TAU-106K: A NEW DATASET FOR COMPREHENSIVE UNDERSTANDING OF TRAFFIC ACCIDENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have demonstrated impressive performance in general visual understanding tasks. However, their potential for high-level and fine-grained comprehension, such as humor or anomaly understanding, remains unexplored. Targeting traffic accidents, a critical and practical scenario within anomaly understanding, we explore the advanced capabilities of MLLMs and introduce TABot, a multimodal MLLM tailored for accident-related tasks. To facilitate this, we first developed TAU-106K, a large-scale multimodal dataset comprising 106K traffic accident-related videos and images, sourced from academic benchmarks and public platforms. The dataset is meticulously annotated through a video-to-image annotation pipeline, ensuring comprehensive and high-quality labels. Upon TAU-106K, our accident-oriented MLLM TABot is trained in a two-step approach to integrate multi-granularity accident understanding tasks, including accident recognition, spatial-temporal grounding, with an additional accident description task to guide the model in comprehending the nature of traffic accidents. Extensive experiments demonstrate the superior performance of TABot in traffic accident understanding, underscoring both its potential for high-level anomaly understanding and the robustness of the TAU-106K dataset. All datasets, annotations, and models will be publicly released for future research.

1 INTRODUCTION

Traffic Accident Detection (TAD) has always been a crucial and practical task in public safety and transportation management. The development of advanced computer vision and deep learning has enabled the automation of TAD, providing real-time accident alerts and facilitating accident analysis. Despite significant research on traffic accident detection (Hasan et al., 2016; Yao et al., 2019; Fang et al., 2022), existing TAD methods often rely on conventional visual feature extraction techniques that lack high-level comprehension and reasoning capabilities in interpreting intricate spatial-temporal dynamics. The introduction of language models in multimodal large language models (MLLMs) brings powerful reasoning capabilities, providing better semantic understanding and contextual reasoning capabilities for fine-grained accident analysis. Besides, MLLMs excel at aligning visual and textual modalities, enabling them to integrate multiple complex tasks in one model, significantly enhancing the interpretability and utility of the system compared to traditional models.

Recent advancements in aligning large language models (LLMs) on visual-language datasets have driven remarkable progress in MLLMs (Achiam et al., 2023). With extensive pre-training and instruction tuning, MLLMs are becoming increasingly proficient at visual representation learning and human-like logical reasoning for the comprehensive general-purpose understanding of visual data (Li et al., 2023; Zhu et al., 2023; Liu et al., 2024). However, due to the gap between general and specific comprehension tasks, MLLMs pre-trained on general domains often underperform on domain-specific tasks. As illustrated in Figure 1, these models may misinterpret the visual content of traffic accidents where a vehicle collision occurs, leading to inaccurate accident detection and reasoning. General-purpose MLLMs may make errors in understanding traffic accidents due to two main reasons: (i) Traffic accident detection requires MLLMs to grasp abstract concepts like *anomaly* and *accident*, which are defined by human criteria, as well as interpret complex motion patterns involving multiple objects. The lack of comprehensive annotated data specific to such high-level semantics comprehension hampers the models' understanding of accident scenarios. (ii) The

Question: Do you notice any traffic accident scenarios in this picture? Please respond with a 'Yes' or 'No'. Following that, describe the image in one sentence.

Ground Truth: Yes. A black car did not brake in time and collided with an electric bicycle rider wearing a white top.
GPT-4o: No, the image shows a busy city intersection with vehicles and motorcycles waiting at a traffic light.
Qwen2-VL: No. The image shows a busy urban street with cars, scooters, and pedestrians, but there is no visible accident.
Gemini-1.5-Pro: No. The image shows a sunny city street scene with several motorbikes and cars at an intersection under a highway overpass.
TABot (Ours): Yes. Because a black car brake was not on time, it collided with a white coat electric bicycle rider.

Figure 1: One example to illustrate the limitations of general MLLM in understanding traffic accidents. In the scenario where a vehicle collision occurs due to a sudden lane change by the leading vehicle, GPT-4o, Qwen2-VL, and Gemini-1.5-Pro fail to detect this issue.

visual representations of accident occurrences differ significantly from general scenes, necessitating realigning these visual representations with the semantic understanding towards traffic accidents.

To address these challenges and pioneer an MLLM specialized in traffic accident comprehension, we first created TAU-106K, a large-scale multimodal traffic accident dataset containing 106K videos and images with detailed accident-related annotations. In particular, we aggregate academic benchmarks and crawl traffic accident videos from public platforms, resulting in a diverse and high-quality dataset. To ensure annotation efficiency and quality, we design a video-to-image annotation pipeline, where the annotations, especially the textual accident descriptions, are manually annotated at the video level and inherited at the image level. Additionally, to further enhance the capabilities of MLLMs in traffic accident understanding and human-like conversation, we utilize the advanced general-purpose MLLMs (Dubey et al., 2024; Achiam et al., 2023) to integrate our data annotations to form multi-turn dialogues.

Using TAU-106K, we reorganize the annotations into instructional data to unlock MLLMs' potential in traffic accident understanding and introduce TABot, a specialized MLLM for traffic accident comprehension across both image and video modalities. We adopt a two-step training approach for TABot: **functional tuning** to engage multi-granularity accident detection capabilities activation, and **instruction tuning** to enhance contextual accident-related comprehension and instruction following capabilities. In particular, during functional tuning, we propose two training strategies to serve temporal localization, the most crucial task in traffic accident understanding: (i) Negative Segment Referring (NSR), which utilizes contrastive learning to heighten the model's sensitivity to accident boundaries, and (ii) Video Spatial Alignment (VSA), which facilitates the model's temporal localization by complementing the spatial grounding at the image level within the same scene. Additionally, we insert task flags into the queries to guide the model's targeted responses to specific tasks such as temporal localization ([TL]) and spatial grounding ([SG]), meanwhile mitigating the catastrophic forgetting during the subsequent instruction tuning. Our TABot not only addresses the limitations of current MLLMs in recognizing and comprehending traffic accidents but also sets a new standard for the fine-grained spatiotemporal analysis of such critical events.

The contributions of our work can be summarized as:

- We introduce TAU-106K, a large-scale multimodal traffic accident dataset comprising 106K videos and images, annotated through a video-to-image annotation pipeline for comprehensive accident understanding. Additionally, we generated multi-turn dialogues using an automated paradigm, enhancing the dataset's utility for training and evaluation.
- We present TABot, an end-to-end MLLM designed for detailed traffic accident understanding. The model is trained using a two-step approach: functional tuning for unlcoking the multi-granularity accident detection capabilities, followed by instruction tuning to align with human intentions and enhance general comprehension.

- Through joint video-image-text annotation, we advance the TABot's semantic alignment and accident understanding. Extensive experiments demonstrate TABot's superior performance in understanding traffic accident scenarios. The dataset, annotations, and models will be released for future research.

2 Related Work

Multimodal Large Language Models. Extensive research works have been conducted to enable LLMs to process visual information. The typical framework adds an adapter between pre-trained visual models and LLMs to align features from different modalities (Li et al., 2023; Zhu et al., 2023; Liu et al., 2024). However, videos, as an advanced form of visual data, introduce visual information that poses greater challenges for LLMs in aligning with video content (Maaz et al., 2023; Lin et al., 2023; Chen et al., 2023a; Zhang et al., 2023; Qian et al., 2024; He et al., 2024; Cheng et al., 2024; Xu et al., 2024; Zhang et al., 2024; Chen et al., 2023d; 2024). As one of the latest efforts in video MLLMs, Qwen2-VL (Yang et al., 2024) models the three dimensions of time, height, and width using Multimodal Rotary Position Embedding (M-RoPE). However, current models still have limitations in segment understanding tasks for high-level semantic data. Datasets towards general video comprehension often lack functional annotations for executing specific tasks, and the absence of temporal reasoning annotations in the pre-training and fine-tuning phases hinders LLMs' ability to understand temporal or segment-centric information.

MLLMs for Spatial-Temporal Grounding. Fine-grained 2D image grounding with MLLMs is one of the initial fields of engagement. The majority of studies have standardized the grounding task coordinates to text format to ensure a unified paradigm. Works such as MiniGPT-v2 (Chen et al., 2023b), Qwen-VL (Bai et al., 2023), Kosmos-2 (Peng et al., 2023), and Shikra (Chen et al., 2023c) have developed visual grounding-related pre-training and instruction-tuning datasets to endow models with the capability for fine-grained localization. Furthermore, Ferret (You et al., 2023) has introduced negative samples to enhance model robustness. In the realm of video MLLMs, VTimeLLM (Huang et al., 2024) has first pushed toward comprehending time boundaries by employing MLLMs. TimeChat (Ren et al., 2024) modeled temporal features using a sliding window Q-former, equipping models to perform dense video description and action localization tasks. GroundingGPT (Li et al., 2024) has merged fine-grained localization tasks with image, video, and speech modalities, achieving a universally applicable multimodal and multi-granularity understanding.

Traffic Accident Detection and Understanding. In traditional deep learning-based TAD, methods are classified into single-stage (Hasan et al., 2016) and two-stage paradigms (Yao et al., 2019; Fang et al., 2022). Recent advances have seen the integration of textual information into the task of TAD. TTHF (Liang et al., 2024) deployed text-driven attention mechanisms to focus on specific representations of anomalous events within videos. On the MLLM front, VisionGPT (Wang et al., 2024a) has unified open-vocabulary grounding with MLLM to create a training-free system capable of performing zero-shot accident alerts. Extensive empirical studies by (Cao et al., 2023) have authenticated the effective accident recall and description capabilities of GPT-4(V) on traffic accident images. Despite these developments, current MLLMs still exhibit limitations in comprehensive traffic accident understanding, restricted by the lack of multi-granularity accident-related datasets. Therefore, we aims to bridge this gap by introducing TAU-106K, a large-scale multimodal traffic accident dataset, labeled with accident categories, temporal and spatial grounding, and textual descriptions, to facilitate the development of traffic accident-oriented MLLMs.

