

A ADDITIONAL DETAILS ON DATA COLLECTION AND STATISTICS

A.1 DATA COLLECTION AND ANNOTATION

Data Source. We collect 106K media data, including 51,544 video clips and 54,767 images sourced directly or extracted from the video clips. In detail, as shown in Table 1, our dataset comprises 13,536 video clips and 34,968 images from 9 open-source traffic benchmarks, and 38,008 video clips and 19,799 images from social media platforms.

Table 1: The detailed distribution of our TAU-106K dataset from different data sources.

Data Source	# Video	# Image
DoTA (Yao et al., 2022)	4,672	9,502
CCD (Bao et al., 2020)	4,464	6,010
DADA (Fang et al., 2021)	1,923	3,799
DashCam (Chan et al., 2017)	1,727	2,348
TAD-1 (Lv et al., 2021)	352	628
TAD-benchmark (Xu et al., 2022)	208	546
Drive-Anomaly106 (Zhu et al., 2019)	105	207
RetroTrucks (Haresh et al., 2020)	56	348
TrafficS (Ghahremannezhad et al., 2022)	29	58
SUTD-TrafficQA (Xu et al., 2021)	—	9,674
CADP (Shah et al., 2018)	—	914
TaskFix (Juan et al., 2021)	—	713
YouTubeCrash (Kim et al., 2019)	—	221
<i>Youtube</i> (Social Media Platform)	29,364	12,345
<i>Bilibili</i> (Social Media Platform)	7,577	6,476
<i>TikTok</i> (Social Media Platform)	1,067	978

Data Annotation Process. We collaborate with a professional data annotation team to conduct the annotation of all traffic video and image data. The overall annotation process is divided into two parts: video-based annotation and image-based annotation. Each part is separated into annotation and verification phases carried out by different annotators. The annotators in the verification phases are tasked with verifying the quality of each data item as “satisfactory” or “unsatisfactory”. Unsatisfactory items were sent back to the annotation pipeline for refinement.

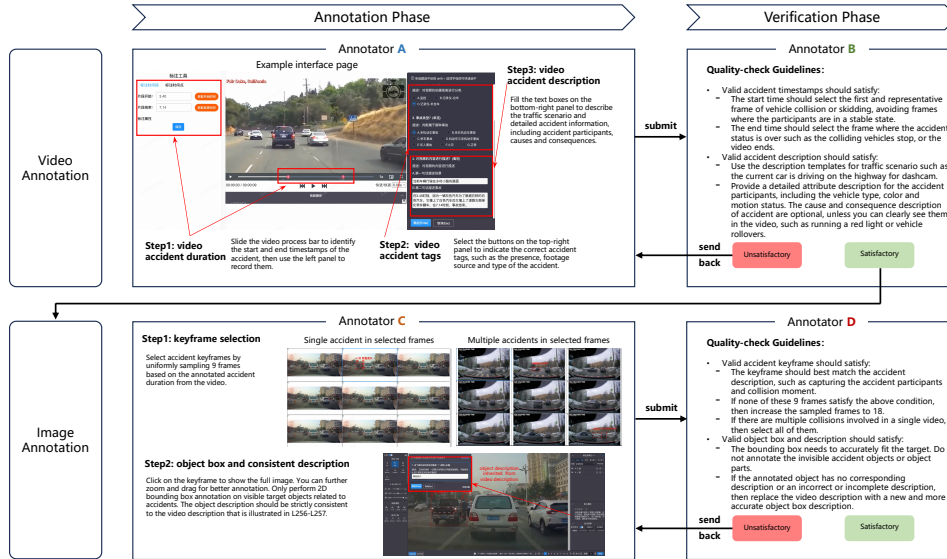


Figure 1: The annotation process and quality control of the TAU-106K dataset.

We use an internal annotation tool to enable interactive use with our annotators and a diagram of the annotation protocol used in our data engine, which is illustrated in Figure 1. Specifically, annotator A in the video annotation phase adopts Steps 1, 2 and 3 to provide the timestamps, semantic tags and detailed description for accident video, and another annotator B focuses on quality verification.

After the video annotation, the fine video data is sent to the image annotation phase, annotator C utilizes Steps 1 and 2 to perform keyframe selection and accident-related object annotation, and then annotator D conducts the image-level quality verification. We employed a team of 50 experienced annotators and all of them followed the same annotation guidelines presented in both video and image verification phases. According to our annotation workflow, each data item involved at least four different annotators to uphold a high standard for annotation. Moreover, benefiting from our proposed video-to-image annotation pipeline, the image annotators and verifiers double-check the annotations from the video phase, ensuring the consistency and accuracy of the annotations across different modalities, which is also label-efficient and cost-effective. We present some annotation samples in Figure 2.

