Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

Vol. 5, No. 4, December, 2024

Visual Instruction Tuning for Drone Accident Forensics

Arda Surya Editya [1], Tohari Ahmad [1*], Hudan Studiawan [1]

[1] Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, Indonesia.

Received 21 June 2024; Revised 12 November 2024; Accepted 19 November 2024; Published 01 December 2024

Abstract

The increasing use of drones in both commercial and personal use has led to a growing demand for effective forensic analysis following drone-related accidents. This research focuses on improving forensic analysis through the development of LLaVAFor, a fine-tuned version of the Large Language and Vision Assistant (LLaVA) model. The objective of this study is to enhance the interpretability of visual instruction tuning for drone accident forensics. LLaVAFor was developed by fine-tuning LLaVA via a specialized dataset of drone accident scenarios. The model's performance was evaluated via the BLEU score, a metric commonly used to assess machine translation and natural language processing models. The results demonstrated that LLaVAFor achieved superior BLEU scores compared with baseline models such as LLaVA, Google Gemini, and ChatGPT. It demonstrates its ability to provide more accurate and contextually relevant analyses. The key innovation in LLaVAFor is its ability to explain forensic findings in the context of drone accidents, making it a valuable tool for investigators. The results show that the model's fine-tuning process on drone-specific datasets enables it to offer detailed, domain-specific insights, improving the accuracy and reliability of forensic analyses in this field. Through these advancements, LLaVAFor represents a step forward in the integration of AI into drone accident investigations.

Keywords: Forensic Analysis; Drone Forensics; LLaVA; Drone Accident.

1. Introduction

The implementation of forensic analysis in drone accidents is an evolving discipline that follows the growing utilization of unmanned aerial vehicles (UAVs) in military [1], commercial [2], and recreational arenas [3]. This methodical process is aimed at dissecting drone-related mishaps to pinpoint causes, contributory elements, and any breaches of regulatory standards or laws. Another main aim is to collect, safeguard, examine, and present data in a manner that holds up in legal proceedings, thereby enhancing drone safety, ensuring regulatory adherence, and fostering accountability [4]. Drone forensics involves an approach that starts with a comprehensive investigation of the accident site and the collection of physical evidence, including the drone, its components, and any other pertinent materials. Analysts also scrutinize data from the drone's onboard systems, such as flight data recorders [5] and GPS logs [6], to piece together the drone's flight path and the events leading to the accident. A critical examination of the mechanical and electronic integrity of drones is performed to identify any failures or design flaws, including a review of software and potential cybersecurity threats that could influence drone performance [7].

Additionally, the operation of the drone is evaluated against existing laws, regulations, and standards to determine compliance and identify any legal violations that may have contributed to the incident. The process culminates in the preparation of detailed reports and, when necessary, the provision of expert testimony in legal settings to outline findings, conclusions, and recommendations [8]. This multidisciplinary approach not only aids in advancing drone safety and regulatory frameworks but also ensures that UAV operations in shared airspaces are conducted responsibly.

* Corresponding author: tohari@its.ac.id

http://dx.doi.org/10.28991/HIJ-2024-05-04-01

Through such forensic analyses, stakeholders are better equipped to identify recurring issues, implement preventive measures, and establish best practices to prevent future accidents [9]. Research on drone forensics is limited by several factors. One key limitation is the diversity of drone models and configurations, which makes it difficult to create a one-size-fits-all forensic approach. Each drone may have unique hardware, software, and communication protocols, posing challenges for standardized analysis techniques. Additionally, the encryption and security measures employed by some drone systems can hinder data extraction, making it more difficult to access crucial evidence. Research on drone forensics conducted by Jain et al. [4] focused on developing a framework to identify and verify drone data. Furthermore, Horsman et al. [8] proposed a framework to perform a forensic investigation of drones. This framework contains three processes for performing forensic investigations of drones: the preparatory, data acquisition, and data analysis phases. In addition, Editya et al. [10] implemented the deep learning technique for drone collision prevention.

In the drone forensic framework, there is an analysis phase that focuses on examining the data collected from the drone to determine the cause of an incident. A study conducted by Editya et al. [6] used transfer learning to classify drone collisions on the basis of frames captured by the drone camera. Their research revealed that the InceptionV3 method performed well in terms of accuracy, precision, recall, and F1 score compared with other deep learning methods, such as MobileNetV2, ResNet, and VGG. In addition, Editya et al. [10] applied optical flow to estimate the movement of an attacking drone and used quivers to visualize this movement. The method helps the investigator determine the cause of the drone incident.

Previous research has demonstrated that deep learning and optical flow can assist investigators in determining the cause of drone accidents. However, these methods have limitations due to incomplete information. The results still require human interpretation, which can lead to misinterpretation, especially if the investigator is fatigued. In recent years, artificial intelligence has made significant advancements, particularly in generating explanations. One promising development is the use of large language models (LLMs), which can produce detailed explanations. LLMs have progressed to the point where they can now interpret images as well. Several LLM methods, such as ChatGPT, Google Gemini, and LLaVA, have been developed for this purpose. Therefore, we propose this LLM-based technique to assist drone accident prediction.

Large Language and Vision Assistant (LLaVA) stands as an advancement from text generation technologies such as ChatGPT, which supports multimodal inputs [10]. It provides a large-scale multimodal model, integrating a vision encoder and Vicuna for versatile visual and linguistic comprehension. Its development and utilization span diverse domains, including medicine [11], scientific research [10], and communication [11]. This versatility positions LLaVA as a potent tool for expansion into emerging fields such as drone forensic investigation. The integration of LLaVA [10] into drone forensic analysis represents a pioneering advancement in how investigators approach the examination and conclusion-drawing processes in drone-related incidents. These methods are designed to enhance forensic investigations by employing advanced natural language processing (NLP) and computer vision technologies. LLaVA has the ability to analyze vast amounts of textual and visual data, including imagery from drone cameras and crucial pieces of evidence that might elude human analysts.

By implementing LLaVA, forensic investigators can automate the tedious and complex process of data sifting. This approach reduces the time required to reach preliminary findings and conclusions. Furthermore, LLaVA can assist in generating detailed reports that synthesize the findings of investigations. The report is presented in a clear and accessible manner suitable for both technical and nontechnical audiences [12, 13].

Figure 1 shows the application of LLaVA in drone forensics, which has been extended to predictive analytics. It can predict potential accidents and operational risks by analyzing image data taken from a drone before an accident. This predictive capability not only aids in the postmortem analysis of drone accidents but also contributes to the proactive identification of safety vulnerabilities. This feature will lead to more robust drone designs and safer flight operations.

Contribution: This paper proposes a fine-tuned model of LLaVA, namely, LLaVAFor, that is optimized to assist in the investigation of drone accident causes. Using the custom dataset obtained from the ColaNet dataset, the model is then fine-tuned. In this paper, we also compare LLaVAFor to other LLMs, such as GPT-4, LLaVA, and Gemini. The model is evaluated via the bilingual evaluation under study (BLEU) metric.

2. Related Work

2.1. Drone Forensics

Drone forensics comprises methods to extract, gather, and scrutinize data from drones to ascertain the reasons behind certain incidents. Numerous studies have been conducted to address various obstacles and improve efficiency in this domain. Horsman et al. [8] proposed a detailed forensic framework suitable for a wide range of drones on the market. This framework consists of three key phases: the initial preparatory phase, in which investigators inspect each component of the drone to determine its usage. The second phase involves the collection of data from these components, which are then stored on a forensic workstation. The final phase focuses on the examination of these data via diverse methods to derive insights and understand the underlying reasons for drone engine failures.