3 TAU-106K: Video-Image Traffic Accident Understanding Dataset

To advance the development of Multimodal Large Language Models (MLLMs) for traffic accident analysis, we introduce TAU-106K, a comprehensive multi-modal dataset integrating video and image data for traffic accident understanding, manually labeled with category, temporal, spatial, and textual description annotations, whose detailed pipeline is illustrated in Figure 2. This dataset is designed to enhance the temporal and spatial grounding capabilities of MLLMs, enabling more precise detection and understanding of traffic accidents.

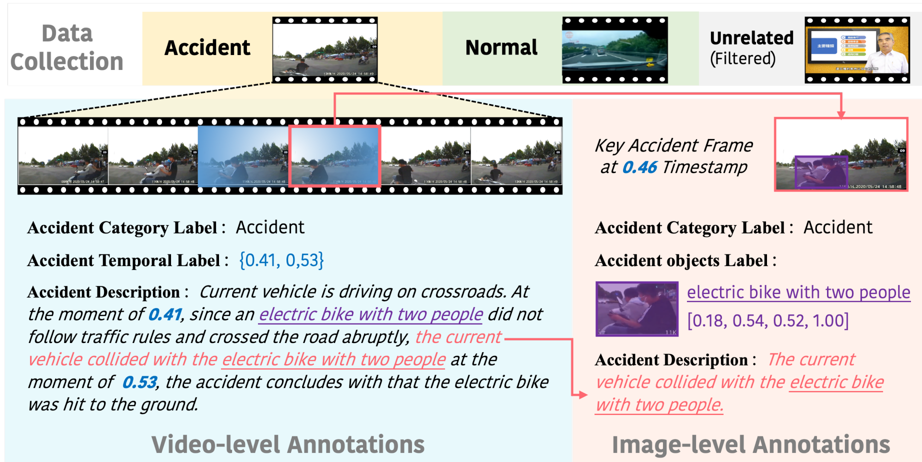


Figure 2: The data collection and annotation pipeline for TAU-106K.

3.1 VIDEO-BASED DATA COLLECTION AND ANNOTATION

Video Data Collection and Preprocessing. While traffic accident understanding is a critical public safety task and has been extensively studied, the available open-source datasets are limited in both scale and diversity, often featuring low-resolution video data. To address this, we aggregate established traffic accident benchmarks such as TAD (Xu et al., 2022), DoTA (Yao et al., 2022), and CCD (Bao et al., 2020), selecting high-quality video clips as a foundational dataset for further annotation. We further expand the dataset by crawling road surveillance and dashcam footage from platforms like *YouTube* and *BiliBili*, capturing diverse real-world traffic conditions. Despite the abundance of traffic accident videos on the Internet, they are often unstructured and lack detailed annotations. For the crawled raw videos, we first crop them into individual clips using scene change detection toolkit `PySceneDetect`, avoiding disruptive scene transitions, and then manually filter out irrelevant or low-quality videos. Consequently, we obtain a collection of 51.5K traffic-focused video clips mixed with academic benchmarks and social media platforms, as illustrated in the first part of Figure 2.

Video-based Accident Annotations. All existing benchmarks for traffic accident understanding lack comprehensive annotations, especially in terms of accident descriptions, which are crucial for enabling MLLMs to understand accident events in detail. To bridge this gap, we annotate or supplement annotations in three key aspects:

1. Accident Category: whether an accident is present in the clip. Each clip is reviewed to determine if an accident is present, labeled either as *Accident* or *Normal*. For clips marked as *Accident*, we further categorize the accident type into *Single Motor Vehicle (SMV) Accident*, *Multiple Motor Vehicle (MMV) Accident*, *Multiple non-Motor Vehicle (MnMV) Accident*, *Motor Vehicle and non-Motor Vehicle (MV&nMV) Accident*, and *Vehicle and Pedestrian (V&P) Accident*.

2. Accident Duration: the specific time points of the accident occurrence. Annotators precisely identify the start and end timestamps of the accident within each clip, yielding the time points $\{t_{start}, t_{end}\}$ where the accident begins and ends. In particular, the start time t_{start} should be the exact frame when the accident event begins, such as the moment of collision, while the end time t_{end} is marked when the event concludes (e.g., when vehicles stop). These timestamps are normalized relative to the respective clip's duration to ensure consistency.

3. Accident Description: a detailed textual description of the nature of the accident. Notably, the accident description is absent in all existing traffic accident benchmarks, which is infeasible for MLLMs to understand the accident event in detail. Due to the arbitrary nature of textual descriptions, these specific annotations are crafted following detailed guidelines to ensure consistency and precision. In detail, the description template for *Accident* is structured to include the traffic scenario (urban, highway, etc.), the objects involved in the accident (vehicles, pedestrians, etc.), the nature of the accident (collision, scrape, etc.), and aftermath, ensuring comprehensive and structured annota-

tions. Beyond the content-based descriptions, annotators are also encouraged to infer the potential causes of the accident, such as traffic rule violations or improper driving behaviors. The detailed annotation template is formulated according to the footage source, either *Dashcam* or *Surveillance camera*, as follows:

```
Description = Footage Source + Traffic Scenario + Cause of the Accident +
Content of the Accident + Aftermath
```

> Current vehicle is driving on (***Dashcam***) / The surveillance camera captures (***Surveillance camera***) the road of [***TODO****: the traffic scenario*]. At t_{start}, since [***TODO****: cause of the accident*], [***TODO****: the content of the accident, including the nature of the accident and the objects involved*] , at t_{end}, the accident concludes with [***TODO****: the aftermath of the accident*].

where the placeholders ***TODO*** are filled by the annotators with the specific information of the accident event. This structured approach ensures clarity, consistency, and coverage of relevant details. Although we depict the annotation process as three discrete tasks, they are performed simultaneously in practice execution. This integrated approach ensures consistency and coherence in annotations, reflecting the interconnected nature of these tasks.

3.2 Image-based Data Derived from Video-based Data

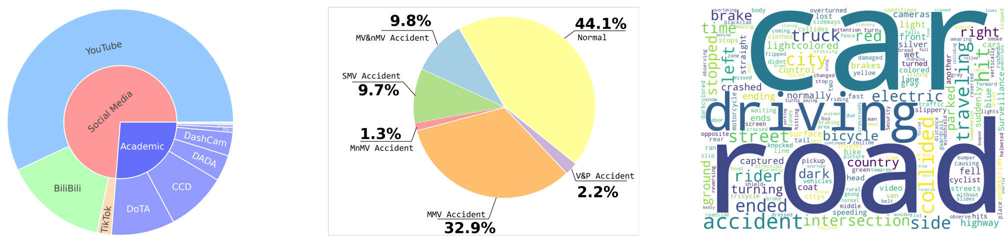
Restricted by the computational overhead and the complexity of video data, MLLMs are incapable of learning fine-grained visual features from video data. To mitigate this, we derive image data from video clips, enabling MLLMs to align accident-related visual information with textual semantics, whose detailed pipeline is illustrated in the third part of Figure 2.

Image Data Collection and Selection. While there are a few image-only accident datasets (e.g., TaskFix (Juan et al., 2021a), YouTubeCrash (Juan et al., 2021b)), most of the image data in our TAU-106K is sampled and derived from the video clips we collected and annotated as the previous section. Guided by the temporal localization annotations in the video clips, we first extract candidate frames by uniformly sampling frames within the labeled accident duration. These frames are then evaluated by annotators to select keyframes that best represent the accident events, based on the *Accident Description* in the video annotations. Notably, the time points of the selected keyframes are preserved to keep the temporal alignment between the video and image data, which also serves our video spatial alignment strategy in the subsequent model training. The selected keyframes are then used as the image data for further spatial grounding annotations. In addition to accident-related frames, we randomly sample accident-free frames to maintain balance between accident and normal instances in the image data.

Image Annotations Derived from Video Annotations. For images sourced from existing benchmarks, we adopt the available annotations and extend them by referring to video-based annotation guidelines. For the images derived from video data, we inherit the accident-related annotations from the video clips, including the *Accident Category*, *Accident Duration*, and *Accident Description*, and annotators only proceed to localize the accident-involved objects in the images. In particular, labels for involved objects are derived directly from the accident descriptions, ensuring that the annotated objects are those explicitly mentioned. For instance, given the accident description as "*A blue car collides with a pedestrian in white clothes*", the corresponding objects will be labeled as *blue car* and *pedestrian in white clothes*, respectively. This instance-specific labeling helps MLLMs focus on the objects directly involved in the accident, minimizing distractions from irrelevant objects of the same category that may appear in the scene. For the image-level accident descriptions, we extract the *content of the accident* in the video-based accident description to maintain consistency across the video and image data and reduce the annotation workload.