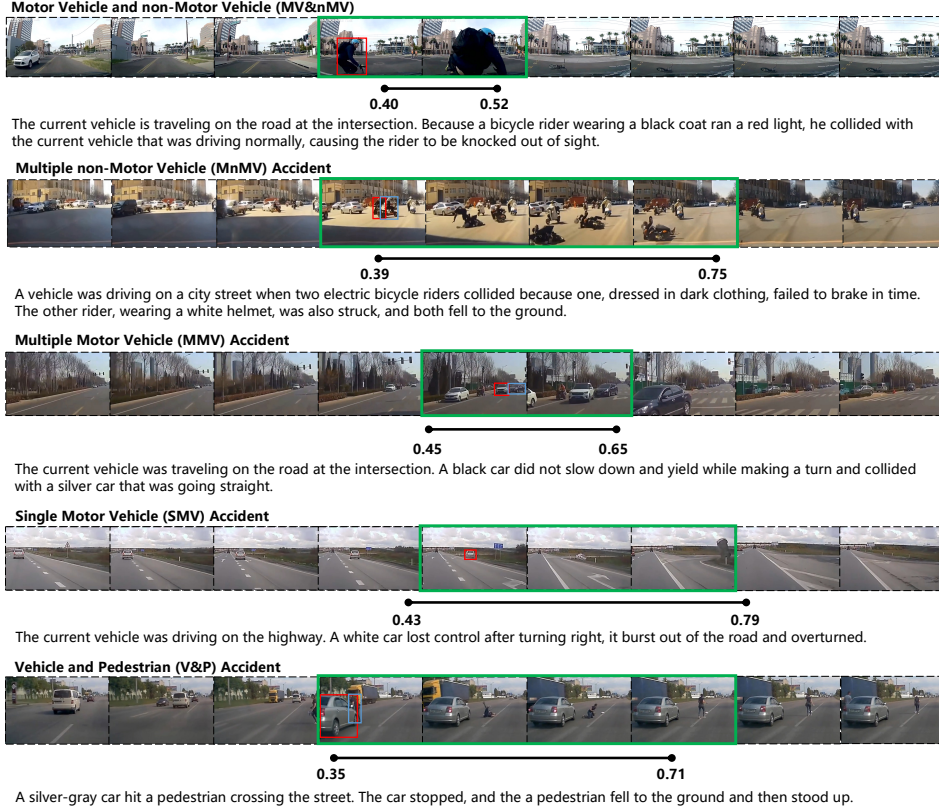


Figure 2: The video annotation examples of different types of accidents.

A.2 DATA STATISTICS

Data Categories. Our dataset covers various categories of traffic scenarios, objects, and accidents:

- *Traffic Scenarios*: urban streets (49%), intersections (19%), country roads (17%), highways (12%), and other traffic scenes (3%).
- *Objects*: cars (58%), trucks (12%), electric bikes (11%), pedestrians (5%), vans (3%), bicycles (3%), buses (2%), guardrails (2%), motorcycles (2%), and other objects (2%).
- *Accident Categories*: multi-motor-vehicle accidents (59%), motor-vehicle & non-motor-vehicle accidents (18%), single-motor-vehicle accidents (17%), vehicle & pedestrian accidents (4%), and multi-non-motor-vehicle accidents (2%).

Video Duration Distribution. As shown in Figure 3 (a), the video duration distribution of the accident-related video clips is visualized. The video clips collected from previous benchmarks or cropped from the raw crawled videos are relatively short, with the majority of the video clips lasting less than 20 seconds. Rarely, some video clips exceed 50 seconds, with the longest video clip lasting 12 minutes. For better visual presentation, we restrict the x-axis to 50 seconds, which covers the majority of the video clips.

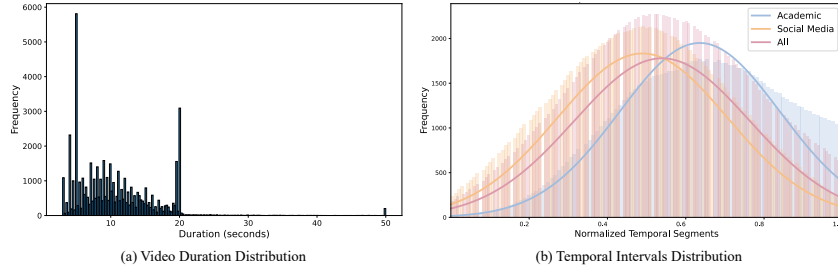


Figure 3: (a) Video duration distribution of the accident-related video clips; (b) Temporal intervals distribution of the accident events in the video clips. “Frequency” refers to the number of occurrences of videos with a certain length in the TAU-106K dataset.