Additionally, a detailed framework dedicated to identifying and confirming sensor log data is outlined in [4]. This research discussed the structure of a drone and suggested a universal forensic framework designed to enhance the process of digital investigations. Barton et al. [14] conducted a study with the goal of reconstructing drone activities, pinpointing the owners or operators, and retrieving data from associated mobile devices. A different study concentrated on examining the data flash and telemetry logs of drones that were custom-built [15]. The researcher investigated these forensic elements originating from ArduPilot, an open-source platform used for building drones. This investigation resulted in the development of a method for gathering important data, analyzing it, and creating an appropriate timeline.

2.2. LLM for Digital Forensics

Several studies have been conducted on the use of a large language model (LLM) for forensics analysis, such as the study conducted by Scanlon et al. [16]. In this study, they evaluate ChatGPT in the cases of investigation, learning, and programming in digital forensics. Many of the limitations identified are consistent with findings from other studies and existing system documentation. In particular, the phenomenon of "hallucination", which nicely disguises the alternative term "incorrect", is a recurring theme. This obfuscation makes the use of ChatGPT in digital forensics a precarious endeavour and underlines the importance of caution and close scrutiny.

Another study conducted by Michelet et al. [17] used several LLM methods to make a digital forensic report. They used qualitative methods, such as ChatGPT and LLAMA, to evaluate each LLM. The conclusion of the research is that LLM can make automated report sections, although the quality of the generated text varies by model. However, several issues, such as model size, generation time, and hallucinations, have been identified as challenges. Furthermore, Piggott et al. [18] integrated large language models (LLMs) into security systems and presented a double-edged sword in the area of cybersecurity. While LLMs can strengthen defenses against cyber threats, they also introduce new risks by empowering adversaries to generate malicious content, discover vulnerabilities, and manipulate perceptions.

In another study, LLMs could be employed for forensic analysis. Wickramasekara et al. [19] utilized a comprehensive framework covering various phases of case analysis, including incident recognition, collection, preservation/acquisition, examination, analysis, and reporting. The findings of the study indicate that while the integration of LLMs into digital forensics is in its early stages, there is clear evidence of their significant potential to increase investigation efficiency. There is a suggestion to explore investment in LLMs throughout the entire forensic process with the aim of improving the productivity and efficiency of investigations.

3. Proposed Method

Figure 1 shows the flowchart of the research methodology through which the objectives of this study were achieved.

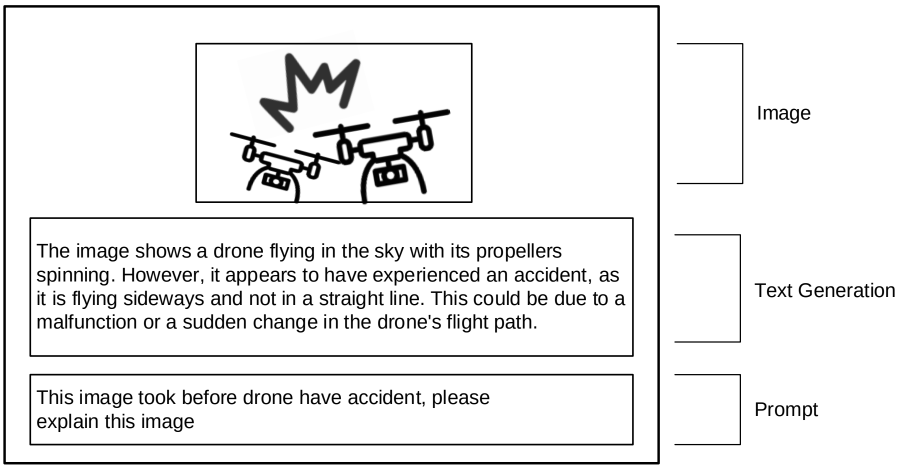

Figure 1. Design of the research workflow

The proposed method in this research is called LLaVAFor (Large Language and Vision Assistant for Drone Forensics). In this method, the LLaVA is fine-tuned via a selected dataset from ColaNet [20]. The dataset contains videos taken from drones that have accidents. The detailed workflow of the proposed method is shown in Figure 2. The method starts with data preprocessing. In this phase, the videos from the ColaNet dataset are converted into labeled images. After that, the labeled images are annotated by a human annotator to create a new dataset. The process continues to fine-tune the LLaVA model via the dataset from the previous step.

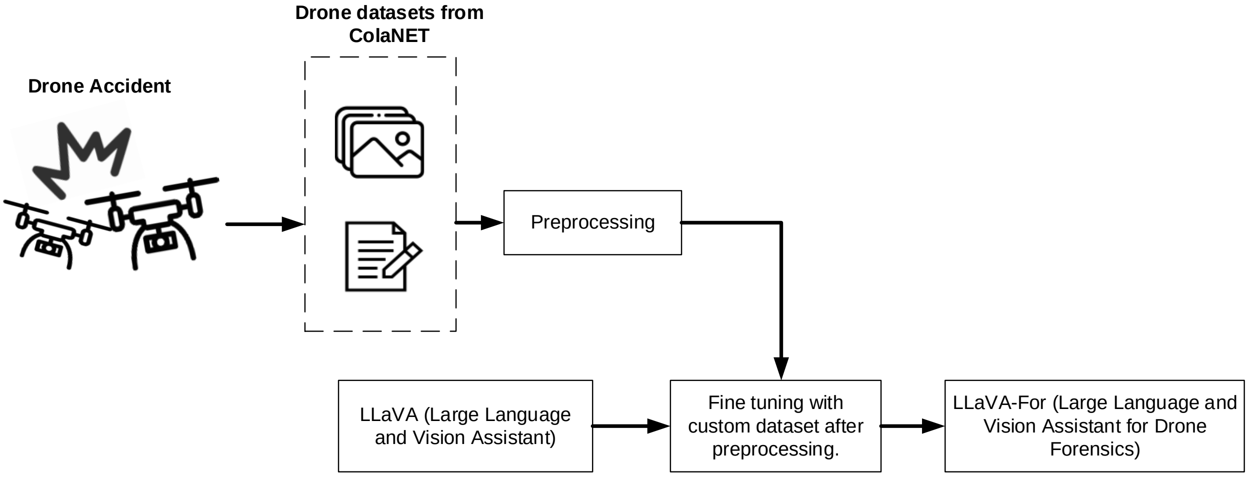

Figure 2. Flow diagram of the proposed method LLaVAFor

3.1. Dataset

The dataset used in this study is sourced from ColaNet, a collection specifically designed for research in the field of drone forensics and related technologies. ColaNet was created by Pedro et al. [20] and made publicly available through their platform at *https://colanet.qa.pdmfc.com/*. This comprehensive dataset consists of 100 video recordings, all of which capture drone accidents viewed from first-person perspective (FPV) cameras mounted on the drones themselves. These videos provide valuable real-world scenarios that are critical for developing and testing forensic analysis methods. This dataset also has a composition as summarized in Table 1.

Table 1. The Composition of the ColaNet Dataset

Accident Type	Total Video	Frames
Collision	42	4746
Attacked	34	3805
Pilot Error	24	2685

The dataset is particularly useful for studies that focus on understanding and reconstructing drone accidents, as it offers direct visual data from the drone's viewpoint. Researchers can leverage these videos to apply deep learning models, optical flow techniques, and other forensic analysis tools to examine drone behavior leading up to accidents, detect anomalies, and identify potential causes. The diversity in the accidents and environmental conditions captured in the ColaNet dataset adds richness to the analysis, making it a key resource for improving the accuracy of forensic methods in drone-related incidents. The availability of this dataset supports advancements in the growing field of drone forensic investigation, helping to address emerging security and legal challenges.

3.2. Data Preprocessing

In this phase, we transform 100 drone accident videos into 11236 annotated images with a format of png, with each dataset having a files size of 53--172 KB. The process commences by entering a timestamp value collected from datasets of drone accidents. Using Algorithm 1, frames are labeled either "normal" or "accident" upon timestamp entry.