3.3 Data Statistics

TAU-106K comprises 106K multimodal data instances, including 51.5K video clips and 54K images, all with high-quality annotations. The majority of the video clips and images are in 720p resolution and are sourced from both open-source benchmarks and social media platforms, as shown in Figure 3.2. Among the TAU-106K, 56% of instances are labeled as *Accident* and 44% as *Normal*, with detailed category distribution shown in Figure 3.2. The balanced distribution of accident-related and accident-free instances ensures that the model is trained robustly, avoiding biases towards accident occurrences. The average video duration of processed and filtered clip is 10.3 seconds, with annotated accidents lasting an average of 3 seconds (approximately 25% of the video clip). As for the image data, 45K accident-involved objects are grounded, with an average of 1.6 bounding boxes per image and an average bounding box area covering 7.9% of the image. Our accident descriptions are detailed and diverse, covering a broad range of traffic scenarios, accident types, and objects



(a) Data source distribution. (b) Accident type distribution. (c) Word cloud of description.

Figure 3: Data source distribution, accident type distribution, and word cloud of accident descriptions in TAU-106K dataset.

involved, as shown in the word cloud of accident descriptions in Figure 3.2. Further statistics, including video duration, accident occurrence distribution, and bounding box area, are provided in the appendix.

4 TABot: A Chatbot for Traffic Accident Understanding

We introduce TABot, a multimodal fine-grained MLLM developed by leveraging instructional data constructed from the TAU-106K dataset. TABot is compatible with both video and image modalities, enabling it to perform fine-grained understanding and reasoning tasks in traffic accident scenarios. The proposed TABot integrates a suite of traffic accident-related tasks, including accident recognition, temporal localization, spatial grounding, and accident description generation, as depicted in Figure 4.

4.1 Model Overview

We advance the TABot upon GroundingGPT (Li et al., 2024), a model known for its strong performance in fine-grained image and video understanding. By fine-tuning this general-purpose MLLM on our annotated TAU-106K dataset, we enhance its capabilities for traffic accident comprehension on several functional tasks:

Accident Recognition. Targeting this fundamental task in traffic accident understanding, TABot is trained to detect the presence of accidents for both video clips and images. For each visual input, TABot is expected to answer whether an accident is present with a binary response, *Yes* or *No*.

Accident Description Generation. Beyond simple recognition, TABot generates supporting evidence for its decisions in the form of accident descriptions. These textual annotations aid the model in understanding and characterizing the nature of the accidents. The combination of accident recognition and description generation ensures that the recognition decision is based on a deep comprehension of the scene.

Accident Temporal Localization In addition to identifying accident occurrences, the practical application of traffic accident understanding often requires precise temporal localization of the accident event. Drawing on the temporal localization annotations in our TAU-106K dataset, TABot is trained to determine the precise temporal boundaries of the accident occurrence. The responses are normalized to the video duration and denoted as $\{t_{start}, t_{end}\}$, with specific tokens "{" and "}", using curly braces to indicate the temporal boundaries.

Accident Spatial Grounding In conjunction with temporal localization, our TABot is also trained to spatially ground the accident-involved objects and the global accident region within images. This task links accident-specific visual representations to corresponding language descriptions, filling the visual-semantic alignment gap in previous general-purpose MLLMs. Similar to temporal localization, the spatial grounding answers are expected to be normalized to the image size and are enclosed in angular bracket tokens, "[" and "]".

Following previous works (Li et al., 2024; Liu et al., 2024), we adopt a two-stage fine-tuning approach: **functional tuning** and **instruction tuning**. Firstly, during the functional tuning stage, TABot is jointly fine-tuned on both image and video data, focusing on the four key tasks mentioned above. We generate structured single-round conversations for each task to facilitate the model's understanding of traffic accidents, and the detailed conversation construction is presented in the appendix. To ensure the model's flexibility in handling multiple tasks, task-specific flag tokens (Accident Recognition & Description [RD], Temporal Localization [TL], and Spatial Grounding [SG]) are inserted at the start of each query to guide TABot's responses. Two additional training strategies are proposed to further improve performance in temporal localization: Negative Segment Referring (NSR) and Video Spatial Alignment (VSA), which promote the performance from the perspective of contrastive learning and spatial understanding, respectively. Negative segment referring involves sampling accident-free segments before the occurrence of an accident and training the model through accident segment

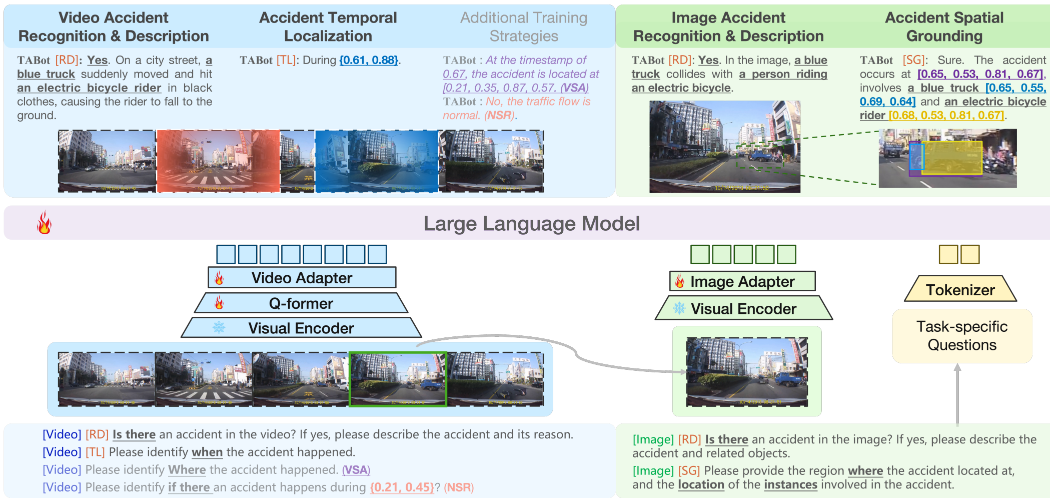


Figure 4: The model architecture and capabilities of the TABot. Additionally, two training strategies designed for temporal localization task, Negative Segment Referring (NSR) and Video Spatial Alignment (VSA), are also illustrated in lighter colors.

referring, serving as contrastive learning to highlight the perception of accident occurrences. Specifically, TABot is queried about both accident-free and accident-labeled durations, training the model to respond with the corresponding decision answer. On the other hand, benefiting from the unified video-to-image annotation pipeline, Video Spatial Alignment (VSA) enables simultaneous training on video and image data from the same scene, complementing spatial information from images into the temporal localization task. This alignment improves TABot's fine-grained spatial understanding of accidents in video contexts. As for the implementation details, we extend the answer of the temporal localization task to include the spatial grounding annotations. For example, the response to the temporal localization task '{*0.30, 0.45*}' may be further extended with '*At the timestamp 0.38, a traffic accident occurs at [0.21, 0.35, 0.87, 0.57].*', enhancing the model's spatial understanding facing videos and improving its fine-grained understanding capabilities.

With the functional tuning as introduced above, TABot is endowed with the capabilities to perform coarse- and fine-grained traffic accident understanding tasks. To further advance the TABot's comprehensive understanding and conversational skills, we draw inspiration from the work of LLaVA (Liu et al., 2024) and generate a multi-round dialogue set based on our textually annotated TAU-106K dataset. Specifically, we utilize the textual captions of the video clips and images as the abstracts to prompt the powerful LLMs such as GPT-4o (Achiam et al., 2023), to conclude the above functional tasks and generate additional accident-oriented dialogue, such as the causes of accidents or prevention suggestions. In our implementation, the open-source Llama3-70B (Dubey et al., 2024) is utilized to produce these dialogues, which are then used for the **instruction tuning** upon TABot, leading to the TABot-Chat model. Through this training process, TABot-Chat gains an integrated understanding of traffic accidents and develops enhanced instruction-following capabilities.

5 Experiments

We set GroundingGPT-7B (Li et al., 2024), a pre-trained general-purpose MLLM with temporal and spatial grounding capabilities, as the baseline model for our TABot. The detailed experimental settings of the two-step approach are described as follows:

Functional Tuning. In this stage, we construct structured single-round queries for the aforementioned accident-oriented visual understanding tasks, resulting in our TABot model. We train LLM and both visual adapters of the GroundingGPT model on the TAU-106K dataset for 3 epochs using 8 × H800 GPUs. The initial learning rate is set to 2e-5 with a batch size of 32, requiring about 20 hours to complete.

Instruction Tuning. To boost traffic-related comprehension and dialogue capability, we extended training with the instruction-tuning dataset generated by LLaMA-70B (Dubey et al., 2024), leading our TABot-Chat model. To avoid catastrophic forgetting, we combine the data from functional tuning and instructing tuning for training. TABot is further trained for 1 epoch on 8 × H800 GPUs for about 9 hours, using the same learning rate and batch size as the functional tuning stage.

For evaluation, the TAU-106K dataset was split into training and testing sets in a 9:1 ratio, ensuring the same distribution of normal/accident instances and scene continuity across both. We evaluate the TABot on tasks including accident recognition, accident description, and temporal localization, as well as spatial grounding at both the image and video levels. The evaluation metrics are as follows:

1) Accident Recognition. Recall, Precision, and F1 scores are used to assess the model's accuracy in distinguishing accidents from normal scenes in both image-level and video-level contexts.

2) Accident Description. BLEU-1 score, Rouge-L F1 score, and BERT F1 score are employed to measure the model's ability to generate coherent and accurate accident descriptions. We further leverage GPT-4o to evaluate and assign scores based on comparing the model's output and the ground truth, referred to as GPT-4 score.

3) Accident Temporal Localization. We reported the Intersection over Union (IoU) between predicted and true temporal intervals, along with Average Precision (AP@30, AP@50, AP@70).

