# AccidentGPT: Large Multi-Modal Foundation Model for Traffic Accident Analysis

Kebin Wu[a], Wenbin Li[b] and Xiaofei Xiao
*Technology Innovation Institute, Abu Dhabi, U.A.E.*

Keywords: Traffic Accident Analysis, Multi-Modal Model, Video Reconstruction, Vehicle Dynamics, Multi-Task, Multi-Modality.

Abstract: Traffic accident analysis is pivotal for enhancing public safety and developing road regulations. Traditional approaches, although widely used, are often constrained by manual analysis processes, subjective decisions, uni-modal outputs, as well as privacy issues related to sensitive data. This paper introduces the idea of AccidentGPT, a foundation model of traffic accident analysis, which incorporates multi-modal input data to automatically reconstruct the accident process video with dynamics details, and furthermore provide multi-task analysis with multi-modal outputs. The design of the AccidentGPT is empowered with a multi-modality prompt with feedback for task-oriented adaptability, a hybrid training schema to leverage labelled and unlabelled data, and a edge-cloud split configuration for data privacy. To fully realize the functionalities of this model, we proposes several research opportunities. This paper serves as the stepping stone to fill the gaps in traditional approaches of traffic accident analysis and attract the research community's attention for automatic, objective, and privacy-preserving traffic accident analysis.

## 1 INTRODUCTION

The rapid and accurate traffic accident analysis is critical in enhancing public safety and shaping effective road regulations. The tasks of the traffic accident analysis, varying from accident process reconstruction, responsibility attribution to traffic management and emergency response, are multifaceted and complex. Conventional approaches (Mohammed et al., 2019), relying on eyewitness testimonies, official police documentation, and footage from surveillance cameras (if any), have been the core of the accident analysis for decades. However, these approaches are constrained by intensive manual labor nature, susceptibility to subjective biases, restricted uni-modal outputs, and the privacy concerns emerging from the handling of sensitive data (Al-ani et al., 2023).

The advent of machine learning techniques have begun to boost the field of the traffic accident analysis, presenting enhanced precision and insights. Models are built by learning vast datasets including video footage, sensor data, and textual reports to achieve specific tasks such as accident detection (Ali et al., 2021), accident prediction (Chand et al., 2021), cause identification (Najafi Moghaddam Gilani et al., 2021). Focusing on the accident process reconstruction, from numerical modelling to software simulations of the collisions (Duma et al., 2022) are applied to determine sliced elements (e.g., pre-collision speed, traveled distance, trajectory) during the accident process. Nevertheless, these works are often uni-modal providing useful but fragmented information, while lacking the capacity to integrate and interpret diverse data sources cohesively to reconstruct all details (e.g., process video, vehicles' dynamics) of the accident and automate the post-accident management such as injury assessment, emergency response, report generation, and insurance claim. Furthermore, these traffic applications have been limited in their adaptability, often requiring extensive customization for each specific use case.

As a step further, the recent emergence of the large language models (LLMs) such as LLaMa2 (Touvron et al., 2023) and large multi-modal models (LMMs) such as GPT-4V (Wu et al., 2023b) not only demonstrate the capability to handle multi-modal inputs and outputs, but also underscores a paradigm shift towards task-agnostic learning frameworks, which generate insights across a myriad of tasks without the necessity

[a] https://orcid.org/0000-0003-4492-4152
[b] https://orcid.org/0000-0002-7836-0052



Wu, K., Li, W. and Xiao, X.
AccidentGPT: Large Multi-Modal Foundation Model for Traffic Accident Analysis.
DOI: 10.5220/0012422100003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 943-950
ISBN: 978-989-758-680-4; ISSN: 2184-433X

for task-specific training. The intrinsic versatility of these models is manifested in their capability to generalize learned knowledge and skills across complex multi-task output scenarios. Although most of LMMs focus on dealing with image and text inputs and outputs, recent work (Zhang et al., 2023), (Wu et al., 2023b) shows the possibility to bridge extended list of modalities (e.g., image, text, video, audio, video, etc.) as inputs and produce corresponding outputs of multi-modalities as a response to the prompt. In the context of traffic accident analysis, the technical foundation of such LMM models and techniques brings forward the possibility to build a foundation model to take into account multi-modal inputs and generate outputs for a multiplicity of traffic accident analysis tasks.