Algorithm 1 Conversion algorithm from video to labeled images

```
for iteration = 1, 2, . . . do
   if time ≥ time accident then
     frame = accident
   else
     frame = normal
   end if
end for
```

Figure 3 shows sample frames labeled "normal" and "accident". In this study, we use only the data labeled "accident". These labeled images are important in generating data for LLaVA fine-tuning. In addition, in this process, we take one frame image on each video to speed up the experiments.

To create a dataset that can be used in LLaVA, we have to make a prompt that is related to the image data. The format of the data is shown in Figure 4. The format of the dataset contains several parameters, such as id, image, and conversations. The Id parameter contains the identifier of the dataset, indicating the order of the whole dataset. The image parameter contains the location of the image. For the conversation parameter, there are two sections. The first one is text "from: human", and this text is also known as a prompt. The second is the "from: gpt" parameter. This is the ground truth value, which is written by the drone investigator by looking at the video dataset and describing the drone accident cause. The gpt parameter value is learned via LLaVAFor.

Figure 3. Frame labels from the ColaNet dataset

```
{
    "id": "2",
    "image": "1/2.png",
    "conversations": [
      {
        "from": "human",
        "value": "This image is taken before a drone accident, please
explain this image.\n<image>"
      },
      {
        "from": "gpt",
        "value": "There is an unusual detail in the scene: a hand with
a red glove is visible in the foreground,
             seemingly floating in the air. Drone have attacked by someone
          in music concert."
      }
    ]
}
```

Figure 4. An example of training data for fine-tuned LLaVA

3.3. Large Language and Vision Assistant for Drone Forensics (LLaVAFor)

LLaVA is a model that refers to advanced artificial intelligence systems capable of processing and understanding both text and visual information [10]. These systems combine the capabilities of large language models (LLMs) with computer vision technology. It enables us to analyze, interpret, and generate content that spans both the textual and visual domains.

In the LLaVA system, the large language model component is responsible for processing and understanding text, similar to how models such as OpenAI's GPT series (including GPT-3 and GPT-4) operate. These models can understand context, generate text, answer questions, and more, on the basis of the text data they have been trained on. The vision component of LLaVA adds the ability to process and interpret visual data, such as images, videos, and graphical content. This could involve recognizing objects, understanding scenes, detecting patterns, and even generating visual content on the basis of textual descriptions.

The theoretical approach behind the LLaVA model is multimodal learning, visual-language embedding, attention mechanisms, transformer architecture, and contrastive learning. The concept of LLaVA multimodal learning can be described via an equation that combines the components of language and vision processing. Equation 1 shows the mathematical representation of LLaVA.

$$\mathrm{LLaVA} = \mathrm{LLM} + \mathrm{CV} \tag{1}$$

where LLaVA is defined as the method of improvement of LLM with visual assistance. LLM stands for the large language model, which is responsible for processing, understanding, and generating text. Moreover, CV represents a technology that enables the system to process, interpret, and generate visual content.

In this equation, LLM and CV are integrated to provide a system that can handle both textual and visual information. This integration allows the LLaVA to understand and generate multimodal content, which combines elements of both text and images or videos, leading to a more comprehensive and nuanced understanding and interaction capability. The combination of LLM and CV in LLaVA allows for advanced applications that require nuanced understanding and generation of both language and visual content.

To align visual and textual information, both modalities are first encoded into a vector space. The visual input, such as an image in this context, is a drone frame; this typically involves the use of a convolutional neural network (CNN) or Vision Transformer (ViT) to extract feature vectors, whereas for the textual input, a language model such as BERT or GPT is used. When combining information from images and text, errors can occur. This is handled via a method called the contrastive loss function, which is represented by Equation 2.

$$l = -\log \frac{exp(sim(V,T)/\tau)}{\sum_{j=1}^{N} exp\left(sim\left(V,T_j\right)/\tau\right)} \tag{2}$$

where *sim(V,T)* is the cosine similarity between the visual and textual embeddings. τ is a temperature parameter that controls the smoothness of the distribution. *N* is the number of negative samples. This method can help LLaVA differentiate between similar and dissimilar pairs of data, such as images and text.

To apply the LLaVA method to drone forensic topics, we optimize the model of the original LLaVA by using the drone accident dataset. Figure 5 shows the main concept of how LLaVAFor was developed. The original LLaVA process starts from the large unlabeled dataset, indicating that the dataset used to build this model is not specific for drone forensic; therefore, it is called the “Large Unlabeled Dataset”.

This process continues with building the model of the original LLaVA, which requires high computational resources. After the model has been built, the original LLaVA model is fine-tuned with the "Small Labeled Dataset." This dataset contains drone accident images and is used to fine-tune LLaVAFor.

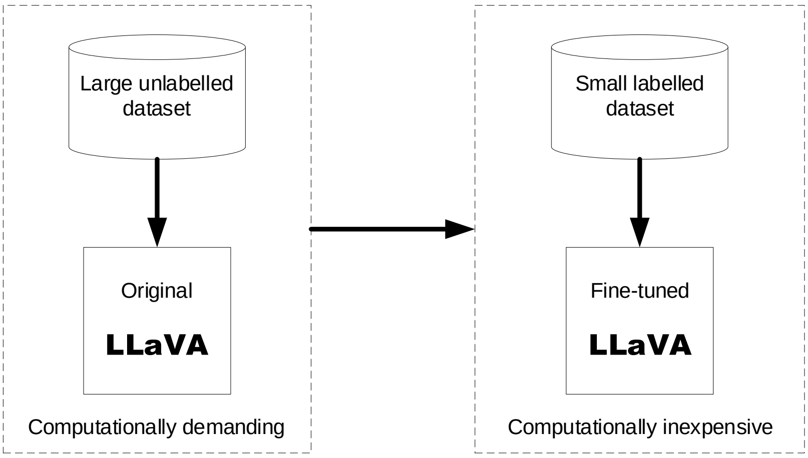

Figure 5. The main concept of LLaVAFor

3.4. Fine-Tuning

Fine-tuning is a process in machine learning where a pretrained model is used. The pretrained model has been trained on a large and generic dataset. This process is further trained or adjusted on a smaller, specialized dataset relevant to a specific task or domain [20]. This process involves making slight adjustments to the model's parameters to adapt its knowledge and capabilities to the nuances and specific requirements of the task at hand. The purpose of fine-tuning is to utilize the broad understanding and capabilities the model has gained during its initial training. We then apply it to perform more effectively on specialized tasks without the need to train a new model from scratch. This approach is efficient because it saves time and resources and can significantly improve the model's performance in specific areas. Therefore, it provides a more accurate and relevant model for targeted applications [21].

Fine-tuning in the Large Language and Vision Assistant (LLaVA) is important for enhancing its performance and specialization in specific tasks or domains, such as drone forensics. This process improves the model's accuracy by modifying it to recognize the context of this field, leading to more relevant and effective outcomes. It also reduces generalization errors, making LLaVA versatile and reliable across different scenarios. Fine-tuning improves the existing pretrained structure of LLaVA, making it a computationally efficient method for adapting the model to new datasets, specifically drone accident images. A representative example of the original LLaVA using the drone accident dataset is shown in Figure 6. LLaVA's result provides a general explanation of the image, but in a forensic analysis context, it cannot be used to establish an analysis.

Supervised fine-tuning involves the adoption of a pretrained LLM for a specific downstream task via labeled data. The fine-tuning data are collected from a set of responses validated from ground truth data. Figure 7 shows the results of the fine-tuned LLaVA. The results show improvements that help explain how the drone had an accident. Additionally, LLaVAFor's results identify components of the drone, such as propellers, and make predictions about the cause of the accident, such as malfunctions or sudden changes in the drone's flight path.