4) Accident Spatial Grounding. We evaluate the model's performance on accident region and object grounding through reporting detection metrics: mean Intersection over Union (mIoU) and Average Precision (AP@30, AP@50, AP@70).

5.1 VIDEO-LEVEL TASKS

Table 1: Experimental results on video accident recognition in traffic scenes. "@A" and "@N" represent the class-wise results on accidents and normal scenes.

Methods	Video Accident Recognition						
	Acc	Rec@A	Pre@A	F@A	Rec@N	Pre@N	F1@N
Video-LLaVA (Lin et al., 2023)	50.20	99.70	50.10	66.69	0.70	70.00	1.39
TimeChat (Ren et al., 2024)	54.65	91.80	52.67	66.93	17.50	68.09	27.84
VTimeLLM (Huang et al., 2024)	50.00	**100.00**	50.00	66.67	0.00	0.00	0.00
GroundingGPT (Li et al., 2024)	50.00	**100.00**	50.00	66.67	0.00	0.00	0.00
Qwen2-VL (Wang et al., 2024b)	72.65	53.46	87.23	66.29	92.08	66.16	77.00
Gemini-1.5-Pro (Reid et al., 2024)	69.61	61.82	74.18	67.44	77.70	66.25	71.52
TABot (Video)	80.95	78.95	84.40	81.59	83.24	77.50	80.27
TABot (Ours)	81.00	78.65	85.10	81.75	83.77	76.90	80.19
TABot-Chat (Ours)	**82.05**	79.70	86.00	**82.73**	84.80	**78.10**	**81.31**

Table 2: Experimental results on video accident description and accident temporal localization.

Methods	Video Accident Description				Accident Temporal Localization			
	BLEU	Rouge	BERT	GPT-4	AP@30	AP@50	AP@70	mIoU
Video-LLaVA (Lin et al., 2023)	22.20	24.81	60.72	26.17	-	-	-	-
TimeChat (Ren et al., 2024)	7.12	18.16	58.77	12.67	23.00	7.90	2.50	18.07
VTimeLLM (Huang et al., 2024)	25.25	23.32	60.84	18.62	0.00	0.00	0.00	0.00
GroundingGPT (Li et al., 2024)	9.77	16.43	55.70	14.00	4.60	2.40	0.90	3.79
Qwen2-VL (Wang et al., 2024b)	15.38	23.64	61.61	39.80	32.91	15.76	5.42	20.75
Gemini-1.5-Pro (Reid et al., 2024)	12.83	19.57	60.79	23.66	13.87	5.14	1.64	9.31
TABot (Video)	54.70	55.79	82.62	54.63	38.20	20.28	9.60	25.16
TABot (Ours)	54.59	57.94	82.31	55.60	**39.44**	20.12	**9.80**	**25.93**
TABot-Chat (Ours)	**55.70**	**58.32**	**83.78**	**55.73**	37.90	**20.70**	7.80	25.33

In this subsection, we present the results on video-level tasks of our proposed models, including TABot, TABot-Chat, and their comparison with several baseline methods: Video-LLaVA (Lin et al., 2023), TimeChat (Ren et al., 2024), VTimeLLM (Huang et al., 2024), GroundingGPT (Li et al., 2024), Qwen2-VL (Wang et al., 2024b), and Gemini-1.5-Pro (Reid et al., 2024). The experiments cover three key tasks: video accident recognition, video accident description, and accident temporal localization.

Table 1 provides the experimental results for video accident recognition. Most baseline models struggle to recognize traffic accidents, with accuracies ranging from 50% to 54.65%. They tend to classify all videos as positive, showing that general-purpose models cannot understand the semantic information related to traffic accidents in videos. Although Qwen2-VL and Gemini-1.5-Pro show some improvement, they still tend to classify the videos as normal, exhibiting a bias toward normal scenes. In contrast, our TABot, trained on our TAU-106K dataset, demonstrates a significant improvement, reaching an accuracy of 80.95% and outperforming all prior methods. Further instruction tuning with multi-round dialogue data, our TABot-Chat variant further presents an accuracy of 82.05% and improved overall performance for both accident and normal scenarios.

For the tasks of video accident description and temporal localization, the performance of our models is detailed in Table 2. TABot excels in generating accurate and contextually relevant accident descriptions, achieving the highest BERT and GPT-4 scores, indicating high semantic alignment with human judgments and conversation preferences. In terms of temporal localization, previous models struggled to pinpoint the occurrence of accidents, and only Qwen2-VL demonstrated a certain capability in fine-grained localization within videos. Our TABot significantly surpasses all existing methods in the video accident description task, establishing a new state-of-the-art (SOTA) in temporal localization performance. However, after instruction tuning, while the TABot-Chat variant shows improved description capabilities, there is a slight decrease in its temporal localization performance. This suggests that the instruction tuning may have introduced a trade-off, improving the language understanding at the expense of precise temporal boundary detection.

Additionally, we evaluated the impact of video-image joint training compared to video-only training. The results show that incorporating image data into video training leads to a minor performance boost across tasks. Nonetheless, the enhancement is not substantial; the marginal gains can be attributed to the inclusion of more conversational data, which enriches the model's contextual understanding. In contrast, as demonstrated in Section 5.2, adding video data to image training yields a significant performance improvement.

5.2 Image-level Tasks

Table 3: Experimental results on image accident recognition and description in traffic scenes.

Methods	Image Accident Recognition							Image Accident Description			
	Acc	Rec@A	Pre@A	F1@A	Rec@N	Pre@N	F1@N	BLEU	Rouge	BERT	GPT-4
MiniGPT4 (Zhu et al., 2023)	64.05	75.57	68.89	72.08	45.73	54.06	49.54	9.63	11.56	45.84	11.67
GroundingGPT (Li et al., 2024)	63.75	79.15	67.45	72.84	39.25	54.20	45.53	7.22	7.81	45.00	21.08
Qwen-VL-Max (Bai et al., 2023)	69.95	87.87	70.48	78.22	41.45	68.23	51.57	4.59	4.27	43.08	28.46
Qwen2-VL (Wang et al., 2024b)	58.35	40.07	83.53	54.16	87.44	47.84	61.84	23.31	24.53	66.12	32.01
Gemini-1.5-Pro (Reid et al., 2024)	80.99	0.00	0.00	0.00	80.99	1.00	89.50	16.28	21.53	64.44	24.54
GPT-4o (Achiam et al., 2023)	63.65	45.44	90.73	60.55	**92.62**	51.62	66.30	4.78	5.18	43.05	35.71
TABot (Image)	77.95	87.80	74.43	80.56	67.26	83.55	74.52	43.93	41.15	74.16	48.22
TABot (Ours)	90.75	94.38	**90.31**	**92.30**	85.58	91.45	**88.42**	48.62	43.31	75.20	55.12
TABot-Chat (Ours)	**90.50**	**94.90**	89.33	92.03	84.48	**92.36**	88.24	**50.28**	**45.67**	**77.26**	**55.73**

Table 4: Experimental results on accident region and object grounding in traffic images.

Methods	Accident Region Grounding				Accident Object Grounding			
	AP@30	AP@50	AP@70	mIoU	AP@30	AP@50	AP@70	mIoU
MiniGPT4 (Zhu et al., 2023)	50.57	34.85	24.67	39.36	70.33	56.65	33.24	49.72
GroundingGPT (Li et al., 2024)	26.55	14.25	7.82	3.84	62.23	49.06	27.34	43.75
Qwen-VL-Max (Bai et al., 2023)	43.73	26.47	12.79	30.72	59.97	45.27	28.25	43.00
Qwen2-VL (Wang et al., 2024b)	60.21	47.52	29.70	43.02	71.66	57.48	35.66	50.38
Gemini-1.5-Pro (Reid et al., 2024)	56.66	37.20	17.42	37.85	46.07	34.99	20.09	31.98
TABot (Image)	79.40	68.97	43.81	57.08	76.74	64.70	38.62	53.78
TABot (Ours)	80.05	**70.03**	**45.52**	**57.83**	**78.05**	**65.86**	**39.88**	**54.95**
TABot-Chat (Ours)	**80.29**	69.87	44.95	57.63	77.64	65.41	39.68	54.78

In addition to the video-level tasks, we also evaluate our proposed models on image-level tasks, including accident recognition, accident description, and accident spatial grounding. The experimental results are presented in Tables 3 and 4, where we compare our models against several state-of-the-artine methods: MiniGPT4 (Zhu et al., 2023), GroundingGPT (Li et al., 2024), Qwen-VL-Max (Bai et al., 2023), Qwen2-VL (Wang et al., 2024b), Gemini-1.5-Pro (Reid et al., 2024), and GPT-4o (Achiam et al., 2023).

Table 3 presents the results of the image accident recognition. Our TABot (Image) model, trained solely on image data, outperforms all baseline models across various metrics, including accuracy, recall, precision, and F1 scores for both anomaly and normal scenes. After incorporating video data during training, TABot further improves upon these results, achieving an accuracy of 90.75% and outperforming all baselines by a significant margin. TABot-Chat, which undergoes instruction tuning, maintains a similar level of accuracy but exhibits a slight decline in other metrics. Table 3 also provides the results for image accident descriptions. Our models excel in generating accurate and contextually relevant descriptions of accidents, as evidenced by the high BERT and GPT-4 scores. TABot-Chat, following instruction tuning, attains excellent values of 77.26 and 55.73. These results demonstrate the superior language understanding and generation capabilities of our models.