However, while the incorporation of multi-modality in traffic analysis presents a promising frontier, it also brings to light significant challenges that have yet to be fully addressed specific to the field:

- Quality and Integrity of Data from Various Sources: In traffic analysis, data can come from a variety of sources, including dashcams, traffic cameras, eyewitness reports, vehicle sensors, and more. The quality and integrity of this data can vary greatly, impacting the accuracy and reliability of the analysis. The quality and integrity of the data from various sources are to be ensured for desired model performance;
- Complexities with Seamless Interpreting and Reasoning: The complexities associated with seamless interpreting and reasoning from diverse traffic accident data and modalities are substantial;
- Model Training and Task-Specific Outputs with Multi-Modal Inputs: The model training and the alignment of task-specific outputs with multi-modal inputs are challenging, which often require intricate customization and tuning;
- Ethical and Privacy Concerns: Ethical and privacy concerns, especially related to the handling and processing of sensitive and personal data, have also been inadequately addressed.

In this work, we propose the idea of AccidentGPT - a foundation model to transform the domain of traffic accident analysis by integrating multi-modal inputs not only to automatically reconstruct accident scenario details but also delivers comprehensive multi-task analysis with a variety of output modalities. The idea extends the existing LLM and LMM solutions with a multi-modality prompt coupled with a feedback mechanism for adaptive task optimization, a hybrid training schema leveraging both labelled and unlabelled data for enhanced model generalization and performance, and a edge-cloud split configuration for data privacy. This paper seeks to tackles the gaps in conventional solutions and unveil the potential of an automated, fast responding, objective, and privacy-preserving traffic accident analysis solution.

The rest of the paper is organized as follows: section 2 discusses the gaps inherent in existing traffic accident analysis approaches. Section 3 outlines our idea of AccidentGPT, highlighting its multi-modal inputs and outputs and multi-task features. Section 4 presents the corresponding research opportunities. Section 5 concludes the paper.

## 2 GAPS IN CURRENT TRAFFIC ACCIDENT ANALYSIS

The traditional approaches and contemporary machine learning techniques, while contributory, present several gaps and challenges that limit their applicability. These gaps highlight the urgent need for a systematic approach to traffic accident analysis.

### 2.1 Data Integration and Analysis

**Manual Efforts:** The traditional approaches (Mohammed et al., 2019) involve substantial manual efforts in post-accident data collection, processing, and analysis. This labor-intensive process is prone to bias of human judgment and thus can lead to inconsistencies and errors impacting the reliability of the analysis. Furthermore, the manual process is time-consuming and the lag in analysis undermines the immediacy of response, impacting emergency services, traffic management, and subsequent investigative processes. Automate the process with timeliness and systematic analysis is one of the key challenge to tackle.

**Privacy Concerns:** Machine learning based approaches (Najafi Moghaddam Gilani et al., 2021) integrates sensitive data sources (e.g., dashcam footage and bystander videos) and raise corresponding privacy and ethical concerns (Butt et al., 2019). These challenges have constrained the scope and depth of accident analysis, leaving a wealth of potentially insightful data untapped. Ensuring the privacy of sensitive data directly improves the effectiveness of the traffic accident analysis.