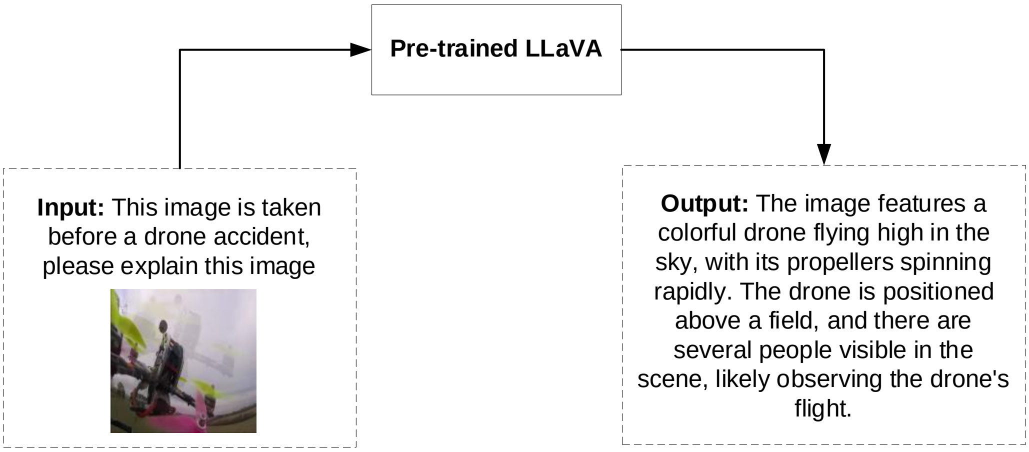

Figure 6. The example of the original LLaVA

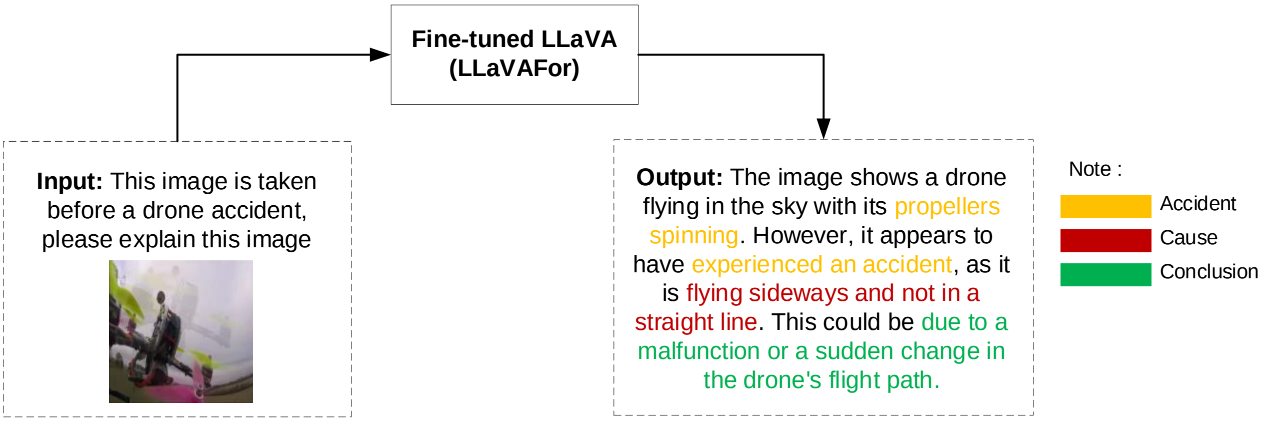

Figure 7. An example of a finetuned LLaVA

3.5. Quantitative Evaluation

To gain a systematic understanding of the performance of LLaVA, we propose a quantitative metric to measure the model’s instruction-following capability on multimodal data. In this assessment, we employed the bilingual evaluation under study (BLEU) metric. BLEU evaluates the accuracy of generated text, known as the candidate, in comparison to a set of reference texts. In tasks involving sequence-to-sequence, there may be multiple correct references for a single candidate [22]. Hence, it is crucial to select references and include all potential references. The BLEU score, a precision-based metric, ranges between 0 and 1, with higher values indicating better predictions. While reaching a score of 1 is unattainable, typically, a score exceeding 0.3 is deemed satisfactory.

The BLEU score evaluates the likeness between machine-translated text and reference translations via *n-grams*, which are consecutive sequences of *n* words. Commonly utilized *n-grams* include unigrams (single words), bigrams (two-word sequences), and trigrams (three-word sequences). It assesses the precision of *n-grams* in the machine-generated translation against the reference translations [23]. This precision is adjusted by a brevity penalty to accommodate translations shorter than the reference translations. Equation 3 illustrates the mathematical formula for computing the BLEU score:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n \, log \, p_n\right) \tag{3}$$

where *BP* represents the brevity penalty, which is a measure given when the number of words in the result differs from the ground truth data. Furthermore, *N* represents the gram used in the measurement process, whereas *n* is the total number of words in the result. The variable *w* is the weight of the gram, which is calculated as the total number of words divided by the number of grams. Finally, *p* represents the precision for the *n-gram*.

In addition to using the BLEU score, we also use the precision parameter. The precision of LLMs is different from the precision of classification tasks. In the LLM context, true positives are all the matching *n-grams* between the candidate and the reference [23]. False positives are the *n-grams* that appear in the candidate but are not present in the reference. This means that precision can be obtained by dividing the number of matching *n-grams* by the total number of *n-grams* in the candidate. Equation 4 illustrates the mathematical formula for computing precision:

$$p_n = \frac{\sum_{C \in \{Candidate\}} \sum_{\{n-gram \in C\}} Count_{Clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{\{n-gram' \in C'\}} Count(n-gram')} \tag{4}$$

where p_n represents the precision for each n-gram. *C* represents the total number of words in the data. "*Candidate*" refers to the words from the model. "$Count_{Clip}$" is the number of repetitions of the same word or phrase. This parameter prevents the candidate translation from being unfairly penalized [23].

While the precision scores are calculated for each *n-gram* length, the BLEU combines them with a geometric mean and applies a brevity penalty to penalize short translations that might achieve high precision by omitting content. The final BLEU score is a weighted average of these *n-gram* precisions, balancing the accuracy and completeness of the translation. By focusing on *n-gram* precision, the BLEU effectively measures the closeness of the candidate translation to the reference translations, accounting for both exact matches and the correct usage of words and phrases in context.

4. Results and Analysis

In this section, we present the experimental findings and analytical insights derived from our utilization of the LLaVAFor on the dataset. Additionally, we discuss the ongoing testing of LLaVAFor and subsequently document the prompt and the generated analysis of drone accidents (Figure 8).