Table 4 showcases the results for accident region grounding and accident object grounding. Our TABot significantly outperforms the baselines in terms of AP and mIoU for both accident regions and objects, and our TABot-Chat also maintains a competitive performance after instruction tuning. These results confirm the effectiveness of our models in accurately localizing and identifying accident-related elements within traffic images.

Furthermore, by comparing the performance of TABot (Image) with TABot, we observe significant improvements in accident recognition and description tasks when incorporating video data into the training process. This suggests that the integration of multiple modalities, particularly video and image data, enhances the model's ability to recognize and describe accidents. However, the improvement in spatial grounding tasks is less pronounced, indicating that the primary benefit of video data is the scale-up in the amount of training data, which is particularly effective for tasks requiring richer contextual information.

5.3 ABLATION STUDY

Table 5: Ablation study on the additional training strategies of the **functional tuning**. "AG", "OG" & "TL" denote the AP@50 of accident region grounding, accident object grounding, and temporal localization.

TABot		Image Understanding					Video Understanding			
VSA	NSR	Acc	BERT	GPT-4	AG	OG	Acc	BERT	GPT-4	TL
✗	✗	88.45	75.09	54.28	68.57	64.06	80.50	82.08	55.23	19.30
✗	✓	88.00	74.73	53.82	70.20	64.21	**81.90**	81.72	54.78	18.90
✓	✗	88.60	74.83	53.91	**70.36**	64.55	80.80	82.26	55.53	19.92
✓	✓	**90.75**	**75.20**	**55.12**	70.03	**65.86**	81.00	**82.31**	**55.60**	**20.12**

Video Spatial Alignment (VSA) To further enhance the spatial understanding of accidents when facing video data, we propose a Video Spatial Alignment (VSA) strategy to incorporate spatial grounding tasks in video dialogues. Prior models often aligned temporal features with the LLM but fell short in capturing the critical spatial details. Due to the unified video-image-text joint annotation, we can explicitly incorporate spatial grounding at specific time frames with the video data within the same scenario. As shown in Table 5, our VSA strategy leads to a consistent improvement in the model's temporal localization capabilities, demonstrating its effectiveness in video spatial alignment.

Negative Segment Referring (NSR) To further refine TABot's ability to distinguish between accidents and non-accidents in frame-level localization, we implemented the Negative Segment Referring (NSR) strategy, which incorporates negative sample-based durations to enable contrastive learning. This addition improves the model's overall performance across both image and video tasks by enhancing its capacity to differentiate accident events from normal content, as indicated in Table 5. However, there is a marginal decline in spatial grounding performance, and we attribute this to the model's focus on temporal localization, which may have led to a slight trade-off in spatial understanding. Despite this, NSR effectively strengthens the model's holistic accident recognition capabilities, making it more adept at filtering out false positives and improving temporal localization accuracy in challenging traffic scenarios.

Table 6: Ablation study on the training strategy of the **instructing tuning**.

TABot-Chat		Image Understanding					Video Understanding			
Mixed Data	Task Flag	Acc	BERT	GPT-4	AG	OG	Acc	BERT	GPT-4	TL
✗	✗	84.55	75.44	50.18	68.71	64.52	79.50	82.43	53.32	5.10
✗	✓	85.50	75.59	52.83	69.14	64.76	79.35	82.14	54.51	13.30
✓	✗	88.30	76.56	52.04	69.22	64.11	80.20	83.10	55.40	18.90
✓	✓	**90.45**	**77.20**	**55.73**	**69.46**	**64.96**	**81.25**	**83.51**	**55.73**	**19.50**

Training Strategies for Chat Version In the TABot-Chat model, we observe that directly performing instruction tuning without additional measures leads to a significant drop in the functional metrics achieved in the previous stage. To address this, we took a data-centric approach by: (1) mix the datasets used for Functional Tuning and Instruction Tuning. (2) introduce task flags to specify the target response for the model in a multi-task framework. Without our mixed data and task flags, the model's performance dropped significantly; for example, the accuracy for image accident recognition decreased to 84.55%. As presented in Table 6, based on our training data paradigm, we successfully improve the conversational performance of TABot-Chat while maintaining excellent functional results.

6 DISCUSSION AND CONCLUSION

To advance the exploration of multimodal language learning models (MLLM) for traffic accident understanding, we introduced video-image-text joint dataset TAU-106K, which includes 51.5K video clips and 54.8K images, with high-quality annotations covering coarse- and fine-grained accident-oriented information. Upon our comprehensive dataset, we proposed TABot, a unified MLLM that is compatible with video and image data and can handle various traffic accident understanding tasks including accident recognition, description, temporal localization, and spatial grounding. Our method and dataset lay the foundation for MLLM to infer and understand fine-grained representations of traffic accident scenarios. Our publicly available data and code will facilitate further research on MLLM for traffic accidents. Future work will include more detailed grounding and addressing the hallucination problem.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2682–2690, 2020.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.

Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. Anticipating accidents in dashcam videos. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*, pp. 136–153. Springer, 2017.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023a.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023b.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023d.

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.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021.

Jianwu Fang, Jiahuan Qiao, Jie Bai, Hongkai Yu, and Jianru Xue. Traffic accident detection via self-supervised consistency learning in driving scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(7): 9601–9614, 2022.

Hadi Ghahremannezhad, Hang Shi, and Chengjun Liu. Real-time accident detection in traffic surveillance using deep learning. In *2022 IEEE international conference on imaging systems and techniques (IST)*, pp. 1–6. IEEE, 2022.

Sanjay Haresh, Sateesh Kumar, M Zeeshan Zia, and Quoc-Huy Tran. Towards anomaly detection in dashcam videos. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1407–1414. IEEE, 2020.

Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742, 2016.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024.

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.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Caitlienne Diane C Juan, Jaira Rose A Bat-og, Kimberly K Wan, and Macario O Cordel II. Investigating visual attention-based traffic accident detection model. *Philippine Journal of Science*, 150(2), 2021a.

Caitlienne Diane C Juan, Jaira Rose A Bat-og, Kimberly K Wan, and Macario O Cordel II. Investigating visual attention-based traffic accident detection model. *Philippine Journal of Science*, 150(2), 2021b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024.

Rongqin Liang, Yuanman Li, Jiantao Zhou, and Xia Li. Text-driven traffic anomaly detection with temporal high-frequency modeling in driving videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing*, 30:4505–4515, 2021.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–9. IEEE, 2018.

Hao Wang, Jiayou Qin, Ashish Bastola, Xiwen Chen, John Suchanek, Zihao Gong, and Abolfazl Razi. Visiongpt: Llm-assisted real-time anomaly detection for safe visual navigation. *arXiv preprint arXiv:2403.12415*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9878–9888, 2021.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.

Yajun Xu, Chuwen Huang, Yibing Nan, and Shiguo Lian. Tad: A large-scale benchmark for traffic accidents detection from video surveillance. *arXiv preprint arXiv:2209.12386*, 2022.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 273–280. IEEE, 2019.

Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. Dota: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):444–459, 2022.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Rixing Zhu, Jianwu Fang, Hongke Xu, and Jianru Xue. Progressive temporal-spatial-semantic analysis of driving anomaly detection and recounting. *Sensors*, 19(23):5098, 2019.

A APPENDIX

In the appendix, we provide additional details and analysis about our work as follows:

- In Section A.1, we present detailed data collection process of the TAU-106K dataset.
- In Section A.2, we provide annotation templates for video accident description for both *Accident* and *Normal* scenes.
- In Section A.3, we present more detailed and fine-grained statistics of the TAU-106K dataset.
- In Section A.4, we deliver instruction-following design paradigms for task-specific single-turn conversation generation.
- In Section A.5, we report how to prompt general-purpose MLLMs to generate multi-turn conversations for traffic accident understanding.
- In Section A.6, we introduce qualitative and quantitative comparisons between our TABot and other general-purpose multimodal language models.
- In Section A.7, we report the comparison of TAU-106K with other accident-specific or general-purpose benchmark.
- In Section A.8, we provide more detailed information about the annotation process and quality control.
- In Section A.9, we address the potential copyright concerns in the dataset.
- In Section A.10, we fine-tune other baselines on TAU-106K and report the results.
- In Section A.11, we present the class-wise performance on the fine-grained categories.
- In Section A.12, we demonstrate the dataset coverage, including the traffic scenarios, objects, and accident categories.
- In Section A.13, we report the experimental results using 7:3 split for training and testing.
- In Section A.14, we evaluate the effectiveness of the reasoning task for the temporal localization task.
- In Section A.15, we further clarify the proper nouns in the experiments.
- In Section A.16, we analyze the potential bias on the binary classification task of accident recognition.

A.1 DATA COLLECTION PROCESS

The video clips collected by previous benchmarks are accident-centric and have been trimmed to single-scene video clips, ensuring that each video clip contains only one traffic accident, without scene changes. However, the raw videos crawled using the keywords, such as *traffic accident*, *road accident*, *dashcam*, *surveillance camera*, etc., are diverse but contain noise. Some are unrelated to traffic accidents, while others are compilations of multiple distinct traffic accident clips.

