot)$  denotes the encoder output from the language model and  $f_{\theta}$  represents the learned regression layer that maps the embeddings to a scalar score. COMET provides a more human-aligned assessment of caption quality, especially in capturing nuanced semantic differences that traditional metrics may overlook.

Finally, the ROUGE score is a set of metrics used to evaluate the summarization quality, specifically through precision (P), recall (R), and harmonic mean (F) scores based on n-gram overlaps. We report ROUGE-1, ROUGE-2, and ROUGE-L scores, corresponding to unigram overlap, bigram overlap, and longest common subsequence (LCS), respectively. For each variant, we compute precision, recall, and F1 as follows:

$$\text{Precision}_{\text{ROUGE-}n} = \frac{|\text{n-gram}_{\text{gen}} \cap \text{n-gram}_{\text{ref}}|}{|\text{n-gram}_{\text{gen}}|}, \quad (6)$$

$$\text{Recall}_{\text{ROUGE-}n} = \frac{|\text{n-gram}_{\text{gen}} \cap \text{n-gram}_{\text{ref}}|}{|\text{n-gram}_{\text{ref}}|}, \quad (7)$$

$$\text{F1}_{\text{ROUGE-}n} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (8)$$

where  $n \in \{1, 2\}$  for ROUGE-1 and ROUGE-2, and  $\text{n-gram}_{\text{gen}}$ ,  $\text{n-gram}_{\text{ref}}$  denote the sets of n-grams in the generated and reference captions, respectively.

For ROUGE-L, which is based on the longest common subsequence (LCS), the scores are defined as:

$$\text{Precision}_{\text{ROUGE-L}} = \frac{\text{LCS}(\text{gen}, \text{ref})}{|\text{gen}|}, \quad (9)$$

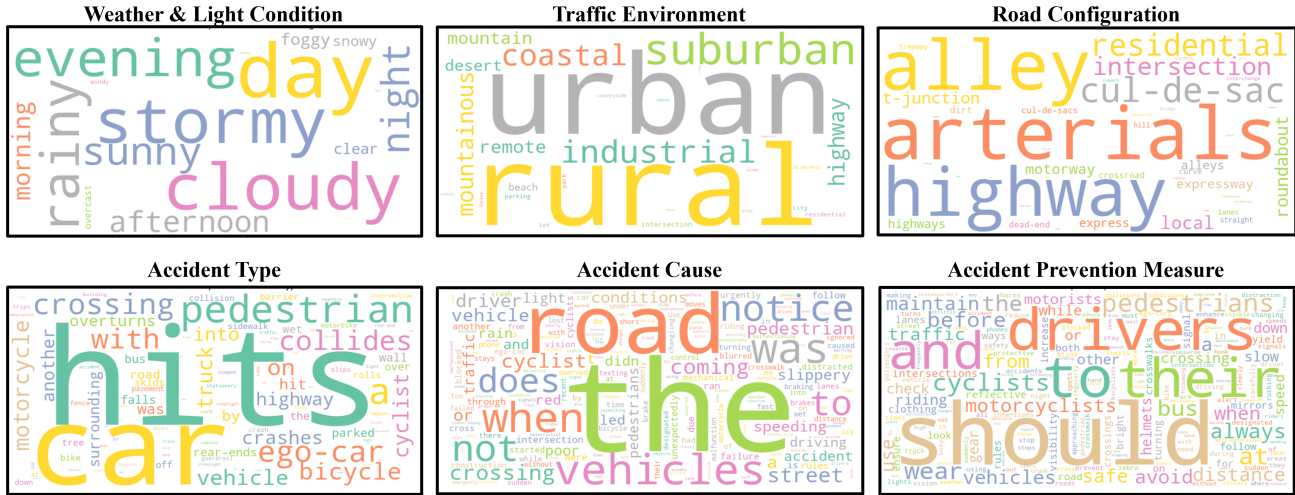
$$\text{Recall}_{\text{ROUGE-L}} = \frac{\text{LCS}(\text{gen}, \text{ref})}{|\text{ref}|}, \quad (10)$$

$$\text{F1}_{\text{ROUGE-L}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (11)$$

### B.3. Annotation Detail

Our VRU-Accident benchmark is designed to evaluate two core tasks: Video Question Answering (VQA) and Dense Captioning. The goal of the VQA task is to select the most appropriate answer from concise options—typically represented as short phrases a word, while the Dense Captioning task aims to generate detailed natural language descriptions that capture the full spatio-temporal context of a traffic accident. Given that real-world accident scenarios often involve complex and intertwined physical, environmental, and behavioral elements, it is inherently difficult to fully represent such events using concise options alone. Therefore, we provide annotation guidelines below that clarify how we determined the ground truth labels for each VQA category, while noting that richer, time-dependent contextual information is separately captured in the dense caption annotations.