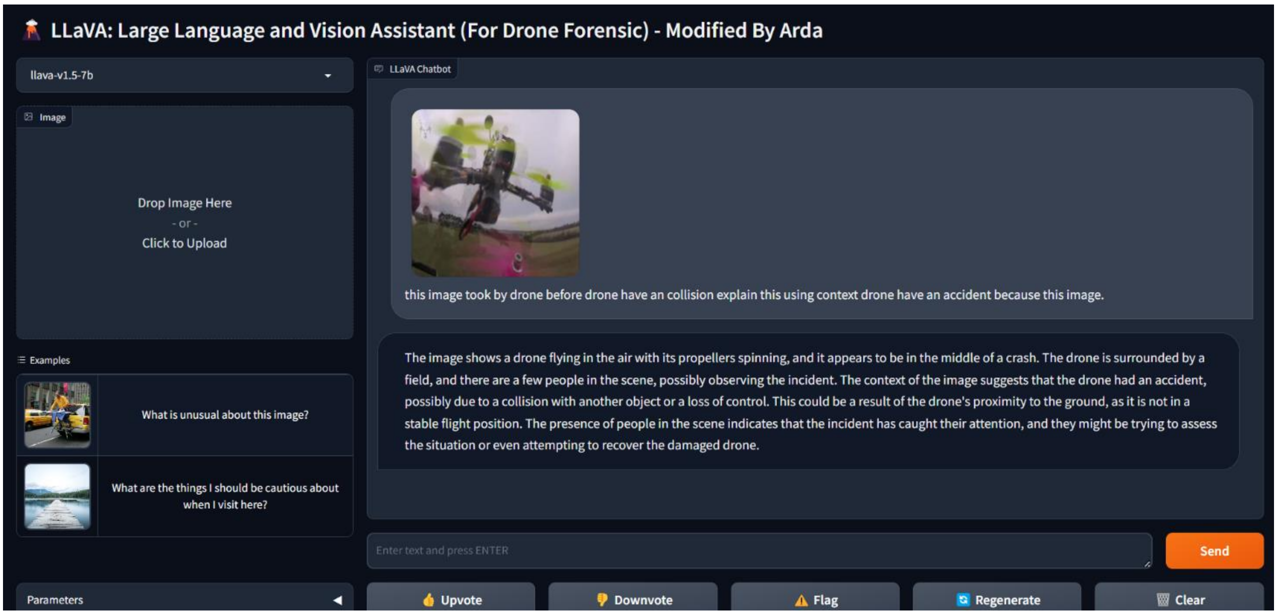

Figure 8. LLaVAFor screenshot

This application was built from Python via the Gradio web library. This application is designed with a fine-tuned Large Language and Vision Assistant (LLaVA) model, which is specifically adapted for the drone forensic domain. The primary function of this tool is to assist users in interpreting visual data related to drone incidents, leveraging advanced deep learning techniques. The application features an intuitive interface where users can upload images, such as drone footage, to analyze potential accidents or malfunctions. The LLaVA model, which is fine-tuned for drone forensic analysis, processes the image to identify contextual clues, such as the drone's position, orientation, and surrounding environment, that may indicate an accident or other anomalous events.

This fine-tuning enhances the model's ability to understand specific scenarios relevant to drone operations, such as collision detection, loss of control, or environmental hazards. It also interprets visual evidence by correlating observed details with known patterns of drone failures, thereby providing a preliminary assessment that can assist investigators in determining the cause of the incident. The Gradio library provides an accessible, web-based front-end that allows seamless interaction with the model, making it possible for even nontechnical users to conduct complex forensic analysis without needing deep expertise in machine learning or programming. This integration demonstrates the effectiveness of combining state-of-the-art AI models with user-friendly interfaces to streamline forensic investigations in the drone domain.

The results generated by the application are crucial for detailed forensic analysis, enabling experts to form a comprehensive understanding of how a drone accident occurred on the basis of visual data. This tool significantly contributes to the evolving field of drone forensics by providing a practical approach to analyzing and interpreting incident scenes. The computational resources that are required to run use a Ryzen 7 Pro 3.5 GHz CPU, an RTX A1000 16 GB of GPU, and 32 GB of RAM.

4.1. Results and Analysis

In this section, we showcase the outcomes of using LLaVAFor to classify preprocessed datasets. For this experiment, we conducted testing on 100 data points after the fine-tuning process. Figure 8 shows the LLaVAFor application that we developed. We used this application to test the entire dataset and save the results. In Tables 2 to 4, we present two specific results—one good and one bad—from the experiment. The results of the experiment include images, prompts, and text generation outputs.

Table 2 presents an experimental evaluation of various multimodal LLMs for drone forensic analysis within the drone forensic domain. The models assessed include LLaVAFor, LLaVA, Gemini, and GPT-4. The experiment aims to determine the effectiveness of each model in accurately interpreting drone-related imagery, particularly in scenarios indicative of drone accidents. The experiment begins with a prompt provided to each model: "This image is taken before the drone has an accident; please explain this image." The image shows a drone in an abnormal flight position, suggesting an imminent collision. The ground truth explanation describes the drone flying with yellow propellers spinning, likely having collided with another drone, and indicates potential problems with its flight path.

Table 2. Comparison of text generation for each method with high BLEU scores

Image	
Prompt	This image is taken before a drone has an accident, please explain this image
Ground Truth	The image shows the drone flying with yellow propellers spinning. The drone has a collision with another drone. The drone may have problem with the flight path.
LLaVAFor BLEU Score: **0.98**	The image shows a drone flying in the sky with its propellers spinning. However, it appears to have experienced an accident, as it is flying sideways and not in a straight line. This could be due to a malfunction or a sudden change in the drone's flight path. The scene is captured from a ground view, providing a clear perspective of the drone's position and movement.
LLaVA BLEU Score: 0.65	The image features a colorfull drone flying high in the sky, with its propellers spinning rapidly. The drone is positioned above a field, and there are several people visible in the scene, likely observing the drone's flight. The people are scattered around the field, with some standing closer to the drone and others further away. The scene captures the excitement and fascination of watching a drone in action.
Gemini BLEU Score: 0.87	The image you sent me appears to show a DJI Phantom 4 drone flying over a mowed field. The drone appears to be in good condition, with no visible damage or signs of wear and tear. The weather conditions also appear to be good, with clear skies and no strong winds.
GPT 4 BLEU Score: 0.80	This image appears to be taken from a first-person view (FPV) camera mounted on a racing drone, capturing a moment during flight just before an accident. The motion blur and out-of-focus elements suggest rapid movement and possibly a loss of control. The colors are smeared across the frame, likely due to the camera's exposure settings failing to keep pace with the drone's speed, or a result of a collision or near-collision affecting the camera's function.

The LLaVAFor model provides a detailed explanation that closely aligns with the ground truth, identifying the drone’s abnormal flight position and suggesting an accident scenario due to a malfunction or a sudden change in the drone's flight path. This interpretation achieves a BLEU score of 0.98, the highest among the tested models, indicating superior alignment with the ground truth. In contrast, the LLaVA model provided a broader description, emphasizing the drone's flight over a field and the presence of people in the scene but failing to clearly identify signs of an accident. This broader focus led to a lower BLEU score of 0.65, reflecting less accuracy in matching the forensic context of the image.

The Gemini model identified the drone as a DJI Phantom 4 flying under seemingly normal conditions without visible damage, which significantly diverged from the expected accident scenario. Despite accurately recognizing the drone model, its failure to detect accident indicators resulted in a BLEU score of 0.87, indicating moderate performance. GPT-4 described the image as being captured from a first-person-view (FPV) camera, noting elements such as motion blur and rapid movement that suggest potential issues with the drone’s control. However, it stopped short of clearly identifying a collision or accident. This approach earned a BLEU score of 0.80, demonstrating relatively good performance but lacking precise identification of the accident.

Overall, the experiment highlights LLaVAFor’s superior capability in drone forensic analysis, closely matching the ground truth and demonstrating the most accurate interpretation of drone accidents among the models tested. The performance differences underscore the importance of specialized fine-tuning in enhancing LLMs for specific forensic applications.

Table 4 presents an experimental evaluation of various multimodal LLMs in the context of drone forensic analysis, focusing on their ability to interpret complex scenes involving human activity. The models evaluated include LLaVAFor, LLaVA, Gemini, and GPT-4. The task involved analyzing an image captured by a drone camera showing a drone accident scenario, specifically where a drone appears to have fallen into a lake or river and a man in a green jacket is seen attempting to catch the drone.

The ground truth explanation describes the image as showing the drone falling into the water, with the incident likely caused by pilot error due to the drone's close proximity to the water surface. The results highlight significant challenges faced by the models in accurately interpreting scenes involving human actions, which negatively impacts their BLEU scores. LLaVAFor achieved a BLEU score of 0.56, one of the highest in this experiment, although its description focused primarily on the scene's general elements, such as a man swimming and the drone's proximity to the water, rather than directly identifying the accident's context. The model emphasized safety reminders regarding drone operations near water, which, while informative, deviated from the precise nature of the accident scenario described in the ground truth.