In order to extract high-quality traffic-related videos from social media platforms, we initially employed the scene change detection library PySceneDetect to automatically segment the raw crawled videos into individual clips, specifically identifying and isolating scene changes. Then, we manually filter out irrelevant and low-quality video clips, such as those with blurred pixels, or traffic-unrelated content, ensuring that each video clip contains only one traffic accident event. In particular, when selecting clips, we only retain video segments depicting clearly visible and severe traffic accidents, such as collisions, scrapes, and pedestrians falling. On the other hand, we discard videos showing traffic violations such as wrong-way driving, speeding, and illegal lane changes, which did not result in accidents according to video contents. It is noteworthy that the video clips capturing the normal traffic scenes are also included in the dataset, as they are essential for training the model to distinguish between normal and abnormal traffic scenarios. For video clips that do contain traffic accidents, annotators must additionally verify the quality of the clip's trimming. If there are redundant frames at the beginning or end due to imprecise trimming, annotators need to provide additional timestamps for the start and end of the core accident segment to ensure the quality of the dataset.

Overall, we collect 106K media data, including 51,544 video clips and 54,767 images sourced directly or extracted from the video clips. In detail, our dataset comprises

- 13,536 video clips and 34,968 images from 9 open-source traffic benchmarks:
  - 4,672 videos and 9,502 images from the DoTA dataset (Yao et al., 2022);
  - 4,464 videos and 6,010 images from the CCD dataset (Bao et al., 2020);
  - 1,923 videos and 3,799 images from the DADA dataset (Fang et al., 2021);
  - 1,727 videos and 2,348 images from the DashCam dataset (Chan et al., 2017);

  - 352 videos and 628 images from the TAD-1 dataset (Lv et al., 2021);
  - 208 videos and 546 images from the TAD-benchmark dataset (Xu et al., 2022);
  - 105 videos and 207 images from the Drive-Anomaly106 dataset (Zhu et al., 2019);
  - 56 videos and 348 images from the RetroTrucks dataset (Haresh et al., 2020);
  - 29 videos and 58 images from the TrafficS dataset (Ghahremannezhad et al., 2022);
  - 9,674 images from SUTD dataset (Xu et al., 2021);
  - 914 images from the CADP dataset (Shah et al., 2018);
  - 713 images from the TaskFix dataset (Juan et al., 2021a);
  - 221 images from YouTubeCrash dataset (Juan et al., 2021b);
- and 38,008 video clips and 19,799 images from social media platforms:
  - 29,364 videos 12,345 images from *Youtube*;
  - 7,577 videos and 6,476 images from *BiliBili*;
  - 1,067 videos and 978 images from *TikTok*.

A.2 Annotation Templates for Accident Description

The annotation process is conducted by human annotators who watch the video clips with labeled *Accident* and describe the accident event in detail. The accident description is required to include the traffic scenario (urban, highway, etc.), the objects involved in the accident (vehicles, pedestrians, etc.), and the nature of the accident (collision, scrape, etc.). Beyond the basic accident description, we also task the annotators to infer the possible causes of the accident, such as traffic rule violations or improper driving behaviors.

To ensure the quality and consistency of the accident description, we provide the annotators with two detailed template designed for *dashcam* and *surveillance camera* footage, respectively, as shown below:

`Description = Footage Source + Traffic Scenario + Cause of the Accident + Content of the Accident + Aftermath`

> Current vehicle is driving on (***Dashcam***) / The surveillance camera captures (***Surveillance camera***) the road of [***TODO***: *the traffic scenario*]. At t_{start}, since [***TODO***: *cause of the accident*], [***TODO***: *the content of the accident, including the nature of the accident and the objects involved*] , at t_{end}, the accident concludes with [***TODO***: *the aftermath of the accident*].

As shown in the template, the annotators are tasked to fill in the square brackets "[*TODO: ...*]" according to the specific content of the video clip. Additionally, the experienced annotators are encouraged to describe the weather conditions, especially when the weather conditions are extreme, such as heavy rain, snow, or fog, which may contribute to the occurrence of the accident.

For normal video clips, since there is no accident event, the annotators are required to label the following simplified template:

`Description = Footage Source + Traffic Scenario` + Content of the Traffic Scene

> Current vehicle is driving on (***Dashcam***) / The surveillance camera captures (***Surveillance camera***) the road of [***TODO***: *the traffic scenario*]. All traffic is normal and no accident occurs in this video clip.

With the help of these templates, the annotators can provide detailed and consistent descriptions of the accident events, ensuring the quality and reliability of the dataset.

A.3 Data Statistics

Video Duration Distribution. As shown in Figure 5, 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.

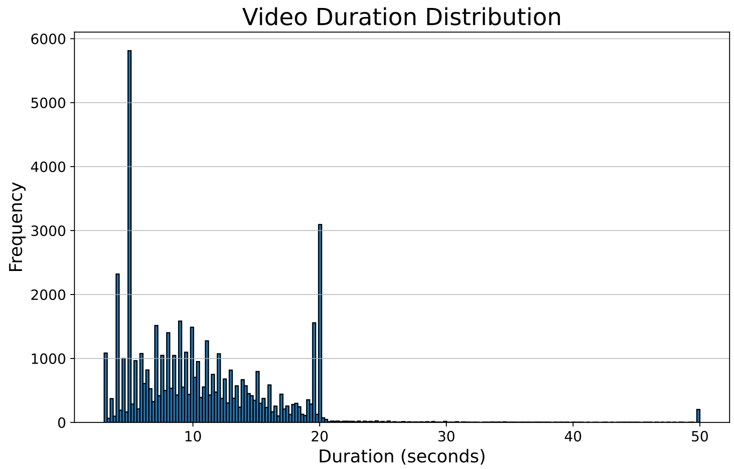


Figure 5: Video duration distribution of the accident-related video clips.

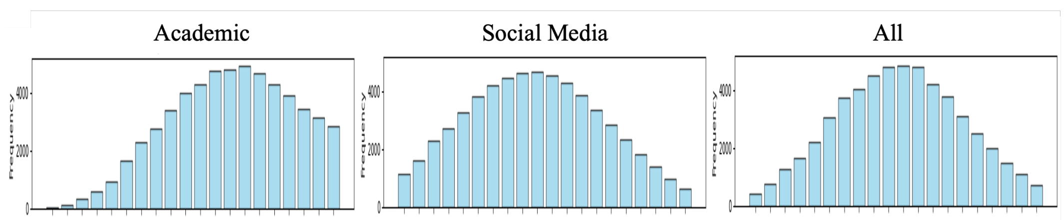


Figure 6: Temporal intervals distribution of the accident events in the video clips.

Temporal Localization Distribution. To statistically analyze the temporal localization distribution of the accident events in the video clips, we visualize the annotated temporal intervals of the accident events, as shown in Figure 6. The accident temporal intervals derived from previous benchmarks are relatively concentrated, with the majority of the accident events occurring within the later episode of the video clips. In contrast, among the crawled video clips, the accident occurrences are more dispersed, following a Gaussian-like distribution. The crawled accident-related video clips form a more balanced distribution of accident temporal intervals, which avoids the bias towards specific temporal segments and ensures the MLLMs to authentically learn the temporal characteristics of the accident events.

Accident-Involved Object Spatial Distribution. The accident-involved object distribution in the images is visualized in Figure 7. Specifically, we respectively cluster the bounding boxes of accident-involved objects in the images according to different footage sources, including dashcam (dashcam-nonself and dashcam-self), surveillance camera, and all sources. When the video clips are captured by dashcams and the dashcam owners are involved in the accidents, the accident-involved objects are concentrated in the location of the front of the current vehicle, mainly the bottom half of the images, which indicates that collisions of the current vehicle with other objects are the primary accident scenarios. In contrast, when the dashcam owners are not involved in the accidents, the accident-involved objects are distributed more evenly across the images with a blank area in the bottle, complying with the real-world traffic scenes where the dashcam owners are bystanders. The accident-involved objects in the surveillance camera footage are more dispersed, and spread across the entire image, which suggests the diversity and complexity of the accident occurrences captured by surveillance cameras. Since the majority of video clips are captured by dashcams, the overall distribution of accident-involved objects in the images is biased towards the bottom half, which is consistent with the dashcam footage characteristics.

A.4 Single-turn Conversation 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.

Figure 7: Accident-involved object distribution in the images.

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"]
```

Single-turn: Temporal Localization. Besides the diverse question set, we also 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 as follows:

```
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
```

The statistics of the generated conversation pairs based on our dataset are summarized in Table 7.

A.5 Multi-tune Complex Dialogue Generation

To generate multi-turn conversations for traffic accident understanding, we design a set of instructions to guide the general-purpose multimodal language model, such as Llama3 in our work, to generate multi-turn conversa-

Table 7: 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 Conprehension	70K	Multi-round conversation
All	332K	-

tions. 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. To achieve this, 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 8.

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. Two examples of the structured caption and the generated multi-turn conversations are presented in Table 9 and Table 10.

```
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 8: 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 examples in Table 9 and Table 10 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 9: One example used in in-context-learning to construct visual conversation data.

