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 technologies, such as 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 methods resort to traditional deep learning models to model the accident features, achieving inferior performance since the lack of comprehension capabilities derived from Large Language Models (LLM).

Recent advancements in LLMs further tuning on visual-language datasets have driven remarkable progress in multimodal large language models (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. For instance, 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 visual representations of accident occurrences differ significantly from general scenes, necessitating the realignment of these visual representations with the semantic understanding of MLLM towards traffic accidents.

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.

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.

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 Traffic Accident Detection (TAD), methods are classified into single-stage (Hasan et al., 2016) and two-stage paradigms (Yao et al., 2019; Fang et al., 2022). Single-stage approaches often rely on frame-to-frame errors, yet they tend to underperform in forecasting non-ego accidents and are sensitive to dynamic backgrounds (Hasan et al., 2016). Two-stage methods first extract visual features from videos with bounding boxes, optical flow, etc., and subsequently apply a TAD model to predict anomalies and deviations (Fang et al., 2022). However, this approach is highly contingent on the quality of feature extraction. 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. SUTD-TrafficQA (Xu et al., 2021) models question-answering and reasoning tasks for traffic accident scenes, although it remains constrained by the closed question-answering datasets. 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 spatial-temporal grounding and reasoning over traffic accident videos, attributable to the training data's bias towards normal scenes and the localized spatial-temporal features of anomalies.

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.

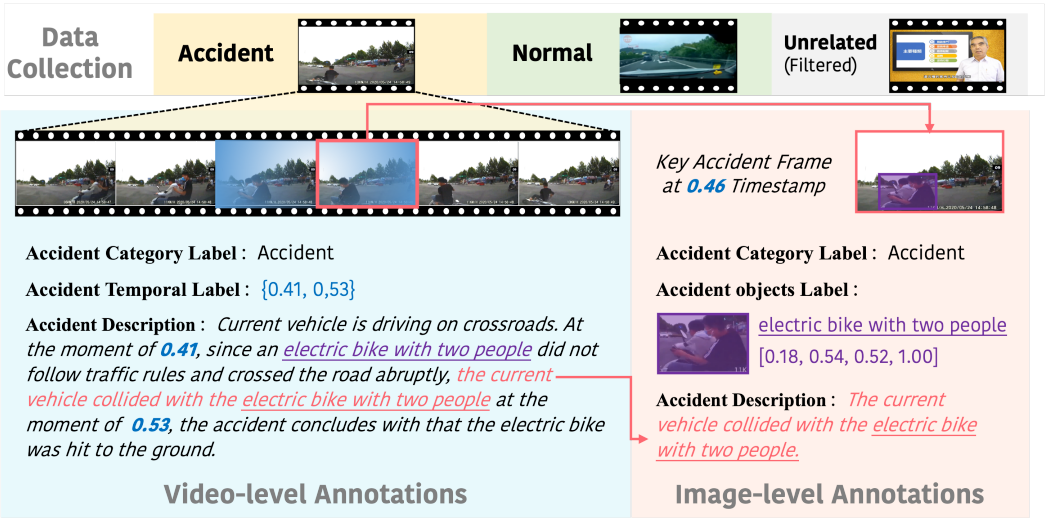


Figure 2: The data collection and annotation pipeline for TAU-106K.

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, hey are often unstructured and lack detailed annotations. For the crawled raw videos, we first crop them into individual clips using scene change detection toolkits, avoiding disruptive scene transitions, and then manually filter out irrelevant or low-quality videos. Consequently, we obtain a collection of 52K 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. To ensure the quality and precision of the Due to arbitrary nature of textual descriptions, this 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 annotations. 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

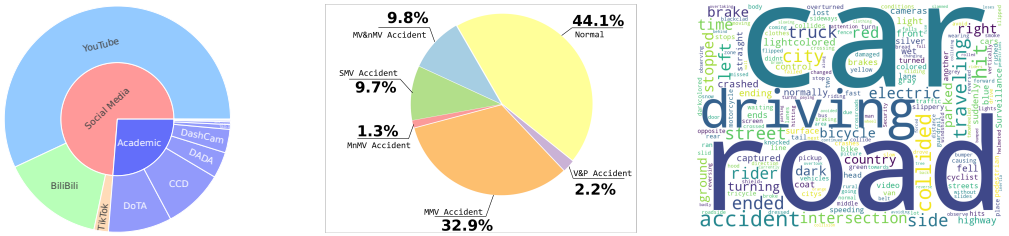
Restricted by the computational overhead and the complexity of video data, MLLMs is incapable of learning fine-grained visual features from video data. To mitigate this, we derive image data from video clips, enabling MLLMs to align accident-related visual information with textual semantics, whose detailed 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 52K 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 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



(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.

of traffic scenarios, accident types, and objects 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 answers 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

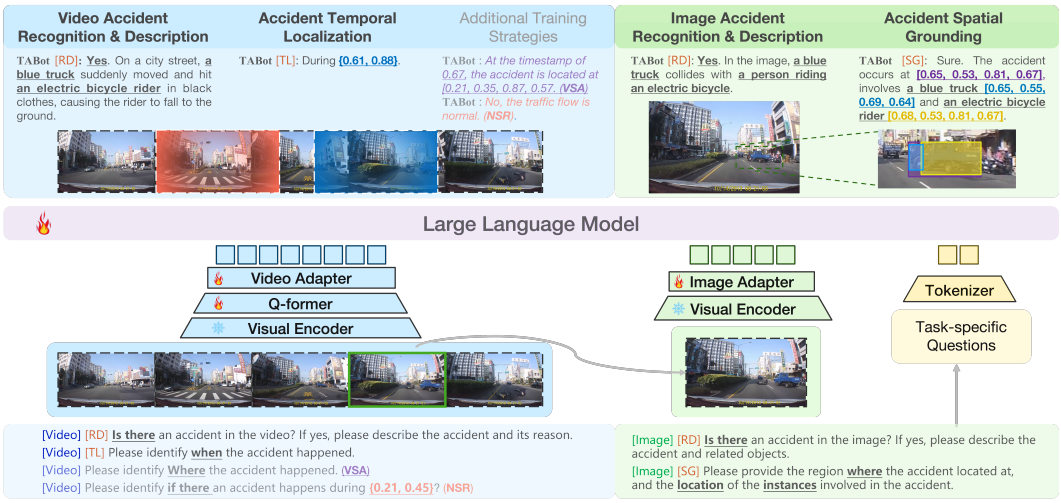


Figure 4: The model architecture and capabilities of the TABot. Additionally, two training strategies designed for temporal localization task, Negative Segment Referring (NSR) and Video Spatial Alignment (VSA), are also illustrated in lighter colors.

accident-free segments before the occurrence of an accident and training the model through accident segment 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 $\times$ 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 a 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 $\times$ 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 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	50.00	**100.00**	66.67	0.00	0.00	0.00
GroundingGPT (Li et al., 2024)	50.00	50.00	**100.00**	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	**9.81**	**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%, indicating that general-purpose models lack the ability to 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, the TABot-Chat variant further enhances performance, resulting in 82.05% accuracy and improved precision, recall, and F1 scores 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 nts 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 accident 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 incorporating 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 52K video clips and 55K 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.

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, 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.

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.

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.

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.

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.