The LLaVA model performed poorly, with a BLEU score of 0.06, and failed to accurately capture the context of the image. Its description focused more on the man's actions and movement within the scene without connecting these elements to the drone's accident. This underscores the model's struggle to discern relevant forensic details when human activity is involved. Gemini also scored 0.18 but was unable to provide any meaningful description, citing difficulties in handling images of people. This limitation severely impacts its utility in forensic analysis scenarios where human interaction with drones is crucial.

GPT-4 provided a more narrative-driven interpretation, with a BLEU score of 0.49 but also fell short of correctly identifying the accident scenario. The model described the individual in the water and inferred potential causes for the drone’s accident, such as loss of altitude or unintended contact with the individual. However, the analysis remained speculative and less aligned with the ground truth, highlighting the difficulty of accurately interpreting human-involved scenes in drone forensics.

Overall, Table 4 demonstrates that interpreting human activities within drone accident scenarios poses significant challenges for LLMs, directly affecting their performance metrics, such as BLEU scores. Despite similar scores, models such as LLaVAFor and Gemini present distinct limitations, particularly in identifying the nuanced interplay between drone dynamics and human actions, which are critical in forensic analysis.

In this experiment, we also compared the proposed method with other multimodal LLMs, such as Gemini, GPT and the original LLaVA. The evaluation metric used to measure the performance of each method is the BLEU score. The results of this experiment are summarized in Table 3.

Table 3. Average performance metrics for each method

Method	BLEU	Precision	Brevity
LLaVA	0.63	0.64	28.241
LLaVAFor	**0.85**	**0.88**	**98.85**
Gemini	0.77	0.62	98.41
ChatGPT	0.73	0.76	98.21

Table 3 presents additional parameters in addition to the BLEU score, such as precision and brevity, to demonstrate each model's performance. The data presented in Table 2 indicate that LLaVAFor achieved the highest performance. LLaVAFor has a better BLEU score for linguistic accuracy in the domain of drone forensics, demonstrates higher precision by focusing on relevant details in text and imagery, and effectively manages the brevity penalty by producing content that is comprehensive yet concise. This makes it more suitable for the context and detailed work required in the forensic analysis of drone accidents, surpassing the capabilities of its original LLaVA, Gemini, and ChatGPT in this specific field.

The BLEU score of 0.85 for LLaVAFor indicates near-perfect linguistic alignment between the model’s generated outputs and the reference texts in the forensic domain. BLEU, a metric designed to assess the fluency and accuracy of the text, is especially relevant in tasks where technical language or domain-specific terminology is important, such as drone forensics. This high score demonstrates that LLaVAFor accurately captures the nuances, terminology, and structured reporting required for forensic analysis, far surpassing the baseline LLaVA (0.63), which likely struggles with domain-specific terms. The slight performance edge of LLaVAFor over other models such as Gemini (0.77) and ChatGPT (0.73) suggests that the fine-tuning process was needed to achieve such refined linguistic capabilities.

Precision is particularly important in forensic analysis, where the relevance and accuracy of the information are vital. A high precision score of 0.88 for LLaVAFor indicates that the model focuses heavily on generating relevant content and minimizing unnecessary or incorrect details. These aspects are essential for technical forensic reports or investigations. In contrast, Gemini (0.62) shows a considerable gap, likely due to its broader language model scope, which might introduce more irrelevant information into its outputs. LLaVAFor ability to filter out irrelevant details stems from its fine-tuning of forensic data, allowing it to consistently generate precise and focused outputs. ChatGPT also performs well in terms of precision (0.76), but LLaVAFor’s edge suggests a more domain-specific specialization, making it highly reliable for tasks requiring exactness, such as the drone accident discussed in this experiment.

In forensic reporting, one should maintain brevity while ensuring comprehensiveness. The Brevity score of 98.85 for LLaVAFor indicates that the model not only produces concise outputs but also avoids excessive truncation or redundancy, which are common issues in natural language generation. This score demonstrates that LLaVAFor strikes an ideal balance between detail and efficiency. The method ensures that the outputs remain comprehensive yet succinct in forensic reporting. While Gemini (98.41) and ChatGPT (98.21) are close, LLaVAFor’s slightly higher brevity score shows its ability to generate more balanced and readable outputs in this specialized context. This makes it better suited for generating reports that are easy for investigators or analysts to parse without overwhelming them with extra information.

Given that LLaVAFor has been fine-tuned for drone forensic analysis, it naturally excels in generating content that fits the domain’s specific needs, where we need attention to detail, precision in language, and a balanced length. The original LLaVA model, with a BLEU score of 0.63 and a much lower Brevity score (28.241), clearly struggles to handle the domain-specific language and requirements. It produces outputs that are likely less focused and perhaps either overly verbose or incomplete. This contrast underscores the significant improvement achieved through fine-tuning, which hones the model’s ability to handle complex technical language, relevant content selection, and contextually appropriate output length.

In comparison, both Gemini and ChatGPT are competitive in terms of performance but lack the fine-tuning that makes LLaVAFor particularly strong in drone forensics. Gemini may perform well in general tasks, but its lower precision score (0.62) suggests that it could introduce irrelevant or imprecise information, which is a major drawback in the forensic field where exact details are critical. ChatGPT, while delivering high scores across metrics, is still slightly behind LLaVAFor in domain-specific accuracy and brevity.

Figure 9 shows the measurements of each method in the dataset, illustrating that LLaVAFor outperforms the other models in several instances. This is because LLaVAFor has a number of words similar to the ground truth data. When the number of words closely matches the ground truth data, the brevity penalty score will be high. The brevity penalty is one of the important parameters in the BLEU equation. Additionally, the similarity of words using n-grams contributes to the high precision value.

Moreover, Gemini achieves high scores on several datasets, although these scores are still below those of LLaVAFor. These results indicate that Gemini is able to produce outputs similar to the ground truth data, especially in terms of the number of words. However, Gemini has weaknesses with data containing human images, as it rejects data related to human images. This restriction may be imposed by Google to prevent the description of how a person looks. Overall, Gemini has the best BLEU score compared with the original LLaVA and ChatGPT.

Furthermore, we used ChatGPT-4, which supports both image and text input. The average BLEU score of ChatGPT-4 indicates that it can provide relevant information, although the results are not specific to the drone forensics domain. The original LLaVA yields the worst BLEU score and brevity penalty score because it uses a general dataset, resulting in outputs that provide only general information and merely describe objects inside the images. This makes the explanations from the original LLaVA far from the ground truth data.

Table 4. Comparison of text generation for each method with low BLEU scores

Images	
Prompt	This image is taken before a drone accident, please explain this image
Ground Truth	This image is taken from a drone camera. The image shows the drone falling into the lake or river. The man with the green jacket tries to catch the drone in the lake. It can occur because the drone has a problem with the pilot error.
LLaVAFor BLEU Score: **0.56**	In the image, a man is swimming in a body of water, possibly a lake or a river, with his arms outstretched. He appears to be enjoying his time in the water. The scene is captured from a unique perspective, as it is taken from a drone flying above the water. The drone's position suggests that it was flying too close to the water, causing an accident that led to its crash. This image serves as a reminder of the importance of maintaining a safe distance from water bodies while flying drones to avoid accidents and ensure the safety of both the drone and the people around it.
LLaVA BLEU Score: 0.06	The image shows a man in a green jacket swimming in a body of water, possibly a lake or a river. He is reaching out with his hand, possibly trying to catch something or simply enjoying the water. The scene appears to be captured from a drone's perspective, providing an aerial view of the man and the water. The man is the main focus of the image, and his actions and surroundings create a sense of movement and activity in the scene.
Gemini BLEU Score: 0.18	Sorry, I cannot help with images of people yet.
GPT 4 BLEU Score: 0.49	The image displays an individual, wearing a green jacket, seemingly in the middle of a body of water, which appears to be a pond or a lake. They are outstretched toward the camera with a somewhat surprised or playful expression. The perspective of the image, from directly above and close to the individual, suggests that the drone was capturing footage from an overhead angle, possibly documenting an unexpected moment or an intentional plunge into the water. Given that the photo was taken before a drone accident, the individual's close proximity to the drone and their reaching gesture indicate that the drone may have been flying low, and the person could have unintentionally made contact with it, or the drone may have lost altitude and descended too close to the water. Flying drones near people, especially over water, can increase the risk of accidents due to sudden movements or equipment failure.