Captions
This video is sourced from a dashcam footage.
In the video, the current vehicle is not involved in the accident.
Current vehicle driving on the road, at 0.23 normalized timestamp, a white car in a straight line collides with a three-wheeler waiting in front, at 0.26 normalized timestamp, white car and three-wheeler parked in place.
At 0.233 normalized timestamp, a white car in a straight line located at [0.000, 0.439, 0.210, 0.772] and a three-wheeler waiting in front located at [0.000, 0.481, 0.040, 0.764] were involved in the accident.
accident_timesteps: {0.23, 0.26}
accident_objects:
  white car in a straight line: [0.000, 0.439, 0.210, 0.772]
  three-wheeler waiting in front: [0.000, 0.481, 0.040, 0.764]
accident_region: [0.000, 0.439, 0.210, 0.772]

Conversation
Question:
What types of vehicles were involved in the collision according to the dashcam footage?
Answer:
The collision involved a white car and a three-wheeler. The white car was in motion, driving in a straight line when it collided with the stationary three-wheeler.
===
Question:
Can you tell me the duration of the accident as seen in the video?
Answer:
The accident as visible in the dashcam footage begins at the normalized timestamp of 0.23 and concludes with both vehicles stopping by 0.26.
===
Question:
Where in the video frame did the collision between the white car and three-wheeler occur?
Answer:
The accident took place at the left edge of the video frame, with specific coordinates of [0.000, 0.439, 0.210, 0.772], where the two involved vehicles came into contact and where they subsequently stopped.
===
Question:
Did the current vehicle have any involvement in the accident shown in the video? ===
Answer:
No, the current vehicle from which the dashcam footage is obtained was not involved in the accident. The collision occurred ahead of the current vehicle and was captured from this perspective.
===
Question:
What actions did the vehicles involved in the accident take immediately after the collision?
Answer:
Following the collision, the dashcam indicates that both the white car and the three-wheeler came to a halt in their respective positions. The action is inferred to occur within a few fractions of a second as per the timestamps provided.
===
Question:
Based on the dashcam footage, can you provide any insights into the cause of the accident?
Answer:
The video evidence shows that the white car collided with the three-wheeler while driving in a straight line. Given that the three-wheeler was stationary and located in close proximity to the white car's trajectory [0.000, 0.439, 0.210, 0.772], it suggests that the driver of the white car may have failed to notice or adequately respond to the three-wheeler's presence. Potential contributing factors could include inattentiveness, distracted driving, or insufficient braking distance. However, without additional information such as the speed of the white car, environmental conditions, or the behavior of both drivers before the collision, a full determination cannot be made solely from the footage.

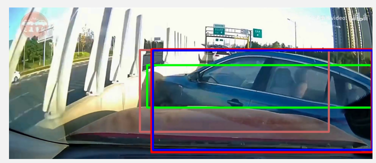
Table 10: One example used in in-context-learning to construct visual conversation data.

A.6 QUALITATIVE AND QUANTITATIVE COMPARISON

Table 11: Ablation study on the training strategy of the **instructing tuning**. 'LoRA' denotes using full parameter training for functional tuning and using LoRA (Hu et al., 2021) for instruction tuning.

Mixed Data	Task Flag	LoRA	Image Understanding Acc	BERT	GPT-4	AG	OG	Video Understanding Acc	BERT	GPT-4	TL
✗	✗	✗	72.75	65.94	48.37	66.12	62.49	68.30	72.84	50.16	4.60
✗	✗	✓	84.55	75.44	50.18	68.71	64.52	79.50	82.43	53.32	5.10
✗	✓	✓	85.50	75.59	52.83	69.14	64.76	79.35	82.14	54.51	13.30
✗	✓	✗	86.15	76.13	53.41	69.27	64.88	80.05	82.79	54.94	16.50
✓	✗	✗	88.65	76.77	54.19	69.15	64.06	81.05	83.38	55.36	20.10
✓	✗	✓	88.30	76.56	53.04	69.22	64.11	80.20	83.10	55.40	18.90
✓	✓	✓	90.45	77.20	54.78	69.46	64.96	81.25	83.51	**56.03**	19.50
✓	✓	✗	**90.50**	**77.26**	**55.73**	**69.87**	**65.41**	**82.05**	**83.78**	55.73	**20.70**

Question [RD]:
Can you identify a traffic accident in this photo? 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 (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 find and 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 (Ours): [0.396, 0.283, 1.000, 0.939]

Figure 8: One qualitative example of the image-level task.

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 (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 (Ours): [0.321, 0.319, 0.878, 0.937]

Figure 9: One qualitative example of the image-level task.

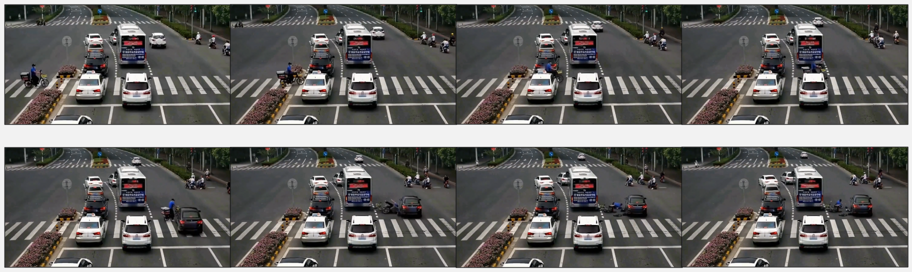
Question [RD]:
Does this video document any vehicular collisions or crashes? Please respond with a 'Yes' or 'No'. Following that, describe the image in one sentence.

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 (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]:
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: {0.39, 0.75}
Qwen2-VL: {0.25, 0.75}
Gemini-1.5-Pro: {0.80, 0.95}
TABot (Ours): {0.39, 0.72}

Figure 10: One qualitative example of the video-level task.



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 (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 (Ours): {0.56, 0.84}

Figure 11: One qualitative example of the video-level task.

A.7 Comparison with Existing Datasets

We provide a more comprehensive comparison between TAU-106K and other datasets, focusing on key features such as size, domain, annotation types, and the characteristics of the textual captions. Here we report the comparison of TAU-106K with other accident-specific or general-purpose benchmarks in Table 12.

Table 12: The comparison of TAU-106K with other accident-specific or general-purpose benchmarks. In the "Annotations" column, 'CLS' indicates Accident Categories, 'TL' indicates Temporal Annotation, 'Bbox' indicates Object Grounding Annotation, 'CAP' indicates Caption Annotation, and 'QA' indicates Question-Answer Pairs.

Dataset	Years	Domain	# Videos	Annotations	# Words per caption	Avg. Duration
Dashcam	2016	Traffic	3,000	TL;	-	5.0 seconds
A3D	2019	Traffic	1,500	TL;	-	8.5 seconds
CCD	2021	Traffic	1,500	TL;	-	5.0 seconds
TAD	2021	Traffic	500	CAP; TL;	-	35.8 seconds
DADA	2021	Traffic	200	CAP; TL; Driver Attention	-	11.0 seconds
SUTD-TrafficQA	2021	Traffic	10,080	QA pairs	-	13.6 seconds
DoTA	2022	Traffic	4,677	CAP; TL; Bbox	-	15.6 seconds
CAP	2023	Traffic	11,727	CAP; TL; Fixed-fourm CAP	6.3 words	6.2 seconds
TAD-106K	2024	Traffic	51,544	CAP; TL; Bbox; Free-Form CAP	32.1 words	10.3 seconds
Charades-STA	2017	Daily activities	9,848	TL; Free-Form CAP	6.3 words	31 seconds
DiDeMo	2017	Open	10,464	TL; Free-Form CAP	7.5 words	30 seconds
ActivityNet-Captions	2017	Open	19,209	TL; Free-Form CAP	13.5 words	180 seconds

According to the comparison, our TAU-106K is the largest dataset in terms of the number of videos and the variety of annotations, supporting a wide range of tasks including accident detection, temporal localization, accident grounding, and accident comprehension. In particular, benefiting from our manual annotation process that is labor-intensive yet worthy, the labeled free-form accident captions in TAU-106K are much more diverse and detailed than other datasets, achieving a largest average length of 32.1 words per caption. This makes TAU-106K a valuable resource for training and evaluating accident-aware models in traffic video understanding.

A.8 Annotation Process and Quality Control

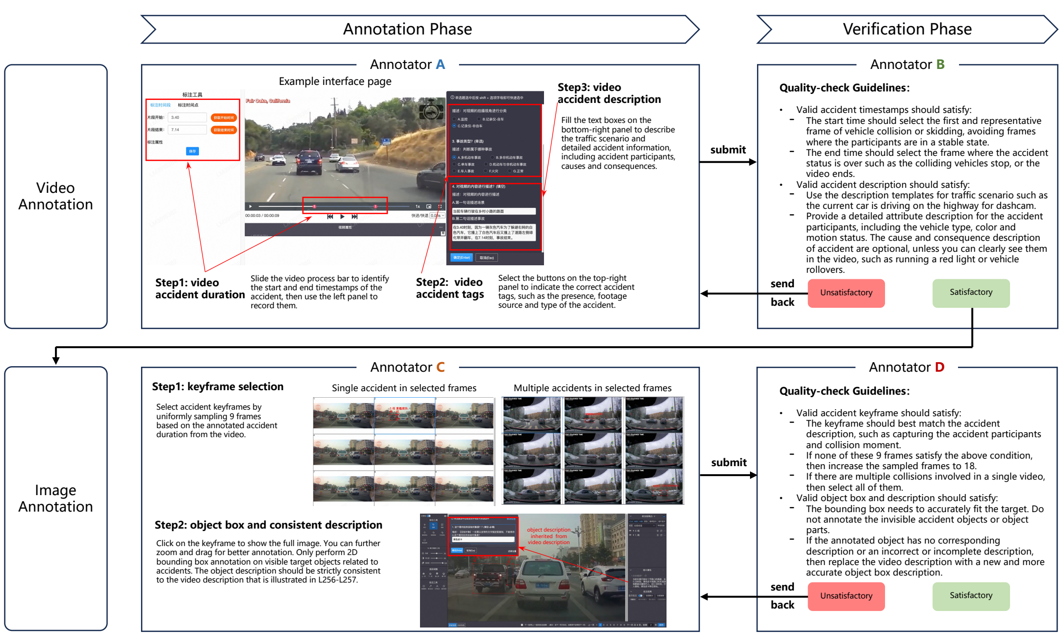


Figure 12: The annotation process and quality control of the TAU-106K dataset.