**Weather & Light Condition** This category is annotated based on how the weather and lighting conditions plausibly contributed to the accident. For instance, if visibility is generally clear and there is no evidence of severe shadows or rain, the condition is annotated as a “sunny day” even if the sunlight appears diffused. Conversely, if the road surface appears wet or snow-covered—even in the absence of ongoing precipitation—we annotate the condition as “rainy” or “snowy” due to the increased risk of slipping and reduced friction. These decisions are made from a safety-critical perspective, prioritizing conditions that affect vehicle control or pedestrian stability. More detailed visual cues regarding surface wetness, lighting, or reflections are included in the dense caption descriptions.



**Traffic Environment** Accidents involving VRUs generally occur in either urban or rural settings. The environmental classification is relatively straightforward based on visual cues such as infrastructure density, type of roadside elements, and presence of intersections or buildings. In most cases, the classification is unambiguous.

**Road Configuration** Although many videos provide a clear view of the road structure, the moving dashcam perspective can sometimes obscure precise location categorization. For example, if a vehicle is approaching an intersection, the same video segment may contain both arterial roads and the intersection itself. In such cases, we annotate the configuration based on where the key event (e.g., collision or near-miss) occurs. If the impact takes place just before entering the intersection, the label is set to “arterial road”; if the event occurs within the crossroad, it is labeled as “intersection.” The sequential progression and transitions in road context are described in detail within the dense captions.

**Accident Type** While most accident types can be clearly and concisely described (e.g., “car hits pedestrian,” “bicycle falls”), there are borderline or ambiguous cases. For instance, when a dashcam-equipped vehicle rapidly approaches a pedestrian and the pedestrian narrowly avoids contact by sidestepping, it becomes unclear whether to classify the event as an actual collision or a near-miss. In such cases, we annotate based on the interaction and directionality of the entities involved. If a pedestrian avoids a moving ego-vehicle by inches, the event is still annotated as “ego-car hits a pedestrian” for the sake of interpretability and consistency. Similarly, if a cyclist accidented into a parked

vehicle, it is labeled as “car hits cyclist,” following common conventions seen in prior datasets such as [4, 5]. These nuances are captured more fully in the dense captions, where the dynamic unfolding of the scenario is described.

**Accident Cause** Causality in traffic accidents is often multi-factorial and temporally distributed. It is difficult to capture such complexity within a short answer choice. Therefore, our annotations focus on the most immediate and visually interpretable cause of the accident—such as driver inattention, sudden lane change, or pedestrian jaywalking. These labels are selected based on high-salience cues that directly precede the event. The dense caption annotations, however, elaborate on the sequence of road user actions, delays in reaction, or misjudgments over time that contributed to the accident.

**Accident Prevention Measure** This category is annotated based on what preventive action could have reasonably avoided the accident, considering the behavior of all involved road users. Rather than speculating on counterfactuals involving unrealistic interventions, we focus on plausible, context-aware measures that align with traffic norms and human capabilities. For example, if a vehicle accelerated despite a pedestrian entering the crosswalk, the prevention measure might be “driver should yield to pedestrian.” In another case, if a cyclist riding without a helmet crosses the road illegally during rainy conditions and collides with a vehicle, the core cause of the accident is the illegal crossing. Among multiple plausible prevention strategies—such as “cyclist should wear a helmet,” “cyclist should wear bright clothing,” “cyclist should reduce speed,” and “cyclist should use designated crosswalk”—we anno-

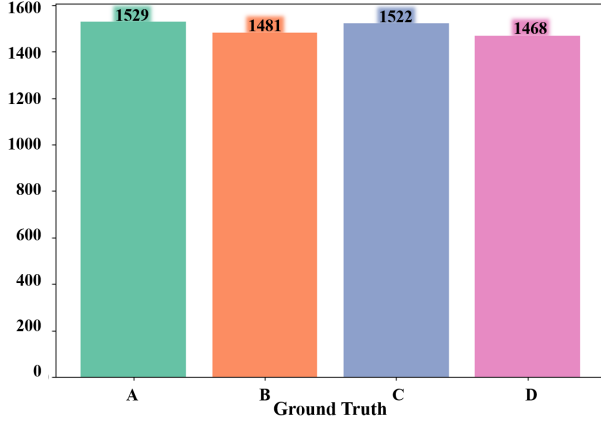


Figure 2. Distribution of ground truth of the VQA task in the VRU-Accident benchmark.

tate the most direct and effective preventive measure, in this case: “cyclist should use designated crosswalk.” This enables our benchmark to assess whether models can move beyond reliance on visually salient but potentially misleading cues, and instead demonstrate high-level understanding by identifying the true cause of the accident and articulating the most appropriate preventive action. In other words, it evaluates the model’s ability to reason about causality in safety-critical scenarios rather than simply describing visually correct but contextually irrelevant information.

#### B.4. VQA Statistics

**Word Distribution per Category** As illustrated in Figure 1, the answer options across six VQA categories exhibit a wide range of lexical diversity. Each category—ranging from *Weather & Light Condition* to *Accident Prevention Measure*—contains semantically rich and diverse phrases. This diversity increases the difficulty of answer selection, requiring high-level scene understanding and contextual reasoning beyond keyword matching. Therefore, the benchmark provides a suitable evaluation setting for probing the generalization capability and multimodal grounding ability of MLLMs.