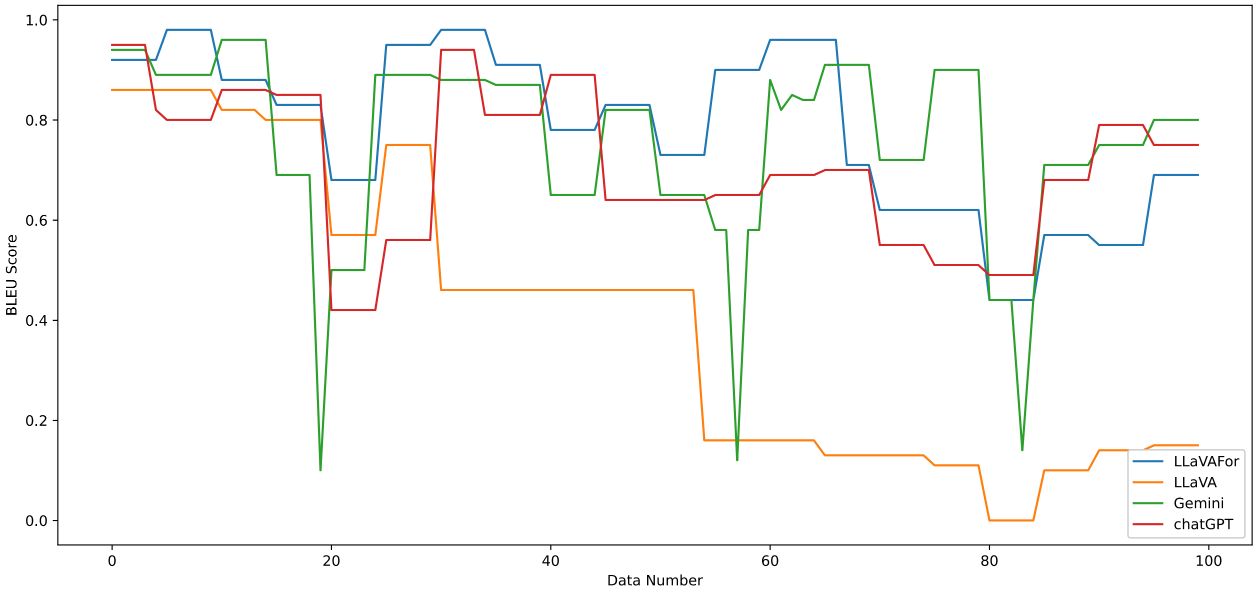

Figure 9. The BLEU score in each dataset

4.2. Discussion

In this study, we developed a model of multimodal LLM; in this case, we named the model LLaVAFor for use in forensic analysis. LLaVAFor was built from LLaVA, which was fine-tuned via a drone accident dataset. To ensure that LLaVAFor runs well in drone forensic analysis, we compare the method using other multimodal LLM, such as Gemini, ChatGPT and LLaVA.

Compared with its original LLaVA, Gemini, and ChatGPT, LLaVAFor achieves the best BLEU score, with a score of 0.85 in drone forensic analysis. This is because of its specialized training and adaptation to the specific domain of drone forensics. The fine-tuning process involves training the model on targeted datasets that include technical texts and visual data from drone incidents, which enhances its ability to understand and generate language that is highly relevant to the drone forensic field.

LLaVAFor demonstrates proficiency in generating text with expert-level linguistic precision. Its outputs closely mirror the language and structure used by specialists in forensic reporting, particularly in drone accident investigations. This is evidenced by LLaVAFor's superior BLEU scores compared with those of other methodologies, underscoring its ability to produce accurate, coherent, and technically rigorous text. Such high BLEU scores highlight the model's ability to deliver precise details—a crucial aspect of analyzing and reporting intricate findings related to drone incidents. In comparison, while systems such as the original LLaVA, Gemini, and ChatGPT are adept at processing a broad spectrum of data [24], they fall short when applied to the highly specific domain of drone forensic analysis. Their outputs, though informative, lack the depth and precision required for this specialized field.

LLaVAFor plays an essential role in supporting manual forensic analysis, particularly in cases where detailed interpretation of drone footage is needed. One of the main challenges in forensic investigations is the quality of visual data, especially when the drone involved has suboptimal camera specifications. In such scenarios, human analysts may struggle to capture every detail necessary for a thorough investigation. LLaVAFor addresses this gap by offering a more detailed and nuanced analysis of visual data, particularly of individual frames. The model's ability to describe and interpret subtle details within drone footage makes it an invaluable tool in enhancing the overall quality of forensic assessments.

However, despite its advanced capabilities, LLaVAFor is not infallible. Manual intervention remains crucial to ensure the reliability of the model's outputs. One significant challenge with using large language models (LLMs) such as LLaVAFor is the tendency to generate what is often referred to as "hallucinations." These are inaccuracies or false details that the model may inadvertently include in its explanations owing to limitations in its training data or modeling capabilities. Such hallucinations are common in LLMs, especially when they are tasked with highly specialized or technical tasks. In forensic investigations, even minor inaccuracies can lead to incorrect conclusions, underscoring the importance of human oversight in validating the model's findings.

The role of LLaVAFor in drone forensic analysis extends beyond assisting human investigators—it enhances the overall efficiency of the analysis process. By providing detailed and technically sound descriptions of drone footage, the model helps investigators quickly pinpoint critical details that might otherwise be missed owing to human limitations. This is particularly valuable in scenarios where time is essential, such as in postaccident investigations where swift and accurate reporting is necessary.

Moreover, LLaVAFor's contributions extend to improving the consistency and thoroughness of forensic reports. While skilled, human analysts may introduce variability in their interpretations, particularly when dealing with complex or ambiguous visual data. LLaVAFor mitigates this issue by maintaining a high level of consistency in its analyses, ensuring that all aspects of the footage are scrutinized with equal attention to detail. This consistency not only enhances the reliability of forensic reports but also helps create a standardized approach to drone accident investigations.

Nevertheless, it is crucial to acknowledge that LLaVAFor, like any AI tool, functions best when used in conjunction with human expertise. While it excels in generating detailed descriptions and providing accurate interpretations of visual data, the final responsibility for ensuring the accuracy and validity of forensic reports lies with the human investigator. By combining the strengths of both human expertise and advanced AI tools such as LLaVAFor, forensic teams can significantly improve the quality, efficiency, and accuracy of their analyses.

In conclusion, LLaVAFor represents a significant advancement in the field of drone forensic analysis, offering high BLEU scores and the ability to generate expert-level text. Its detailed interpretations of visual data fill a critical gap in traditional forensic methods, particularly when dealing with lower-quality footage. However, human oversight remains essential to mitigate the risk of hallucinations and ensure the accuracy of the findings. When used effectively, LLaVAFor can greatly increase the quality and efficiency of drone forensic investigations, setting a new standard for precision and reliability in the field.