### 2.2 Model Modality and Generalization

**Model Specialization:** Current machine learning models in the filed of traffic accident analysis are often specialized and task-specific (Chand et al., 2021). These models excel in their designated tasks but face

challenges when exposed to scenarios or data that are different from their training environments. The generalization capability of these models is limited, and this specialization hinders their adaptability and flexibility, reducing their applicability to a diverse range of accident scenarios and conditions. There exists a significant gap in developing models endowed with task-agnostic learning mechanisms that can seamlessly adapt and perform across a variety of tasks and conditions without the need for retraining or extensive customization.

**Uni-Modal Analysis:** Automatic traffic accident analysis on specific tasks (Ali et al., 2021) predominantly relies on uni-modal data sources, such as textual reports or image evidence. These uni-modal approaches lack the capacity to provide a holistic view of accident scenarios, often missing out on crucial contextual and dynamic information that multi-modal data can offer. The lack of versatility to adapt to different data types and analysis requirements leads to a fragmented and compartmentalized understanding of accident scenarios. There is a pressing need for models that can assimilate diverse data sources, understand the intricate interplay of dynamic factors, and provide a comprehensive analysis.

**Output Limitations:** The outputs of existing models (Duma et al., 2022) for traffic accident analysis is typically limited in solo modality (e.g., responsibility, text report) as well. The uni-modality restricts the detailed insights that stakeholders, including investigators, traffic planners, and victims, can extract from the outputs. Furthermore, the lack of interoperability between different analysis systems and technologies can hinder comprehensiveness and intuitiveness of accident analysis across machine learning models. Models are expected to produce multi-modal outputs (e.g., visual representation, numerical dynamics, text reports and news) especially in a multi-task scenario in order to meet diverse stakeholders' requirements (e.g., responsibility attribution, video reconstruction) for a traffic accident analysis system.

In the light of these gaps and challenges, this paper introduces AccidentGPT as a multi-modal foundation model capable of automatically interpreting a diverse range of data modalities and delivering comprehensive, multi-faceted outputs on multiple traffic accident analysis tasks.

## 3 AccidentGPT OVERVIEW

The general idea of the AccidentGPT is depicted in Figure 1, and the model core follows a preprocessing&encoding, alignment&fusion, and decoding process. To revolutionize the field of traffic accident analysis, the joint of use of data from diverse sources is critical to provide robust and insightful analyses. The model inputs can include a) pre- and post-accident site photos, b) CCTV camera recordings, c) dashcam footage, d) statements about the accident process from the involving parties (e.g., drivers, witness) , e) information from Inertial Measurement Units (IMUs) of the movement dynamics during the accident, f) the contextual data containing the accident's related GPS location, time, road signal and condition (e.g., wet, dry, icy), historical traffic data of the traffic sites, and the details related to vehicles and their insurances, and g) most importantly, the task-oriented prompt to instruct AccidentGPT on the desired analysis and outputs. AccidentGPT does not expect the comprehensive set of the input data in every accident analysis, but dynamically adapt to the available data for analysis with partial inputs as similar to the work (Moon et al., 2023). The statements, context and the prompt can be described in a multi-modal fashion (e.g., speech, text, image, etc.), leveraging both textual and non-textual data for a more holistic interpretation.

The model inputs encompass a variety of modalities, including audio, image, video, text, spatial and/or temporal tabular data, and other modalities such as remote sensing spectrum. Each input modality is subject to modality-specific preprocessing steps and encoders (e.g., CLAP (Wu et al., 2023c) for audio, DinoV2 (Oquab et al., 2023) for image, AnyMAL-Video (Moon et al., 2023) for video, IMU2CLIP (Moon et al., 2022) for spatial/temporal series). During the model inference, the preprocessing and encoding process is to be carried out on the users' edge devices for the sake of privacy, and the decoding process is to be performed by the AccidentGPT model on cloud server. The edge machine learning techniques (Li et al., 2023) can be applied to the encoders for computational efficiency and performance; in the case that specific encoders remain significantly demanding in computational resources after model compression, split learning (Vepakomma et al., 2018) can be leveraged by executing only the initial layers of the encoder on the edge devices, while the remaining layers can be offloaded and ran in the cloud environment.