Temporal Localization Distribution. To statistically examine the temporal distribution of accident events in the video clips, we analyze and visualize the annotated temporal intervals of these events, as illustrated in Figure 3 (b). The temporal intervals corresponding to accident events in existing benchmarks tend to exhibit a relatively concentrated distribution, with the majority of accidents occurring towards the later portions of the video clips. In contrast, the temporal intervals of accident events in the newly collected video clips display a more dispersed pattern, resembling a Gaussian-like distribution. This broader temporal spread results in a more balanced distribution of accident intervals across the video clips. Such a distribution helps to mitigate potential biases toward specific temporal segments, enabling MLLMs to more effectively and authentically learn the temporal dynamics and characteristics of accident events. By ensuring a diverse temporal representation, the proposed dataset enhances the robustness and generalizability of temporal localization models.

B ADDITIONAL DETAILS ON MODEL TUNING DATA

B.1 FUNCTIONAL TUNING DATA TEMPLATES

To avoid the trained model only responding to the specific instruction in the training data, we pre-defined a set of questions for each task to facilitate the model to be activated facing diverse queries.

Single-turn: Temporal Localization. Besides the diverse question set, we provide a set of answer templates to prompt the model to generate human-like responses in the temporal localization task:

```
question_templates = [
    "Do you know the exact times the traffic accident kicked off and wrapped up?",
    "Can you give me the start and end times of the traffic accident in the video?",
    "Any idea about the start and end time of that traffic accident we saw?",
    "Show me when the traffic accident gets going and when it's all over?",
    "What is the start and end time of the traffic accident in the video?",
    "Could you specify the timing of traffic accident's onset and conclusion?",
    "Please specify the precise timing of the traffic accident's onset and conclusion.",
    "At what timestamps does the traffic accident commence and finish?",
    "Can you delineate the duration of the traffic accident from beginning to end?",
    "When is the traffic accident initiated and terminated in the footage?"
]

answer_templates = [
    "Between {}.format,
    "In the time period {}.format,
    "During the span of {}.format,
    "It happens in {}.format,
    "At {}.format,
    "Exactly at {}.format,
    "Through {}.format,
    "Within the window of {}.format,
    "In the {} mark.format,
    "Around {}.format
]

User: random.choice(question_templates) <video>

GPT: random.choice(answer_templates)(annotation["accident-segments"])
```

Additionally, as we introduced in the main text, we also prompt the model to describe the content within the particular temporal segments: random-sampled normal segments or labeled accident segments. The pseudo code of temporal referring question-answer pair generation is presented:

```
question_templates = [
    "What's happened during {} in the video?".format,
    "What's the incident in the period of {}".format,
    "Maybe something wrong happened during {} in the provided video?".format,
    "What's the traffic situation in the period of {}".format,
    "Is the traffic flow captured by the video normal during {}".format,
    "Dose the accident happen during {} in the video?".format,
    "Does the video record any traffic disruptions or accidents around {}".format,
    "Is there any indication of an abnormal traffic event during {}".format,
    "Could you identify any mishaps in the time frame of {}".format,
    "Are there signs of vehicular distress or accidents within {}".format
]

User: random.choice(question_templates)(annotation["accident_segments"]) <video>

GPT: annotation['accident_description']
```

Single-turn: Accident Spatial Grounding. There are two spatial grounding tasks during our training process: accident-involved object grounding and accident region grounding. As for the accident-involved object grounding task, the pseudo code of generating conversations is presented as follows:

```
question_templates = [
    "Where is the {} involved in the accident?".format,
    "Where is the {} involved in the accident in the image?".format,
    "Provide the coordinates of the {} involved in the accident in the image?".format,
    "Can you point out the {} involved in the accident in the image and provide the coordinates of its location?".format,
    "Help me to locate the {} involved in the accident in the image and give me its coordinates, please.".format,
    "In the given image, could you find and tell me the coordinates of the {} involved in the accident?".format,
    "Guide me to the location of the {} involved in the accident within the image by providing its coordinates.".format,
    "I'd like to know the exact coordinates of the {} involved in the accident in the photo.".format,
    "Would you kindly provide the coordinates of the {} involved in the accident located in the picture?".format,
    "Can you find the {} involved in the accident in the image and give me the coordinates of where it is located?".format
]

for object in annotation["accident_objects"]:
    User: random.choice(question_templates)(object["label"]) <image>
    GPT: object['bbox']
```