**Balanced Ground Truth Distribution** Figure 2 presents the overall distribution of ground truth answers (A–D) in the VRU-Accident benchmark. The distribution remains relatively balanced across all four options, with less than 5% deviation between the most and least frequent labels. This balance mitigates bias toward any specific answer index and ensures reliable evaluation without prior answer frequency bias, making the benchmark statistically robust for multi-choice classification tasks.

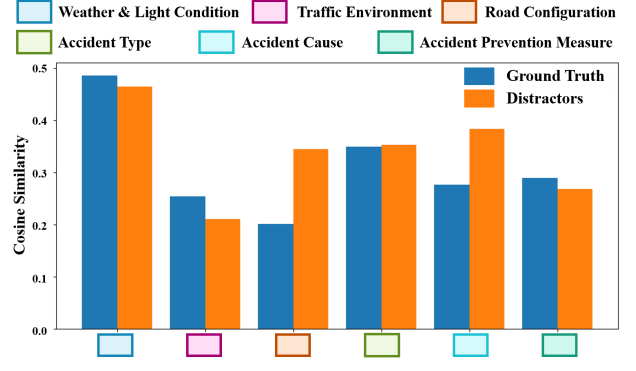


Figure 3. Comparison of average Q-A cosine similarity for ground truth answers and distractors across VQA categories in VRU-Accident benchmark.

#### Semantic Similarity Between Question and Answers

To assess the semantic relevance between questions and answer options in VRU-Accident, we compute the cosine similarity between their embeddings using the `all-MiniLM-L6-v2` model [16]. We denote the embedding model as a function  $F(\cdot)$ , which maps each sentence into a fixed-dimensional vector. For a given question  $q$  and its four answer options  $a_1, a_2, a_3, a_4$ , the cosine similarity between the question and each option is calculated as:

$$\text{sim}(q, a_k) = \frac{\langle F(q), F(a_k) \rangle}{\|F(q)\| \cdot \|F(a_k)\|}, \quad (12)$$

where  $k \in \{1, 2, 3, 4\}$ . For each option  $k$ , we evaluate the average similarity of category  $j$  between the question and the ground truth answer over all  $N$  samples. This average ground truth similarity is computed as:

$$\text{Sim}_j^{\text{GT}} = \frac{1}{N} \sum_{i=1}^N \text{sim}(q^{(i)}, a_{g^{(i)}}^{(i)}), \quad (13)$$

where  $a_{g^{(i)}}^{(i)}$  is the ground truth answer for the  $i$ -th sample. To assess the relevance of distractors, we take the three incorrect options for each sample, compute their similarity to the question, and average them across the dataset:

$$\text{Sim}_j^{\text{Dist}} = \frac{1}{3N} \sum_{i=1}^N \sum_{\substack{k=1 \\ k \neq g^{(i)}}}^4 \text{sim}(q^{(i)}, a_k^{(i)}). \quad (14)$$

Figure 3 shows that distractors in several categories exhibit comparable or even higher similarity to the question than the ground truth. This demonstrates that the distractor options are semantically plausible and not easily distinguishable without detailed reasoning. Therefore, this benchmark poses a challenging evaluation setting for assessing the semantic discrimination and contextual under-



standing capabilities of multimodal large language models (MLLMs).

### C. Qualitative Examples

To further analyze model performance, we present qualitative comparisons between the ground truth annotations in VRU-Accident and the responses generated by state-of-the-art MLLMs [2, 3, 7, 8, 10, 12–15, 17, 18] for both the VQA and dense captioning tasks. Figure 4 illustrates a representative set of cases, offering insight into the models’ capabilities and limitations in understanding complex accident scenarios.

We observe that many models correctly identify basic visual attributes (e.g., rainy day or urban setting). However, several limitations remain evident. For instance, while dense captions often include fluent descriptions of the scene layout and pedestrian appearances, models frequently hallucinate motion trajectories or misattribute collision responsibility, especially in complex scenarios involving occlusion or sudden movements.

These examples underscore the value of VRU-Accident as a challenging benchmark. It not only requires visual recognition but also demands high-level temporal and causal reasoning. By contrasting ground truth annotations with model predictions, our qualitative analysis reveals both the current capabilities and the remaining gaps in modern MLLMs’ ability to comprehend and explain real-world accidents involving vulnerable road users.