The alignment among different modalities (Girdhar et al., 2023) harmonizes different data sources and ensures diverse modalities are properly integrated and correlated for accurate and cohesive analysis. In this edge-cloud split configuration, the alignment can be flexibly executed in the edge devices and/or the cloud environment. This provides adaptability based on the

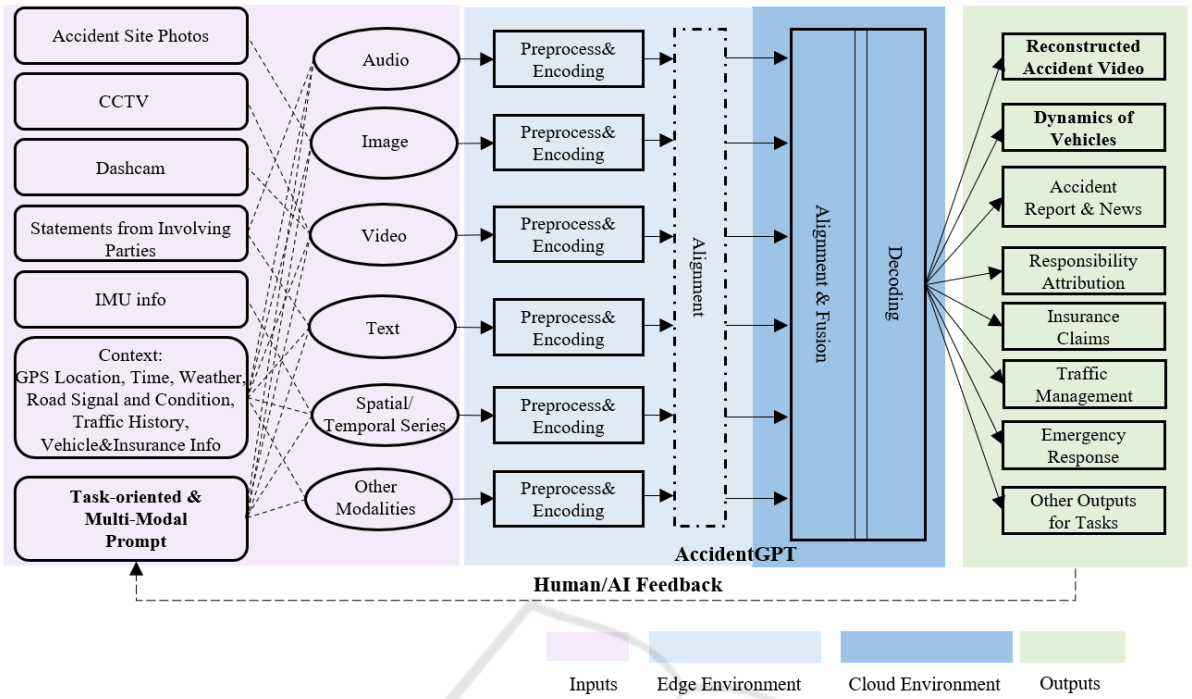


Figure 1: AccidentGPT Overview.

specific requirements of tasks, the computational resources available at the edge, and the desired response time. For simpler alignments on edge, the involving parties can quickly access the data and indicate the temporal and spatial properties of each data item (e.g., pre-accident, in-accident, post-accident). Conversely, for precise alignments involving multiple modalities and the complete input data, the cloud environment can be leveraged with superior computational capabilities to encapsulate the intricate cross-modal interactions among individual components spanning various modalities via representation fusion, coordination and fission (Liang et al., 2023).