We define the accident region as the union of all the bounding boxes of the involved objects. The question-answer pair generation for the accident region grounding task is formulated as follows:

```
accident_region = UNION(object["bbox"] for object in annotation["accident_objects"])

question_templates = [
    "Where is the accident?",
    "Where is the accident in the image?",
    "Provide the coordinates of the accident in the image?",
    "Can you point out the accident in the image and provide the coordinates of its location?",
    "Help me to locate the accident in the image and give me its coordinates, please.",
    "In the given image, could you find and tell me the coordinates of the accident?",
    "Guide me to the location of the accident within the image by providing its coordinates.",
    "I'd like to know the exact coordinates of the accident in the photo.",
    "Would you kindly provide the coordinates of the accident located in the picture?",
    "Can you find the accident in the image and give me the coordinates of where it is located?"
]

User: random.choice(question_templates) <image>

GPT: accident_region
```


Single-turn: Accident Recognition and Description. Here is the pseudo code of generating conversations for the task of accident recognition and description:

```
question_templates = [
    "Is there a traffic accident in the given video clip?",
    "Does this video capture a traffic accident?",
    "Is a traffic accident occurring at any point in this video?",
    "Can you detect any traffic collisions in this video?",
    "Is there evidence of a road traffic accident visible in this video clip?",
    "Throughout this video, is there an incident involving a traffic accident?",
    "Do you observe a car accident happening in the sequence of this video?",
    "Does this video document any vehicular collisions or crashes?",
    "Can you point out if there's a traffic-related accident depicted in this video?",
    "Is there any part of this video that shows a traffic mishap or collision?",
    "Watch this video and confirm if a traffic accident takes place at any moment."
]

User: random.choice(question_templates) <video>

GPT: Yes (for Accident) / No (for Normal). annotation["accident-description"]
```

B.2 INSTRUCTION TUNING DATA GENERATION

To generate instruction tuning data for traffic accident understanding, we design a set of instructions to guide the general-purpose multimodal language model, such as LLaMA-70B (Dubey et al., 2024) in our work, to generate multi-turn conversations. Trained on the generated multiple rounds of conversations, our TABot is expected to be endowed with a more comprehensive understanding of traffic accidents and equipped with the capability to provide more contextually relevant responses. The statistics of the generated conversation pairs based on our dataset are summarized in Table 2.

Specifically, we follow the in-context-learning (ICL) paradigm of LLaVA (Liu et al., 2024) and adapt it to our traffic accident understanding scenario. The detailed ICL prompting instruction for Llama3 is illustrated in Table 3. In particular, the caption of the video is provided for the model to imagine the visual content of the video, which should be accident-oriented in our work. Therefore, we organize the caption in a structured way: Sentence 1 describes the video source; Sentence 2 describes the accident-related content, in other words, the labeled accident description in the TAU-106K dataset; Sentence 3 complements the accident event with more detailed information, such as the temporal information and the objects involved in the accident.

Besides these complete sentences, we also list the annotated temporal segments and the bounding boxes of the accident-involved objects in the video clips, leading the model to refer to the specific content when generating responses to the user queries. An example of the structured caption and the generated multi-turn conversations are presented in Table 4.

Table 2: Generated conversation pairs based on our TAU-106K.

Task	Size	Response formatting
Detection & Description (Image)	55K	Detect and describe the accident
Detection & Description (Video)	52K	Detect and describe the accident
Temporal Localization	28K	$\{t_{start}, t_{end}\}$
Temporal Referring	54K	Describe the accident
Accident Grounding	28K	$[x0, y0, x1, y1]$
Object Grounding	45K	$[x0, y0, x1, y1]$
Complex Comprehension	70K	Multi-round conversation
All	332K	—

```
messages = [ {"role": "system", "content": f""You are an AI visual assistant, and
you are seeing a single video. What you see is provided with a few sentences, describing the same
video you are looking at. Answer all questions as you are seeing the video. The video mainly
focuses on the traffic situation.
```

In particular, if there is an accident, the timesteps of the accident are provided in the format of {start_time, end_time} with normalized time values.

In addition, if there is an accident, specific object locations involved in the accident are given, along with detailed coordinates. The accident region is the area where the accident occurred, presented as the union of all the bounding boxes of the involved objects. These coordinates are in the form of bounding boxes, represented as [x1, y1, x2, y2] with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the video and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the video, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:

- (1) one can see the content in the video that the question asks about and can answer confidently;
- (2) one can determine confidently from the video that it is not in the video. Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the video, for example, **asking about background knowledge of the objects in the video, asking to discuss accidents happening in the video, reasoning about the possible causes of the accident, analyzing the traffic rules that have been violated**, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.”}

```
]
for sample in fewshot_samples:
    messages.append({"role": "user", "content": sample['context']})
    messages.append({"role": "assistant", "content": sample['response']})
)
messages.append({"role": "user", "content": '\n'.join(query)})
```

Table 3: For each query, we illustrate the prompt construction process for Llama3 to collect `query['response']` from `query['context']`, using few-shot in-context-learning, where examples are from `fewshot_samples`, each example including input `sample['context']` and output `sample['response']`. Note that `messages` is the final prompt. The **bold** text indicates the tailored instruction for the accident-oriented conversation generation. In this illustration, we provide the instruction used to generate the conversation response, please see also see its in-context-learning example in Table 4 for details.