### D. Prompts for VRU-Accident Curation

**Prompts for VQA Curation** Figure 5 illustrates the detailed prompt used to curate the VQA benchmark in VRU-Accident. Human annotators first reviewed each event video and manually labeled the correct answer for each reasoning category. Based on this ground truth and the corresponding question, GPT-4o was employed to generate three plausible but incorrect options (distractors) that preserve semantic relevance but differ from the correct answer. The example in Figure 5 involves a rainy-day accident where a pedestrian suddenly emerges from between parked vehicles, resulting in a collision. For the **Accident Reason** category, the ground truth is “C. The ego-car’s vision is blocked or blurred, and there is no time to brake.” GPT-4o generates the following distractors: “A. Poor road conditions led to the accident,” “B. The brakes failed unexpectedly during the drive,” and “D. A cyclist suddenly crossed the road.” Option A, while weather-related, is incorrect because the road condition was not the direct cause. Option B is wrong as the brakes were functioning, and D is incorrect since the involved road user was a pedestrian, not a cyclist. These semantically plausible yet incorrect choices help assess whether MLLMs can go beyond surface-level cues and

demonstrate high-level causal reasoning.

**Prompts for Dense Caption Curation** Figure 6 shows the prompt used to curate dense accident descriptions. The prompt instructs the model to describe each accident video with detailed and coherent narratives, covering weather conditions, road type, vehicle and pedestrian appearances, vehicle dynamics (e.g., speed and trajectory), pedestrian behavior, spatial and temporal relationships among road users, and final collision outcomes. GPT-4o takes the input video and prompt rules to generate a single paragraph that mirrors the descriptive style and level of detail found in human-written annotations. This design ensures that the generated captions are not only visually grounded but also demonstrate temporal reasoning and contextual understanding of the accident dynamics. As a result, the VRU-Accident benchmark enables qualitative assessment of MLLMs’ ability to comprehend and articulate complex accident scenarios beyond mere visual recognition.

### E. Reproduction of Experiment

To encourage widespread use of our VRU-Accident benchmark and to support reproducibility of our experiments, we release all necessary resources used for all experiments in the main manuscript. This includes the inference scripts, model outputs, and evaluation results. Please visit our Github<sup>1</sup> and refer the **ReadMe** file for reproduction.

### F. Acknowledgements

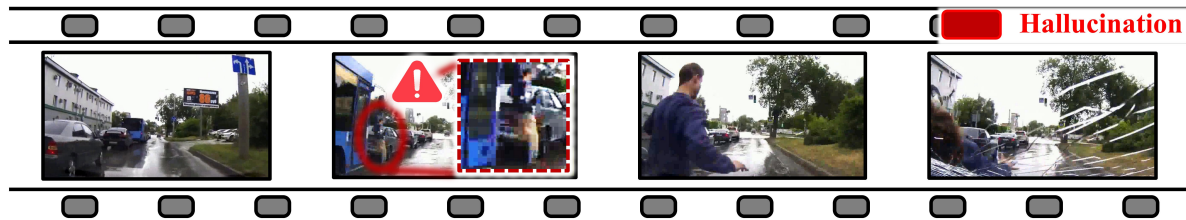
We would like to express sincere gratitude to Uibeom Chun (University of Central Florida) for his invaluable assistance with the conceptualization and experiments in this work. His contributions greatly supported the development of this research.

---

<sup>1</sup><https://github.com/Kimyounggun99/VRU-Accident>

## References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 1
- [2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5
- [3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 5
- [4] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. Cognitive accident prediction in driving scenes: A multimodality benchmark, 2023. 3
- [5] Jianwu Fang, Lei lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. Abductive ego-view accident video understanding for safe driving perception, 2024. 3
- [6] Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, page 228–231, USA, 2007. Association for Computational Linguistics. 1
- [7] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 5
- [8] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 5
- [9] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 1
- [10] Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding, 2025. 5
- [11] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. Comet: A neural framework for mt evaluation, 2020. 1
- [12] Abdelrahman Shaker, Muhammad Maaz, Hamid Rezatofighi, Salman Khan, and Fahad Shahbaz Khan. Mobile-videogpt: Fast and accurate video understanding language model. *arxiv*, 2025. 5
- [13] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024.
- [14] Qwen Team. Qwen2.5-vl, 2025.
- [15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5
- [16] Chen Yin and Zixuan Zhang. A study of sentence similarity based on the all-minilm-l6-v2 model with “same semantics, different structure” after fine tuning. In *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pages 677–684. Atlantis Press, 2024. 4
- [17] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 5
- [18] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingdong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 5



VRU-Accident  
Ground Truth

- ☒ A. rainy day
- ☒ D. urban
- ☒ A. Arterials
- ☒ D. ego-car hits a crossing pedestrian
- ☒ C. The ego-car's vision is blocked or blurred, and there is no time to brake
- ☒ C. Ego-cars should slow down or honk their horns when they stop at intersections or trunk roads where their vision is blocked to prevent other vehicles or pedestrians from rushing out suddenly

The incident occurs on a wet, two-way urban road under overcast skies, with parked vehicles and patches of greenery lining both sides. The road surface is slick and reflective due to recent rainfall, and the dark-colored vehicle gradually accelerates from approximately 33 km/h to 42 km/h while following queued traffic. Just before the collision, a pedestrian appears partially in the left edge of the frame and begins to cross the road unexpectedly from the left curb to the right. The pedestrian does not appear to check for oncoming traffic, while the vehicle continues straight in the leftmost lane, closing the gap rapidly. Without significant deceleration, the vehicle strikes the pedestrian head-on at around 42 km/h. The impact results in severe damage to the windshield, including both radial and concentric cracks, indicating a high-energy collision. This sequence highlights how the pedestrian's crossing attempt transitions into a collision within a very brief time window, with no evident evasive action by the driver.