After preprocessing, encoding and alignment, the data are fed into the AccidentGPT for modality-specific decoding to automatically generate outputs corresponding to multiple tasks. AccidentGPT targets the following outputs of traffic accident analysis:

- **Reconstructed Video:** this output creates a visual 2D or 3D representation of the complete accident process. The resulting representation offers a temporal and spatially accurate depiction, providing investigators with a sequential understanding of the events leading up to, during, and after the accident.
- **Dynamics of Vehicles:** The dynamics of vehicles are associated with each video frame containing the following information of the involved vehicles: coordinates, velocity, direction, actions of each vehicle involved (i.e., braking, acceleration, turning, no action) and the point of impact and the damage descriptions.
- **Accident Report & News:** The accident report servers as the official documentation and details the sequence of events, involved parties, identified causes, and potential preventive measures. Based on the report, an accident news is tailored for dissemination to news agencies for public awareness.
- **Responsibility Attribution:** This output methodically identifies and attributes responsibilities to involved parties.
- **Insurance Claims:** This output automates the insurance claim assessments by providing a data-driven breakdown of the accident, which highlights damages, identifies potential policy violations, and offers estimations of repair costs based on the severity and nature of the damages.
- **Traffic Management:** This output primarily focuses on the immediate and long-term implications of traffic flow and infrastructure. Post-accident, the output provides real-time recommendations on traffic rerouting, crowd control, and area isolation to ensure minimal disruption and prevent secondary accidents. In the long term, based on recurrent patterns, the output identifies weaknesses in current infrastructure and traffic regulations, and suggest interventions to ensure

smoother and safer traffic flow in the future.

- **Emergency Response:** Emphasizing immediate post-accident actions, this output assists in determining the severity of injuries, potential hazards (e.g., fuel leaks), and the requirement of specialized resources such as medical teams, fire brigades, or specialized rescue units. Additionally, the output provides essential information to first responders, like the number of vehicles involved, hazards, and the access to the accident site. This ensures that the response is not only swift but also tailored to the specific needs of the incident, minimizing harm and damage.
- **Other Tasks:** The adaptability and expansiveness of the AccidentGPT's design based on multi-modality and multi-task modelling makes the model suited for additional task-specific outputs not covered in the primary list. Such flexibility ensures that the model remains relevant and scalable, accommodating evolving traffic safety needs and technology advancements.

Furthermore, the model provides avenues for multi-modal prompt based on reinforcement learning with human feedback (RLHF) (Christiano et al., 2017) or AI feedback (RLAIF) (Bai et al., 2022), ensuring a continuous loop of learning and refinement to improve the model performance with the task-oriented and multi-modal prompt.

# 4 RESEARCH OPPORTUNITIES

## 4.1 Opportunity 1: Multi-Modal Traffic Data Collection and Integration

Gathering and integrating a comprehensive dataset for traffic accident analysis is essential for pretraining.

**Collection and Standardization:** Similar to the paradigm shift in the computer vision domain with the introduction of ImageNet (Deng et al., 2009), the traffic accident analysis field expects a transformation through the establishment of a comprehensive and standardized multi-modal dataset. However, the complexity and multifaceted nature of traffic accidents, along with the discrepancies in data collection methods across regions, make this endeavor challenging. The collaboration and standardization of the data gathering to consolidate such data require synchronized efforts from various stakeholders, in order to ensure the analysis uniformity and solution scalability. Alternatively, leveraging simulation software for autonomous driving such as (Cognata, 2023) can produce standardized datasets with controlled variables, which can act as a base for model training across different scenarios and conditions.

**Data Preprocessing:** Real-world traffic data can be noisy due to various interference sources like weather conditions affecting sensors or low-quality traffic cameras, and thus necessitate extensive proprocessing efforts (e.g., cleaning, filtering). On the other hand, collecting high-quality supervised data can be expensive and, at times, unfeasible. Although semi-supervised learning approaches resort to leveraging unlabeled or weakly labeled data, they still demand specialized filtering procedures (Radenovic et al., 2023). In a multi-modal scenario, the challenges compound even more, as each modality has its own inherent noise and discrepancies. The integration of varied data streams necessitates not only modality-specific preprocessing but also meticulous alignment and synchronization. This is essential to guarantee that inputs from disparate sources accurately represent a singular event. Additionally, an inter-modality harmonization mechanism (Wu et al., 2023a) is crucial to ensure that the composite representation holistically encapsulates the phenomenon under study, with no single modality disproportionately influencing the analysis.