Captions

This video is sourced from a traffic surveillance camera.

Surveillance cameras captured the road surface of a rural road, at 0.41 normalized timestamp, a truck collided with an electric rider as it drove straight ahead, at 0.53 normalized timestamp, the owner of the electric bike fell to the ground.

At 0.450 normalized timestamp, a man in a red coat riding an electric bike located at [0.324, 0.301, 0.488, 0.683] and a straight freight car located at [0.472, 0.101, 0.810, 0.849] were involved in the accident.

accident_timesteps: {0.41, 0.53}

accident_objects:

man in a red coat riding an electric bike: [0.324, 0.301, 0.488, 0.683]

straight freight car: [0.472, 0.101, 0.810, 0.849]

accident_region: [0.324, 0.101, 0.810, 0.849]

Conversation

Question:

Can you describe the types of vehicles involved in the accident captured in the video?

Answer:

The accident involved a freight truck and an electric bike. The freight truck was moving straight ahead on a rural road when it collided with the electric bike.

===

Question:

At what point in the video does the accident start and end?

Answer:

The accident starts at the normalized timestamp of 0.41 and ends at 0.53.

===

Question:

Where exactly did the accident occur in relation to the video frame?

Answer:

The accident occurred within the coordinates [0.324, 0.101, 0.810, 0.849] in the video frame. This region represents the unified area of both involved objects, the electric bike and the freight truck, during the accident.

===

Question:

How did the rider of the electric bike respond to the collision?

Answer:

After the collision, the rider of the electric bike fell to the ground. This incident happened around the 0.53 normalized timestamp in the video.

===

Question:

What is the appearance of the person involved in the accident?

Answer:

The person involved in the accident, specifically the rider of the electric bike, was wearing a red coat. This detail is captured around the 0.450 normalized timestamp in the video when describing their location in the frame.

===

Question:

Can you analyze the possible cause of the accident shown in the video?

Answer:

Based on the video, one possible cause of the accident could be the relative positions and movements of the freight truck and the electric bike. The truck was moving straight ahead on a rural road, suggesting that the visibility or the reaction time could have been compromised for either the truck driver or the electric bike rider. Without specific video evidence of traffic signals or signs, one might infer that perhaps there was a failure to yield the right of way or an issue with speed control. However, definitive causes would require further investigation into the traffic conditions, driver behavior, and environmental factors at the time of the accident.

Table 4: One example used in in-context-learning to construct visual conversation data.

C SUPPLEMENTARY ABLATION EXPERIMENTS

Fine-tuning on Other Baselines. We extend our experiments to include recent MLLMs, Video-LLaMA-2 (Cheng et al., 2024) and Qwen2-VL (Yang et al., 2024), fine-tuned on our proposed TAU-106K dataset to evaluate its quality and significance. The results, shown in Table 5, demonstrate significant performance improvements across all tasks after fine-tuning, validating the effectiveness of our dataset and training recipes. After fine-tuning, the performance gap between Video-LLaMA-2 and GroundingGPT is reduced, highlighting the importance of fine-tuning on target datasets for accident understanding tasks. Our TABot model based on GroundingGPT, still outperforms fine-tuned Video-LLaMA-2, particularly in image understanding tasks. As for the SOTA model Qwen2-VL, the pre-trained model already achieves competitive performance, and the fine-tuning on TAU-106K further boosts the performance, reaching the highest performance in all tasks. These findings emphasize the necessity of fine-tuning on domain-specific datasets, demonstrating the effectiveness and quality of our comprehensive TAU-106K dataset.

Table 5: The results of fine-tuning other baselines on TAU-106K.