LLaVA-  
OneVision(0.5B)

- ☒ A. Rainy day
- ☒ D. Urban
- ☒ B. Residential
- ☒ D. ego-car hits a crossing pedestrian
- ☒ C. The ego-car's vision is blocked or blurred, and there is no time to brake
- ☒ D. Delivery vans should maintain a safe distance from intersections to avoid sudden stops that could surprise other drivers.

The video captures a moment of a car accident on a wet road. The scene is set in a residential area with buildings on one side and a road on the other. The sky is overcast, and the road is slick with rain, reflecting the surrounding environment. The car in the foreground is a blue bus, and it is approaching a pedestrian crossing. The pedestrian, dressed in dark clothing, is crossing the road, and the car is making a sharp turn to avoid hitting the pedestrian. The pedestrian appears to be in a hurry, as indicated by their posture and the fact that they are crossing the road. (The car is making ... [repeating] )



Mobile-  
VideoGPT(0.5B)

- ☒ A. Rainy day
- ☒ D. Urban
- ☒ C. Highway
- ☒ A. Car collides with another vehicle
- ☒ A. Poor road conditions led to the accident
- ☒ C. Ego-cars should slow down or honk their horns when they stop at intersections or trunk roads where their vision is blocked to prevent other vehicles or pedestrians from rushing out suddenly

The blue bus was seen driving on the road with a sign that read 20% OFF. The driver of the bus then turned to the right and passed by a white car, which is parked in front of the bus.



InternVL2.5  
(1B)

- ☒ C. Cloudy evening
- ☒ D. Urban
- ☒ B. Residential
- ☒ B. Pedestrian was hit by surrounding vehicle
- ☒ C. The ego-car's vision is blocked or blurred, and there is no time to brake
- ☒ A. Bus drivers should ensure they signal well in advance before making any turns to alert other road users

The video captures an interaction between a vehicle and a pedestrian at a wet street intersection, characterized by a sense of urgency and motion. Initially, the scene is shot from a vehicle's dashboard camera, showing a wet pavement and adjacent buildings. The weather is overcast, with grey skies contributing to a somber mood. The scene then shifts to show pedestrians, with one crossing the street while being caught by the moving vehicle. The speedometer reads 38 km/h; this implies a moderate collision. The vehicle successfully avoids a potential hit before the pedestrian stumbles and tumbles over the windshield due to the speed of the vehicle chasing the pedestrian, eventually losing visibility and moving the camera away from the scene. The windshield is cracked, suggesting the force of impact.



Mobile-  
VideoGPT(1.5B)

- ✓ A. Rainy day
- ✓ D. Urban
- ✗ B. Residential
- ✓ D. ego-car hits a crossing pedestrian
- ✓ C. The ego-car's vision is blocked or blurred, and there is no time to brake
- ✓ C. Ego-cars should slow down or honk their horns when they stop at intersections or trunk roads where their vision is blocked to prevent other vehicles or pedestrians from rushing out suddenly

The video captures a sequence of events unfolding on an overcast day, starting with a view from inside a vehicle moving through a rainy urban environment. Initially, the scene is set with raindrops visible on the windshield and various vehicles including a blue bus ahead, parked cars, and a white van in the background. The road appears wet, suggesting recent or ongoing rainfall. As the video progresses, there's minimal change initially; the rain continues to fall steadily, and the vehicles maintain their positions relative to each other, indicating slow movement or a stop at a traffic light. A new element introduced is a signboard displaying "80" which could be interpreted as speed limits or directions, but its significance remains unclear due to the lack of context. Further into the video, the camera angle shifts slightly, revealing more details such as a pedestrian crossing the street and a red car entering the frame, hinting at possible stops or changes in traffic signals. Despite these slight movements, the overall scene remains largely unchanged except



InternVL3 (2B)

- ✓ A. Rainy day
- ✓ D. Urban
- ✗ B. Residential
- ✓ D. ego-car hits a crossing pedestrian
- ✓ C. The ego-car's vision is blocked or blurred, and there is no time to brake
- ✗ B. Motorcyclists should wear reflective vests at night to enhance their visibility to other vehicles on the road.

The video shows a white car moving from the right side of the frame and coming up towards a blue trolley bus. The car collides with the bus, causing the glass to become shattered. The blue trolley bus crosses from the left side of the frame towards the right and comes to a stop. A person wearing a blue jacket and beige pants crosses the road from right to left. The camera is the driver's point of view. The road has green trees on the left; and a pole and grass on the right. Vehicles are driving on the road, and the weather is sunny, and it is raining. The speed is km per hour in the first frame, and the speed is km per hour at the end of the video. The license plate is seen on the first car and on the bus.