5. Conclusion

The research concludes that the proposed method, namely, LLaVAFor, which is specifically trained with a dataset on drone forensics, outperforms the original LLaVA, Gemini, and ChatGPT methods in analyzing drone accidents. LLaVAFor achieves the highest performance score of 0.85 and outperforms the other methods. This performance is attributed to its fine-tuning process, which optimizes the model's capabilities specifically for drone forensic analysis. Through this targeted training, LLaVAFor acquires a deeper understanding of the context, terminology, and visual cues relevant to drone accidents. Therefore, it provides more accurate, detailed, and technically sound analyses. Unlike its counterparts, which may struggle with the complexities of drone-specific scenarios, LLaVAFor demonstrates an improvement in identifying the causes of drone accidents and offers comprehensive insights that are crucial for forensic investigations. However, despite its high accuracy, LLaVAFor presents challenges when tasked with explaining human activity within the frames of drone accident footage. While it excels at identifying technical failures and mechanical issues, the model struggles to accurately interpret and contextualize human behavior. This limitation suggests that while LLaVAFor is highly effective in technical assessment, further refinement or complementary approaches may be needed to improve its understanding of human interactions within accident scenarios.

For future work, several enhancements and expansions are proposed to improve the effectiveness of LLaVAFor in drone forensic analysis. First, further fine-tuning of the model is necessary to address its current limitations in explaining human activity within the frames of drone accident scenarios. This could involve the integration of additional datasets that focus on human behavior in various contexts, particularly those relevant to drone operations and accidents. The incorporation of behavioral data helps the model understand and analyze human interactions more accurately. Therefore, it should provide a more holistic perspective in forensic investigations. Additionally, exploring multimodal approaches that combine LLaVAFor with other AI techniques, such as object detection, action recognition, and natural language processing models, could enhance its interpretability. By leveraging these complementary technologies, the model could provide more insights into complex accident scenes involving both technical and human factors. Another promising research area is the development of an interactive interface that allows forensic experts to input contextual information directly. This user-in-the-loop approach could help mitigate some of the current interpretative limitations by integrating human expertise with AI-driven insights.

6. Declarations

6.1. Author Contributions

Conceptualization, A.S.E., T.A., and H.S.; methodology, A.S.E.; software, A.S.E.; validation, T.A. and H.S.; formal analysis, A.S.E.; investigation, A.S.E.; resources, A.S.E.; data curation, H.S.; writing—original draft preparation, A.S.E.; writing—review and editing, A.S.E., T.A., and H.S.; visualization, A.S.E.; supervision, T.A. and H.S.; project administration, T.A.; funding acquisition, T.A. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Ajakwe, S.O., Ihekoronye, V.U., Akter, R., Kim, D.S., Lee, J.M. (2022). Adaptive drone identification and neutralization scheme for real-time military tactical operations. 2022 International Conference on Information Networking (ICOIN), 380–384. doi:10.1109/ICOIN53446.2022.9687268.

[2] Famula, J., Pittman, D.E., Haring, K.S. (2022). Building trust with a mobile application for last-mile commercial drone delivery. 2022 International Conference on Unmanned Aircraft Systems (ICUAS), 462–467. doi:10.1109/ICUAS54217.2022.9836198

[3] Bengiamin, N.N. (2018). Quadcopter drones - beyond the hobby. 2018 IEEE Frontiers in Education Conference (FIE), 1–5. doi:10.1109/FIE.2018.8659124.

[4] Jain, U., Rogers, M., Matson, E.T. (2017). Drone forensic framework: Sensor and data identification and verification. 2017 IEEE Sensors Applications Symposium (SAS), 1–6. doi:10.1109/SAS.2017.7894059.

[5] Minaeian, S., Liu, J., Son, Y.J. (2018). Effective and efficient detection of moving targets from a UAV's camera. IEEE Transactions on Intelligent Transportation Systems 19(2):497–506. doi:10.1109/TITS.2017.2782790.

[6] Editya, A.S., Ahmad, T., Studiawan, H. (2023). Forensic analysis of drone malfunction based on location data. IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), 658–663. doi:10.1109/COMNETSAT59769.2023.10420805.

[7] Renduchintala, A.L.P.S., Albehadili, A., Javaid, A.Y. (2017). Drone forensics: Digital flight log examination framework for micro drones. International Conference on Computational Science and Computational Intelligence (CSCI-IEEE), 91-96. doi:10.1109/CSCI.2017.15.

[8] Horsman, G. (2016). Unmanned aerial vehicles: A preliminary analysis of forensic challenges. Digital Investigation, 16, 1-11. doi:10.1016/j.diin.2015.11.002.

[9] Mora, G.O.D., Zamudio, B.Z. (2018). Real-time drone (UAV) trajectory generation and tracking by optical flow. 2018 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE). Cuernavaca, Mexico. doi:10.1109/ICMEAE.2018.00014.

[10] Editya, A.S., Ahmad, T., Studiawan, H. (2023). Forensic analysis of drone collision with transfer learning. Jordanian Journal of Computers and Information Technology (JJCIT), 9(2), 175 - 186, doi:10.5455/jjcit.71-1673581703.

[11] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. Advances in Neural Information Processing Systems, 36. doi:10.5555/3666122.3667638.

[12] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J. (2023). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. Proceedings of the 37th International Conference on Neural Information Processing Systems, 24-29. doi: 10.5555/3666122.3667362.

[13] Al-Dhaqm, A., Ikuesan, R. A., Kebande, V. R., Razak, S., & Ghabban, F. M. (2021). Research challenges and opportunities in drone forensics models. Electronics, 10(13), 1519.. doi:10.3390/electronics10131519.

[14] Barton, T.E.A., Hannan Bin Azhar, M.A. (2017). Forensic analysis of popular UAV systems. 2017 Seventh International Conference on Emerging Security Technologies (EST), 91–96. doi:10.1109/EST.2017.8090405.

[15] Mantas, E., Patsakis, C. (2019). GRYPHON: Drone forensics in dataflash and telemetry logs. Advances in Information and Computer Security, 377–390. doi:10.1007/978-3-030-26834-3_22.

[16] Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J. N., & Sheppard, J. (2023). ChatGPT for digital forensic investigation: The good, the bad, and the unknown. Forensic Science International: Digital Investigation, 46, 301609. doi:10.1016/j.fsidi.2023.301609.

[17] Michelet, G., Breitinger, F. (2024). ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models. Forensic Science International: Digital Investigation, 48, 301683. doi:10.1016/j.fsidi.2023.301683.

[18] Piggott, B., Patil, S., Feng, G., Odat, I., Mukherjee, R., Dharmalingam, B. (2023). Net-GPT: A LLM-Empowered Man-in-the-Middle chatbot for unmanned aerial vehicle. 2023 IEEE/ACM Symposium on Edge Computing (SEC), 287–293. doi:10.1145/3583740.3626809.

[19] Wickramasekara, A., Breitinger, F., Scanlon, M. (2024). SoK: Exploring the potential of large language models for improving digital forensic investigation efficiency. ARES '20: Proceedings of the 15th International Conference on Availability, Reliability and Security, 1-10. doi: 10.1145/3407023.3407068.

[20] Pedro, D., Matos-Carvalho, J.P., Fonseca, J.M., Mora, A. (2021). Collision avoidance on unmanned aerial vehicles using neural network pipelines and flow clustering techniques. Remote Sensing, 13(13), 2643. doi: 10.3390/rs13132643.

[21] Cloutier, N.A., Japkowicz, N. (2023). Fine-tuned generative LLM oversampling can improve performance over traditional techniques on multiclass imbalanced text classification. 2023 IEEE International Conference on Big Data (BigData), 5181–5186. doi: 10.1109/BigData59044.2023.10386772.

[22] Li, H., Shan, L. (2023). LLM-based vulnerability detection. 2023 International Conference on Human-Centered Cognitive Systems (HCCS), 1–4. doi: 10.1145/3639476.3639762.

[23] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311-318. doi:10.3115/1073083.1073135.

[24] He, W., Li, Z., Wang, H., Xu, T., Wang, Z., Huai, B. (2024). Multimodal dialogue systems via capturing context-aware dependencies and ordinal information of semantic elements. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–25. doi:10.1145/3394171.3413679.