	Video Understanding			Image Understanding		
	CLS (Acc)	TL (AP@50)	CAP (BERT)	CLS (Acc)	AG (AP@50)	CAP (BERT)
GroundingGPT	50.00	2.40	55.70	63.75	14.25	45.00
+ TABot	81.00	20.12	82.31	90.75	70.03	75.20
Video-LLaMA-2	64.00	2.10	62.20	63.30	31.57	63.21
+ TABot	79.90	19.30	83.36	77.80	57.25	73.71
Qwen2-VL	72.65	15.76	61.61	58.35	47.52	66.12
+ TABot	82.65	22.50	83.09	92.00	77.61	76.32

Experiment Results of 7:3 Split. Since the amount of our TAU-106K dataset is large enough, 1/10 of the data (5K videos and 5K images) is sufficient for testing. However, we have conducted additional experiments with a 7:3 train/test split to verify the model’s generalization ability to overfitting, as reported in Table 6. The model’s performance remains consistent across different tasks with a slight decrease in accuracy, demonstrating its robustness to different train/test splits.

Table 6: The results of fine-tuning TAU-106K with the 7:3 split.

	Video Understanding			Image Understanding		
	CLS (Acc)	TL (AP@50)	CAP (BERT)	CLS (Acc)	AG (AP@50)	CAP (BERT)
TABot (9:1)	81.00	20.12	82.31	90.75	70.03	75.20
TABot (7:3)	79.95	19.08	81.57	88.97	68.29	74.55

Data Imbalance in Fine-grained Categories. The class imbalance issue is inevitable in data collection and our data distribution also fairly reflects real-world situations. The primary purpose of this paper is to facilitate the development of MLLM on large-scale traffic datasets and learn models that closely resemble real-world conditions. Besides, we provide Table 7 to show the accuracy of different types of traffic accidents in both video/image accident recognition tasks. While categories like MnMV show slightly lower performance due to the limited amount of training data, the model’s performances across different accident types are generally comparable. Overall, the model achieves consistent and satisfactory results across the normal class and the five accident classes.

Table 7: Accident proportion and class-wise accuracy for video and image accident recognition.

Accident Category	Proportion in TAU-106k	Accuracy for Video Accident Recognition	Accuracy for Image Accident Recognition
Normal	44.1%	78.80	92.23
SMV	9.7%	80.56	79.75
MMV	32.9%	84.39	87.40
MV&nMV	9.8%	92.02	82.87
V&P	2.2%	94.29	84.78
MnMV	1.3%	77.78	78.26

Qualitative Visualization and Comparison. We provide visualizations and comparisons on the functional image and video understanding tasks in Figure 4, 5, 6, and 7. Besides, the multi-turn dialogue samples on traffic accident videos are presented in Figure 8 and Figure 9.

Question [RD]: Can you confirm the occurrence of a vehicle accident in this picture? Please respond with a ‘Yes’ or ‘No’. Following that, describe the image in one sentence.

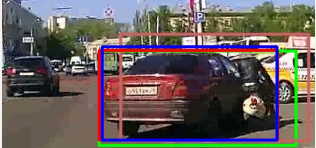
Ground Truth: Yes. An electric bicycle rider wearing black collided with a red car making a right turn due to faulty brakes.

GPT-4o: Yes, the image shows a traffic accident involving a red car that appears to have collided with another vehicle in a busy intersection.

Qwen2-VL: No. The image shows a busy street with multiple cars and a pedestrian crossing the road, but there is no visible accident.

Gemini-1.5-Pro: No. A red car is turning right at an intersection in a city.

TABot-Chat (Ours): Yes. Because a black-clad electric bicycle rider missed the brakes, it collided with a red car turning right.



Question [SG]: Where is the traffic accident? The answer should be given in normalized $[x_{min}, y_{min}, x_{max}, y_{max}]$ format.

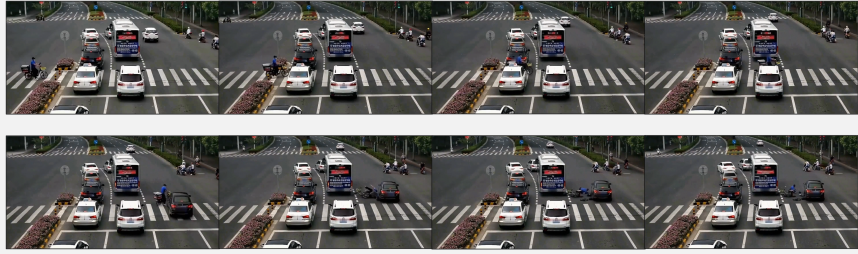
Ground Truth: [0.310, 0.343, 0.881, 0.935]

Qwen2-VL: [0.31, 0.35, 0.94, 0.97]

Gemini-1.5-Pro: [0.38, 0.24, 0.98, 0.94]

TABot-Chat (Ours): [0.321, 0.319, 0.878, 0.937]

Figure 4: One qualitative example of the functional image understanding tasks.



Question [RD]: Does this video document any vehicular collisions or crashes? Please respond with a ‘Yes’ or ‘No’. Following that, describe the video in one sentence.