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. 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 12. The annotation example is presented in Figure 13.

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.

Figure 13: The video annotation examples of different types of accidents.

A.9 Copyright Issues of Dataset

Our data source is from open-source academic benchmarks or public platforms. To avoid copyright and face privacy issues, we only provide video URLs and annotation data to the corresponding raw videos. To facilitate data download, we will also provide an open-source toolkit to access and download the related videos. This is a common practice in prior literature when publishing multimedia datasets, e.g. Youtube8M, HD-VG-130M, and Panda70M.

A.10 Fine-tuning on Other Baselines

We extend our experiments to include additional MLLMs fine-tuned on our proposed TAU-106K dataset, to highlight the significance and quality of our dataset. In detail, the very recent MLLMs, Video-LLaMA-2 and Qwen2-VL, are selected as the alternative baselines for comparison, and we report the results of accident understanding tasks after fine-tuning on the TAU-106K dataset.

Table 13: 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

The results are reported in the Table 13. When fine-tuned on our TAU-106K dataset, the Video-LLaMA-2 and Qwen2-VL models achieve a significant improvement in all tasks, demonstrating the effectiveness of our proposed training recipes and the quality of our dataset. Notably, the CAP performance of Video-LLaMA-2 is even better than our proposed TABot model, and we conjecture that the performance gap is due to the enhanced captioning capabilities of Video-LLaMA-2 pre-trained on a larger-scale general-purpose dataset. After fine-tuning, the performance gap between Video-LLaMA-2 and GroundingGPT is narrowed, which indicates the importance of fine-tuning on the target dataset and the specificity of accident understanding tasks. In particular, the overall performance of our proposed TABot model based on GroundingGPT is still superior to fine-tuned Video-LLaMA-2, especially 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. We conjecture that the competitive performance of the pre-trained Qwen2-VL is due to its advanced architecture and pre-training strategies, and maybe some accident-related or traffic-related data in the pre-training dataset. Even so, the performance improvement after fine-tuning on TAU-106K demonstrates the necessity of fine-tuning on the domain-specific dataset, highlighting the effectiveness and quality of our comprehensive TAU-106K dataset.

A.11 Data Imbalance in Fine-grained Categories

Table 14: 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

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 a table showing 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.

A.12 Dataset Coverage Analysis

We have conducted a detailed analysis of the dataset coverage to demonstrate its diversity and representativeness. In detail, our dataset covers a wide range of traffic scenarios, objects, and accident categories:

- **Traffic Scenarios**: urban streets (49%), intersections (19%), country roads (17%), highways (12%), and other traffic scenes (3%) like parking lots, pedestrian crossings, gas stations, tunnels, roundabouts, etc.
- **Objects:** cars (58%), trucks (12%), electric bikes (11%), pedestrians (5%), vans (3%), bicycles (3%), buses (2%), guardrails (2%), motorcycles (2%), and other objects (2%) like traffic lights, trees, and animals, etc.
- **Accident Categories**: 5 pre-defined categories: multi-motor-vehicle accidents (MMV, 59%), motor-vehicle & non-motor-vehicle accidents (MV&nMV, 18%), single-motor-vehicle accidents (SMV, 17%), vehicle & pedestrian accidents (V&P, 4%), and multi-non-motor-vehicle accidents (MnMV, 2%).

A.13 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 and robustness to overfitting. The results are reported in Table 15. The model's performance remains consistent across different tasks with a slight decrease in accuracy, demonstrating its robustness to different train/test splits.

Table 15: 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

A.14 Effectiveness of Reasoning Tasks

In the application of MLLMs to traffic accident understanding, the most critical task is to achieve precise temporal localization of accidents in videos. The labeled reason caption is a portent of the content of the accident, which makes accident detection and localization more trackable. Here we added an experiment to show the effectiveness of the reasoning caption on the temporal localization task:

Table 16: The ablation study of reasoning captions on the temporal localization task.

Model	AP@30	AP@50	AP@70	mIoU
TABot	39.44	20.12	9.80	25.93
- Reason Caption	34.20	16.90	6.60	21.67

As shown in the table, the removal of reasoning captions leads to a significant performance drop (4.26% on mIoU) on the temporal localization task, validating our claim that reasoning captions serve as valuable cues for accident understanding. And our future work will focus on developing more reasoning tasks based on the reasoning captions in TAU-106K, to achieve the accident forecasting and causality analysis tasks.

A.15 Clarification of Proper Nouns in Experiments

Video Spatial Alignment: Video Spatial Alignment is our novelty in dialogue data construction. GroundingGPT's training data for image tasks emphasizes image understanding and spatial grounding, while for video tasks, it only focuses on video understanding and temporal localization. However, we argue that spatial understanding is equally important when performing video understanding and localization. As discussed in Section 4.1 of the paper, we created QA data that links video understanding with spatial understanding. This approach led to significant performance improvements when compared against baseline methods.

Negative Segment Referring: Conventional temporal localization tasks typically involve two types of QA formats:

- The question asks the model to provide the temporal localization boundaries of the accident.
- The question provides a specific time interval and asks the model to judge whether a traffic accident occurred within that interval.

In type 2, typical temporal localization tasks usually contain only positive samples (segments containing accidents) for training. In contrast, when we construct our type 2 data, we include negative samples in the questions, which are normal time intervals without accidents, and include this part of the data for training. By explicitly asking the model to differentiate between these normal segments and accident segments, we enhanced the model's ability to identify traffic accidents more effectively.

Hybrid Data: The purpose of hybrid data design is to address the potential catastrophic forgetting in our two-step fine-tuning framework. The second stage, instruction tuning, can potentially degrade the functional performance of the four downstream tasks (e.g., image accident recognition, image spatial grounding, video accident recognition, and video temporal localization) compared to the first stage. This is similar to the catastrophic forgetting phenomenon. To mitigate this issue, we mixed the data from both stages (similar to adding rehearsal data in continual learning) and used it for instruction tuning. This approach ensures that the model shall improve its conversational capabilities and also retain its functional performance across the four downstream tasks.

Task Flag: The employed task flags are a widely used technique for multi-task LLMs/MLLMs. In our task accident scenarios, our model is capable of addressing three different kinds of tasks in traffic video understanding (i.e., RD: accident recognition & description, SG: image spatial grounding, TL: video temporal localization), and the response of the model shall be completely different when coming to these three kinds of tasks. To enable the model to respond appropriately to different tasks, we incorporated a task flag into the questions. This allowed the model to produce task-specific responses for each task.

TABot (Video): *Stage-1* model and trained only with *video* data.

TABot (Image): *Stage-1* model and trained only with *image* data.

TABot (Ours): *Stage-1* model and trained with *video & image* data.

TABot-Chat (Ours): *Stage-2* model and trained with *video & image* data.

Multi-turn Dialogues: The multi-turn dialogues in Section 5 refer specifically to the data used for instruction tuning. During functional tuning, all the data is single-turn. It is only in the instruction tuning stage that multi-turn dialogues are introduced.

Tasks and Metrics in Video Understanding:

- Video Accident Recognition: Similar to the image-based task, this is a binary classification task that determines whether a given video contains a traffic accident. The evaluation metrics are the same as those used for the image task.
- Video Accident Description: This task evaluates whether the model-generated accident descriptions match the ground truth. The evaluation metrics are identical to those used for the image-based description task.
- Video Temporal Localization: The model predicts the start and end time coordinates of a traffic accident on a normalized 1D timeline. We use IoU to measure the difference between the predicted temporal boundaries and the ground truth, reporting mIoU as the main metric. Additionally, we calculate AP@30, AP@50, and AP@70 in the same way as in the image spatial grounding task.

Tasks and Metrics in Image Understanding:

- Image Accident Recognition: This is a binary classification task that determines whether a given image contains a traffic accident. The evaluation metrics include overall accuracy as well as class-wise recall, precision, and F1-score for positive (accident) and negative (normal) samples.
- Image Accident Description: This task evaluates whether the model-generated accident description accurately matches the ground truth. The evaluation metrics include the BLEU-1 score, ROUGE-L F1 score, and BERT F1 score to measure the similarity between the predicted sentences and the ground truth. Additionally, we leverage GPT-4-Turbo to assess the alignment between the predictions and the ground truth.
- Image Spatial Grounding: The model outputs bounding boxes to localize traffic accidents and related objects within the image. The evaluation metrics include mIoU. Furthermore, we consider a target to be successfully detected if its IoU exceeds 30, 50, and 70, respectively, and report AP@30, AP@50, and AP@70 to evaluate the model's detection success rate under different IoU thresholds.

A.16 Bias in Yes/No Responses for Accident Recognition

In our TAU-106K dataset, 56% of the instances are labeled as "Accident," and 44% as "Normal," resulting in a relatively balanced Yes/No distribution overall. Specifically, in the test set for the video data, which consists of 2,000 video samples, 50% are positive samples (Yes) and 50% are negative samples (No). The model produces 53.2% Yes responses and 46.8% No responses. This indicates that the model's responses to the video data are balanced, showing no significant bias when performing video understanding tasks.

The test set of the image data also contains 2,000 samples, with 55.2% positive samples (Yes) and 44.8% negative samples (No). The model outputted 61.4% Yes responses and 38.6% No responses. Overall, the model shows a very slight bias toward positive samples, but this bias is minimal and remains within an acceptable range. We understand that this is likely due to a slight imbalance in the training data, which is reflected in the test results.