Qwen2.5-VL(3B)

- ✓ A. Rainy day
- ✓ D. Urban
- ✗ B. Residential
- ✓ D. ego-car hits a crossing pedestrian
- ✗ D. A cyclist suddenly crossed the road
- ✓ C. Ego-cars should slow down or honk their horns when they stop at intersections or trunk roads where their vision is blocked to prevent other vehicles or pedestrians from rushing out suddenly

The video captures a sequence of events involving a bus and a pedestrian on a wet urban street. The scene is set during daylight hours, with overcast skies suggesting recent rain. The road appears to be a two-lane street with moderate traffic, including several cars and a blue bus. Initially, the bus is seen approaching from the left side of the frame, moving at a moderate speed indicated by the speedometer reading 35 km/h. The bus is a standard city bus, painted in blue and white, with visible advertisements on its side. As it approaches the pedestrian crossing, it slows down slightly, maintaining a speed of around 40 km/h. A pedestrian, dressed in casual attire, is seen crossing the street from the right side of the frame towards the left. The pedestrian appears to be in motion, possibly walking briskly, given the slight blur of their figure. The sidewalk is lined with trees and parked cars, contributing to a typical urban environment. As the bus continues its



Video-XL-Pro(3B)

- ✓ A. Rainy day
- ✓ D. Urban
- ✗ B. Residential
- ✓ D. ego-car hits a crossing pedestrian
- ✓ C. The ego-car's vision is blocked or blurred, and there is no time to brake
- ✗ A. Bus drivers should ensure they signal well in advance before making any turns to alert other road users

The video captures a sequence of events on a rainy day in an urban setting, focusing on a street intersection. Initially, the scene is set with a wet road reflecting the overcast sky, and several vehicles are parked along the curb, including a blue bus and a black car. The background features buildings with green accents and a sign indicating "Park City" and a speed limit of 80 km/h. The timestamp reads "2019/07/08 03:28:25," and the camera's position suggests it is mounted inside a vehicle. As the video progresses, the blue bus begins to move forward, while the black car remains stationary. The bus approaches the intersection, and a pedestrian wearing dark clothing is seen crossing the street from right to left. The pedestrian appears to be in motion, possibly walking or running across the road. The camera angle shifts slightly to follow the bus as it moves closer to the intersection, capturing the pedestrian's movement more





InternVL2.5 (4B)

- ✓ A. Rainy day ✓ D. Urban ✗ B. Residential ✓ D. ego-car hits a crossing pedestrian  
✓ C. The ego-car's vision is blocked or blurred, and there is no time to brake  
✗ A. Bus drivers should ensure they signal well in advance before making any turns to alert other road users

The video presents an indoor shooting environment, starting with multiple gunshots and showing smoke emanating from various firearm calibers. It then transitions to an outdoor urban scene. The video begins indoors with smoke visible, possibly from firearms, transitioning into an outdoor scene on a rainy day. The road appears wet, with reflections on the pavement. The surroundings include buildings and greenery along the side of the road. A person in blue jacket and jeans is observed crossing the road in front of a blue bus. Another pedestrian, wearing dark pants and brown shoes, is also seen crossing the road. Traffic conditions are observed on a multi-lane road typical of cities, with vehicles traveling in various directions. The pedestrian then moves further across the road near parked vehicles in a busy urban setting. The footage captures a moment of a pedestrian crossing in the roadway as vehicles pass by.



Video-XL2 (7B)

- ✓ A. Rainy day ✓ D. Urban ✗ B. Residential ✓ D. ego-car hits a crossing pedestrian  
✗ D. A cyclist suddenly crossed the road  
✗ A. Bus drivers should ensure they signal well in advance before making any turns to alert other road users

The video clip shows a first-person perspective from inside a vehicle, likely an ambulance given the title reference to 'CMC' (ambulance). The timestamp indicates it's early morning. The environment is urban with wet roads suggesting recent rain. A blue bus and various cars are parked along the street. A pedestrian in dark clothing suddenly crosses the road from right to left, seemingly unaware of oncoming traffic. The vehicle approaches at speed, indicated by the timestamp and speedometer reading 40 km/h. As the vehicle gets closer, the pedestrian is hit, causing a loud impact sound. The camera shakes violently, and the windshield cracks, obscuring the view momentarily. The pedestrian falls to the ground, and the vehicle comes to a stop. The colors are muted due to overcast weather, with grey dominating the scene.