Ground Truth: Yes. The surveillance camera showed a black car colliding with a blue-clad electric bike rider at a city intersection, knocking the rider to the ground.

Qwen2-VL: Yes. A motorcycle is hit by a car and the rider falls to the ground.

Gemini-1.5-Pro: Yes. A cyclist is hit by a car at a crosswalk.

TABot-Chat (Ours): Yes. Surveillance cameras captured the road surface of a city street, as a black car ran through the intersection, it collided with an electric bicycle rider dressed in blue, the cyclist was knocked to the ground.

Question [TL]: Show me when the traffic accident gets going and when it’s all over. The answered timestamp should be formatted in a normalized manner, using {start_time, end_time}.

Ground Truth: {0.55, 0.86}

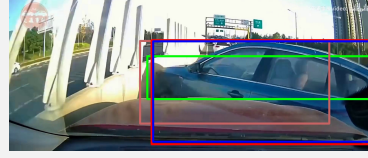
Qwen2-VL: {0.50, 0.80}

Gemini-1.5-Pro: {0.80, 0.88}

TABot-Chat (Ours): {0.56, 0.84}

Figure 5: One qualitative example of the functional video understanding tasks.

Question [RD]: Are there any indications of a traffic accident in this image? Please respond with a ‘Yes’ or ‘No’. Following that, describe the image in one sentence.



Ground Truth: Yes. A blue car did not pay attention to the road conditions while making a left turn, which caused the current vehicle to take evasive action, leading to a collision between the blue car and the current vehicle. Subsequently, both cars crashed into the left road guardrail.

GPT-4o: Yes, the image shows a car colliding with a barrier while another vehicle is in the foreground.

Qwen2-VL: No. The image shows a car driving on a road with other vehicles turning left and traffic signs visible.

Gemini-1.5-Pro: No. The image shows a car driving on a highway, approaching a bridge.

TABot-Chat (Ours): Yes. Because a blue car did not notice the condition of the road when turning left, causing the current vehicle to avoid it, the current vehicle collided with the blue car, then the blue car and the current vehicle collided with a fence on the left side of the road.

Question [SG]: In the given image, could you tell me the coordinates of the blue car recorded in crash incidents? The answer should be given in normalized $[x_{min}, y_{min}, x_{max}, y_{max}]$ format.

Ground Truth: [0.392, 0.275, 0.998, 0.956]

Qwen2-VL: [0.38, 0.38, 1.0, 0.66]

Gemini-1.5-Pro: [0.36, 0.28, 0.88, 0.82]

TABot-Chat (Ours): [0.396, 0.283, 1.000, 0.939]

Figure 6: One qualitative example of the functional image understanding tasks.



Question [RD]: Do you observe a car accident happening in the sequence of this video? Please respond with a ‘Yes’ or ‘No’. Following that, describe the video in one sentence.

Ground Truth: Yes. A vehicle was driving on a city street when two electric bicycle riders collided because one, dressed in dark clothing, failed to brake in time. The other rider, wearing a white helmet, was also struck, and both fell to the ground.

Qwen2-VL: Yes. A motorcycle falls down on the road in the video.

Gemini-1.5-Pro: Yes. A dashcam video shows a motorcyclist falling from their vehicle after colliding with another vehicle on a city street.

TABot-Chat (Ours): Yes. The vehicle was traveling on the streets of the city, because an electric bicycle rider in dark clothes did not brake in time, it collided with a white helmeted electric bicycle rider, both riders fell to the ground.

Question [TL]: Can you delineate the duration of the traffic accident from beginning to end? The answered timestamp should be formatted in a normalized manner, using {start_time, end_time}.

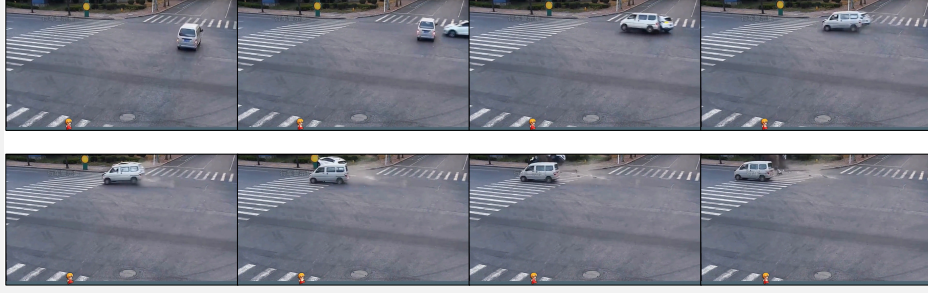
Ground Truth: {0.39, 0.75}

Qwen2-VL: {0.25, 0.75}

Gemini-1.5-Pro: {0.80, 0.95}

TABot-Chat (Ours): {0.39, 0.72}

Figure 7: One qualitative example of the functional video understanding tasks.