## 4.2 Opportunity 2: Multi-Modal Model Structure and Core Components

Although the general idea of the AccidentGPT follows the encoding, alignment, fusion, and decoding process to generate multi-task multi-modal outputs, no dominant design of model structure exists yet either in existing vision-language pretraining (Liu et al., 2023; Alayrac et al., 2022; Zhu et al., 2023a; Zhu et al., 2023b) or the multi-modality works (Zhang et al., 2023; Wu et al., 2023b), and no singular structure has been conclusively demonstrated to significantly outperform others. In addition to the model structure, the large multi-modal and multi-task model for traffic accident analysis involves four fundamental components that require further research innovations:

**Alignment:** Alignment deals with the synchronization of data from various modalities to ensure that they represent the same event or phenomenon. Although emerging works (Girdhar et al., 2023) demonstrate promising results, the extent and dimensions where the traffic accident data are shared across modalities can lead to: a) non-uniformity across modality alignment (e.g., one-to-one, one-to-many, or not exist at all), and b) long-range dependencies that a particular element from one modality corresponds to an element in another modality that is temporally or spatially distant. Effective alignment methods are expected for temporal matching, spatial calibration and

semantic bridging across modalities.

**Fusion:** Once aligned, the fusion component form a unified representation of the data from different modalities and learn representations that capture the interactions between individual elements spanning various modalities (Man et al., 2023). Due to the fact that multi-modal data are heterogeneous in characteristics, distribution, carried information and relevance toward specific tasks, this component is intrinsically challenging. In the filed of traffic accident analysis, the fusion process becomes even more important due to the critical spatial-temporal relations of the sequencing, timing, and positioning of events and actions that lead up to, occur during, and follow an accident.

**Decoding:** The decoding process produces human-understandable outputs that reflect cross-modal interactions and coherence. While certain modality-specific decoders (e.g., text) are mature and widely used, the AccidentGPT decoding components do not merely construct raw outputs from model's internal representation, but also involves summarization of contents, translation between modalities and creation of new contents (i.e., reconstruction of accident process video). Video generation, as a modality, poses multiple challenges, especially when aiming for high fidelity and temporally coherent sequences. This is yet one of the most challenging but popular research direction. Recent advances (Xu et al., 2023) offer potential solutions, but further research is essential to enhance the granularity, accuracy, and realism of generated video content, particularly in the nuanced domain of traffic accident analysis due to its dynamics complexity, physical consistency and multi-modal integration.

## 4.3 Opportunity 3: Multi-Modal Reasoning

During the AccidentGPT's entire process of traffic accident analysis, reasoning with the fused representation is the key capability of to reconstruct the accident sequence, derive critical insights, and formulate logical conclusions about the incident's dynamics. The reasoning dimension is vast and complex for AccidentGPT involving the culmination of a series of events and interactions among multiple entities. The reasoning function is expected to: a) determine and learn the relationships and interactions within the accident scene, b) understand the contribution of each multi-modal data within the reasoning sequence, c) extrapolate increasingly abstract ideas from the individual pieces of multi-modal evidence. Existing works show contradicting results (Stechly et al., 2023; Huang and Chang, 2023) on how well multi-modal models can perform on reasoning tasks, and yet a further step on the reasoning leveraging external large-scale knowledge and components can yield significant advancements in accurate accident reconstruction and understanding.

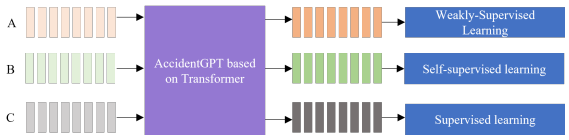

Figure 2: Framework for Pretraining.