LLaVA-Next-Video (7B)

- ✓ A. Rainy day ✓ D. Urban ✓ A. Arterials ✗ A. Car collides with another vehicle  
✗ A. Poor road conditions led to the accident  
✗ A. Bus drivers should ensure they signal well in advance before making any turns to alert other road users

The video depicts a car crash that occurred in a city street on a rainy day. The scene is captured from the perspective of a car's dashboard camera, showing a blue car driving down the street. The car is moving at a moderate speed, and the driver is wearing a black jacket and appears to be focused on the road ahead. As the car approaches a pedestrian crossing, the pedestrian suddenly steps onto the road, seemingly without looking for oncoming traffic. The driver of the blue car brakes abruptly to avoid a collision, but the pedestrian does not move out of the way in time, resulting in a head-on collision. The impact is severe, causing the pedestrian to fall onto the road. The car's trajectory is altered by the collision, and the driver's attention is drawn to the pedestrian's sudden appearance. The weather conditions are rainy, which may



Qwen2-VL(7B)

- ✓ A. Rainy day ✓ D. Urban ✗ B. Residential ✓ D. ego-car hits a crossing pedestrian  
✓ C. The ego-car's vision is blocked or blurred, and there is no time to brake  
✗ A. Bus drivers should ensure they signal well in advance before making any turns to alert other road users

The video depicts a traffic crash on a wet road, likely during or after a rainstorm. The weather conditions appear to be rainy, as evidenced by the wet road surface and the presence of rain streaks on the windshield. The road is a multi-lane street with vehicles parked on both sides, indicating it might be a suburban or urban area. A pedestrian is seen crossing the street from the right side to the left side. The pedestrian is wearing casual clothing, including a light-colored shirt and dark pants. The pedestrian's posture suggests they are in motion, likely walking or running across the street. A vehicle, possibly a car or a bus, is approaching the pedestrian from the right side of the street. The vehicle's speed appears to be moderate, as it is not overtaking other vehicles. The vehicle's trajectory is straight, and it is moving towards the pedestrian. As the vehicle approaches the pedestrian, the pedestrian suddenly changes direction and runs towards the left side of the street. This sudden change

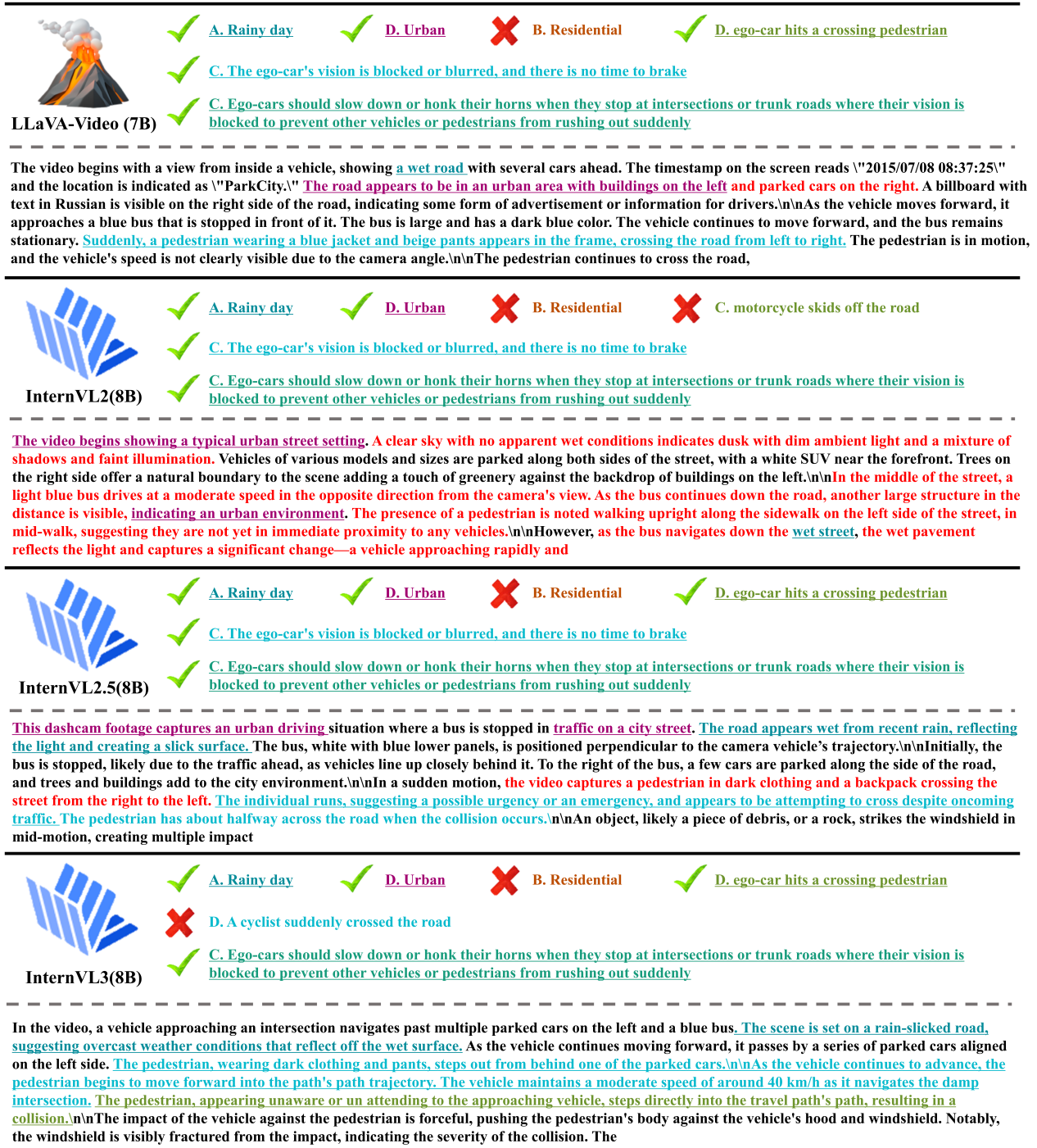


Figure 4. Qualitative comparison of VQA and dense captioning outputs from MLLMs on the VRU-Accident benchmark. Each example shows the ground truth annotation (top row) and the responses from models.