User: How many vehicles were involved in the accident captured by the traffic surveillance camera?

TABot-Chat: There were two vehicles involved in the accident: a silver bread truck and a white car. Additionally, the white car also crashed into a sign on the side of the road.

User: What is the duration of the accident as seen in the video?

TABot-Chat: The accident starts at the normalized timestamp of 0.12 and ends at 1.0, indicating a relatively long duration.

User: Can you describe the sequence of events in the accident?

TABot-Chat: The accident unfolded in two stages. Initially, at 0.218 normalized timestamp, the white car collided with the silver bread truck at the intersection. Later, at 0.609 normalized timestamp, the same white car crashed into a sign on the side of the road.

User: Where did the accident occur in relation to the video frame?

TABot-Chat: The accident occurred within the region bounded by the coordinates [0.317, 0.000, 0.941, 0.279] in the video frame. This region encompasses both collision scenes, including the intersection where the white car hit the silver bread truck and the area where the white car crashed into the sign.

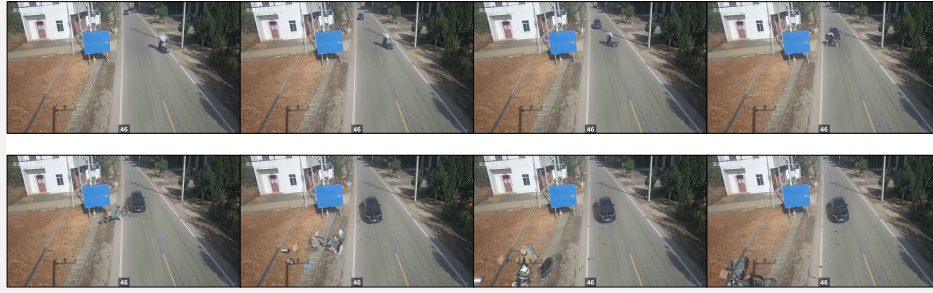
User: What can be inferred about the white car's behavior during the accident?

TABot-Chat: Based on the video, it appears that the white car was involved in two separate collisions. The first collision occurred when it hit the silver bread truck, and the second collision happened when it crashed into the sign on the side of the road. This sequence of events suggests that the white car may have lost control or failed to stop after the initial collision, leading to the subsequent crash into the sign.

User: Can you provide any insights into the possible causes of the accident?

TABot-Chat: The video evidence suggests that the white car may have been traveling at an excessive speed or failed to obey traffic rules, leading to the initial collision with the silver bread truck. The subsequent crash into the sign could be attributed to the driver's loss of control or failure to react appropriately after the initial collision. However, without additional information, such as the speed of the vehicles, road conditions, or driver behavior, a definitive cause cannot be determined solely from the footage.

Figure 8: One qualitative example of the multi-turn dialogue on the traffic accident video.



User: What type of vehicles were involved in the accident captured by the traffic surveillance camera?

TABot-Chat: The accident involved a black car and a blue tricycle. The black car was moving on the country road when it collided with the blue tricycle, which was turning left.

User: What was the cause of the accident according to the video?

TABot-Chat: The accident occurred because the black car failed to brake on time, resulting in a collision with the blue tricycle.

User: At what point in the video does the accident start and end?

TABot-Chat: The accident starts at the normalized timestamp of 0.30 and ends at 0.89.

User: Where exactly did the accident occur in relation to the video frame?

TABot-Chat: The accident occurred within the coordinates [0.490, 0.223, 0.582, 0.410] in the video frame. This region represents the unified area of both involved objects, the blue tricycle and the black car, during the accident.

User: What happened to the tricycle and the cyclist as a result of the accident?

TABot-Chat: The tricycle overturned, and both the tricycle and the cyclist were hit from the road surface.

User: Can you analyze the possible consequences of the accident shown in the video?

TABot-Chat: Based on the video, the consequences of the accident could be severe. The tricycle overturning and the cyclist being hit from the road surface suggest potential injuries to the cyclist. The accident could also have caused damage to both vehicles involved. Additionally, the accident may have caused a disruption to traffic flow on the country road. A thorough investigation would be necessary to determine the full extent of the consequences.

User: What traffic rule might have been violated by the black car driver?

TABot-Chat: The black car driver might have violated the rule of maintaining a safe following distance or failing to yield to the blue tricycle, which was turning left. The driver's failure to brake on time suggests a possible failure to exercise due care and caution while driving.

Figure 9: One qualitative example of the multi-turn dialogue on the traffic accident video.