## 4.4 Opportunity 4: Data Efficient Training Paradigm

Data from different sources for traffic accident analysis can be categorized into three types: labeled, unlabeled data, and weakly-labeled noisy data (even after preprocessing). Since the data related to traffic accident is scarce in general, it is worthwhile to investigate how to maximize the utilization of (pseudo) supervision or priors in the multi-modal data.

One potential solution is to adopt a combined loss that enables supervised learning for labeled data (Dosovitskiy et al., 2020), self-supervised learning for unlabeled data (He et al., 2022; Chen et al., 2020) and weakly-supervised learning for weakly-labeled noisy data (Radford et al., 2021). Without losing generalization, illustrative examples are shown in Figure 2. In contrast to single strategy training, the hybrid training paradigm is another research opportunity allowing for the comprehensive exploitation of valuable and diverse data information, offering a flexible trade-off between the cost of data collection and the performance of the model.

## 4.5 Opportunity 5: Task-Oriented Multi-Modal Prompt with Feedback

The concept of "prompting" has demonstrated remarkable utility in LLMs (Brown et al., 2020) and LMMs (Lyu et al., 2023). By managing various tasks with task-specific descriptive prompts, attaching them to the input for downstream processing, and then jointly feeding them into a pre-trained, frozen foundational model, this approach offers a unified solution for diverse tasks. However, the full potential of prompting within the realm of large multi-modal models has yet to be fully explored. One of the paramount challenges lies in the vast complexity and diversity of multi-modal data. Unlike textual data where prompts can be relatively straightforward,

defining an ideal prompt in multi-modal scenarios becomes intricate. And the alignment of task objectives with the modality specifics make the process non-trivial, as the misinterpretations or biases can have significant real-world consequences in the traffic accident analysis, ensuring the accuracy, interpretability, and contextual relevance of multi-modal prompts becomes absolutely critical.

Further, the process of feedback plays a vital role in shaping the effectiveness of the prompt. While RLHF (OpenAI, 2023) can provide nuanced insights and guide the model towards desired outcomes, relying solely on it can be costly and time-consuming. On the other hand, recent work on RLAIF (Bai et al., 2022) demonstrates that AI systems can potentially self-regulate, refine, and provide feedback only with the help of human oversight in terms of a list of rules or principles. This presents an intriguing paradigm where multi-modal prompts can be self-optimized and critiqued by a balance between human and AI feedback. The potential evolution of a feedback-driven prompting mechanism could pave the way for more granular and context-aware prompts, thereby enhancing the model's efficacy and responsiveness.

### 4.6 Opportunity 6: Validation Methods and Reliability Metrics

The evolution of multi-modal models in traffic accident analysis opens new avenues for research, particularly in the development of sophisticated validation techniques. Future studies should focus on creating methodologies that can accurately assess and ensure the reliability of outputs from complex systems like AccidentGPT. Another critical area of research is the formulation of robust metrics tailored to multi-modal, multi-task models in high-stake scenarios. These metrics would serve as benchmarks for evaluating the trustworthiness of the model's interpretations, which is paramount in traffic accident analysis.

## 5 CONCLUSION

In this paper, we have introduced AccidentGPT, an innovative foundation model tailored for the intricate domain of traffic accident analysis, leveraging multi-modal data streams and a multi-tasking paradigm. AccidentGPT synthesizes these varied data streams and processes them seamlessly through a unified analytical framework, thereby enabling comprehensive and insightful outputs that span multiple modalities and tasks. The potential of this approach represents a significant paradigm shift, promising to revolutionize the methodologies and tools available for traffic accident analysis.

Our work marks a first step towards an automatic, systematic and privacy preserving traffic accident analysis solution. Research efforts are required to refine these opportunities, fully realize their potential, and rigorously evaluate their performance in real-world scenarios. Future work will focus on exploring the related research opportunities and enhancing the effectiveness of the proposed approach.