Figure 5. Prompt used for curating VRU-Accident VQA benchmark, including category-specific rules and a generated multiple-choice question with one correct answer and three semantically plausible distractors.

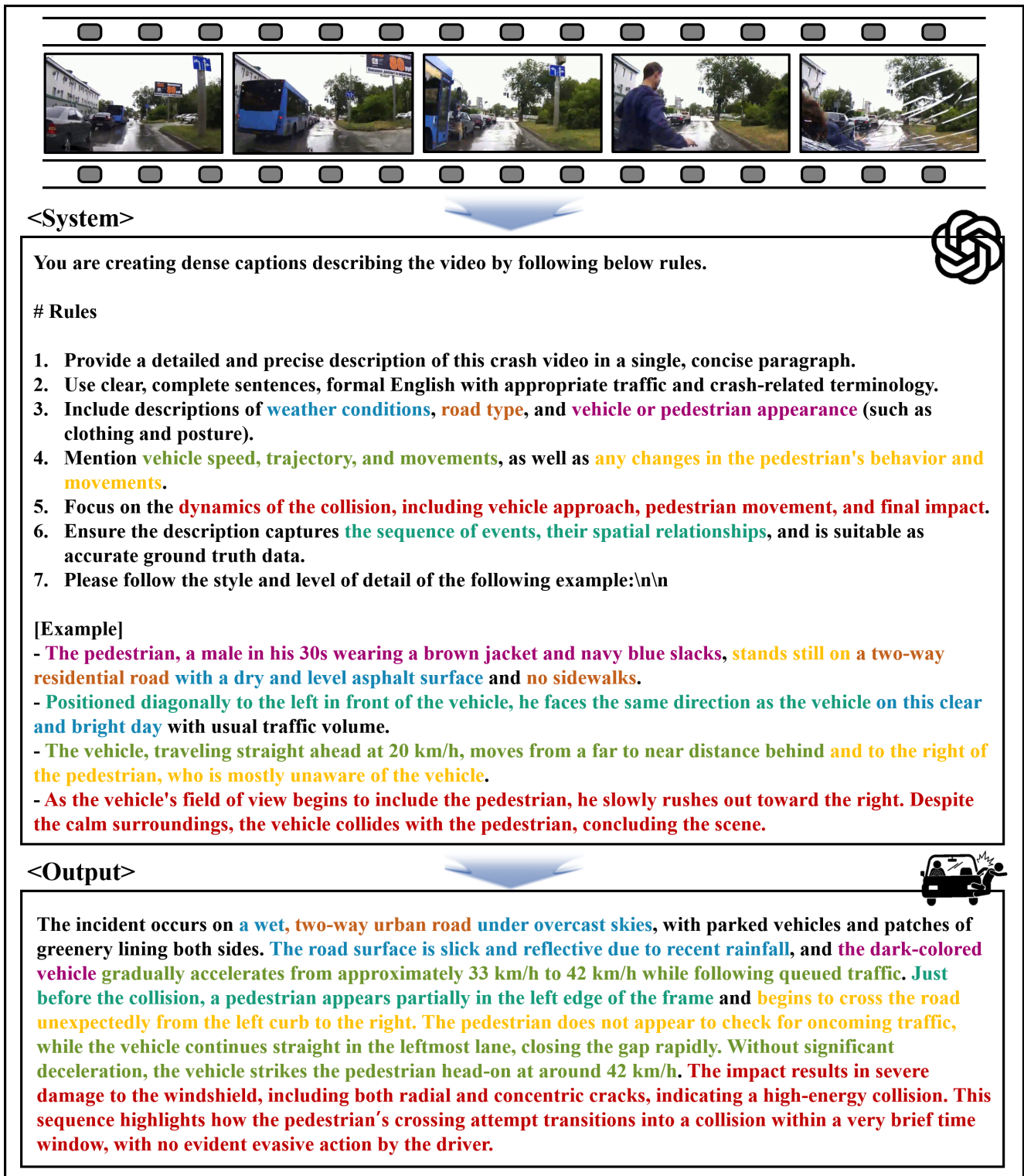


Figure 6. Prompt used for curating dense caption annotations in VRU-Accident. The prompt guides the model to describe accident videos with detailed references to weather, road configuration, agent appearance, kinematic features, and collision dynamics.