## REFERENCES

Al-ani, R., Baker, T., Zhou, B., and Shi, Q. (2023). Privacy and safety improvement of VANET data via a safety-related privacy scheme. *International Journal of Information Security*, 22(4):763–783.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Ali, F., Ali, A., Imran, M., Naqvi, R. A., Siddiqi, M. H., and Kwak, K.-S. (2021). Traffic accident detection and condition analysis based on social networking data. *Accident Analysis & Prevention*, 151:105973.

Bai, Y., Kadavath, S., and et al. (2022). Constitutional ai: Harmlessness from ai feedback.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Butt, T. A., Iqbal, R., Salah, K., Aloqaily, M., and Jararweh, Y. (2019). Privacy management in social internet of vehicles: Review, challenges and blockchain based solutions. *IEEE Access*, 7:79694–79713.

Chand, A., Jayesh, S., and Bhasi, A. (2021). Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. *Materials Today: Proceedings*, 47:5135–5141. International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Cognata (2023). Cognata — Autonomous and ADAS Vehicles Simulation Software.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duma, I., Burnete, N., and Todorut, A. (2022). A review of road traffic accidents reconstruction methods and their limitations with respect to the national legal frameworks. *IOP Conference Series: Materials Science and Engineering*, 1220(1):012055.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. (2023). Imagebind: One embedding space to bind them all.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Huang, J. and Chang, K. C.-C. (2023). Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Li, W., Hacid, H., Almazrouei, E., and Debbah, M. (2023). A comprehensive review and a taxonomy of edge machine learning: Requirements, paradigms, and techniques. *AI*, 4(3):729–786.

Liang, P. P., Zadeh, A., and Morency, L.-P. (2023). Foundations and trends in multimodal machine learning: Principles, challenges, and open questions.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., and Tu, Z. (2023). Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration.

Man, Y., Gui, L.-Y., and Wang, Y.-X. (2023). Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21960–21969.

Mohammed, A. A., Ambak, K., Mosa, A. M., and Syamsunur, D. (2019). A Review of the Traffic Accidents and Related Practices Worldwide. *The Open Transportation Journal*, 13(1):65–83.

Moon, S., Madotto, A., Lin, Z., Dirafzoon, A., Saraf, A., Bearman, A., and Damavandi, B. (2022). Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text.

Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.-F., Murugesan, P., Heidari, P., Liu, Y., Srinet, K., Damavandi, B., and Kumar, A. (2023). Anymal: An efficient and scalable any-modality augmented language model.

Najafi Moghaddam Gilani, V., Hosseinian, S. M., Ghasedi, M., and Nikookar, M. (2021). Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models. *Mathematical Problems in Engineering*, 2021:9974219.

OpenAI (2023). Gpt-4 technical report.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision.

Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., and Mahajan, D. (2023). Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Stechly, K., Marquez, M., and Kambhampati, S. (2023). Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems.

Touvron, H., Martin, L., and et al. (2023). Llama 2: Open foundation and fine-tuned chat models.

Vepakomma, P., Gupta, O., Swedish, T., and Raskar, R. (2018). Split learning for health: Distributed deep learning without sharing raw patient data.

Wu, P., Wang, Z., Zheng, B., Li, H., Alsaadi, F. E., and Zeng, N. (2023a). Aggn: Attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion. *Computers in Biology and Medicine*, 152:106457.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. (2023b). Next-gpt: Any-to-any multimodal llm.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. (2023c). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xu, Z., Peng, S., Lin, H., He, G., Sun, J., Shen, Y., Bao, H., and Zhou, X. (2023). 4k4d: Real-time 4d view synthesis at 4k resolution.

Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., and Yue, X. (2023). Meta-transformer: A unified framework for multimodal learning.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023a). Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023b). Minigpt-4: Enhancing vision-language understanding with advanced large language models.