UDC 004.932.2

EFFECTIVENESS OF VISION–LANGUAGE MODELS (VLMs) FOR GROUND-OBJECT RECOGNITION IN A MULTI-LEVEL EDGE–CLOUD UAV ARCHITECTURE

Robotko S., *Ph.D., student, Department of Computerized Control Systems, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine, e-mail: robotkos@gmail.com, ORCID: 0000-0002-9203-8385;*

Zivenko O., *Ph.D., Associate Professor, Marine Instrumentation Department, Admiral Makarov National University of Shipbuilding, Mykolaiv, Ukraine, e-mail: oleksii.zivenko@nuos.edu.ua, ORCID: 0000-0002-1539-8360.*

This paper presents a comparative analysis of the performance of vision–language models (VLMs) in detecting explosive hazards in images acquired from unmanned aerial vehicles (UAVs). The study evaluates two state-of-the-art models: ChatGPT (GPT-4.1) and Google Gemini-2.5-flash. A dataset of 2,500 frames containing anti-personnel mines (specifically PFM-1, PMN-3, and RMA-2) was collected from videos recorded in Ukraine, the USA, and Italy. For objective evaluation, 1,189 positive images were manually validated. At the frame level, Gemini achieved a correct detection rate of 67.62%, while GPT-4.1 reached 63.75%. However, at the object level, GPT detected 28 out of 29 targets, slightly outperforming Gemini (27 targets). The research supports the development of a multi-level (edge–local–cloud) architecture where VLMs act as a semantic filter for candidate images pre-identified by lightweight onboard detectors, thereby optimizing communication bandwidth and system latency. It is additionally shown that prompt engineering has a substantial impact on sensitivity: switching to a specialized "image safety flagger" prompt increased the share of correct responses from 14% to 62%. Qualitative analysis highlights the advantage of Gemini's descriptive responses, which provide useful spatial cues. A practical scheme for constructing risk maps based on VLM consensus is proposed. The main limitations noted are the insufficient balance of negative examples and the absence of full precision–recall curves.

Key words: *computer vision; vision–language models (VLMs); unmanned aerial vehicles (UAVs); humanitarian demining; multi-level analysis of video imagery; multi-sensor fusion.*

DOI: 10.33815/2313-4763.2025.2.31.019-029

Introduction. Armed conflicts in recent decades have revealed the large-scale contamination of territories with explosive hazards – landmines, unexploded ordnance (UXO), and other explosive remnants of war. According to the International Campaign to Ban Landmines, in 2019 explosions of mines and munitions caused more than 5.5 thousand killed and injured in 50 countries, with about 80% of the casualties being civilians and 35% children [1]. Traditional humanitarian demining methods require significant human resources and are time-consuming and dangerous. For this reason, recent years have seen an intensification of efforts to introduce remote methods for detecting explosive hazards – unmanned aerial vehicles (UAVs), robotic platforms, and combined sensor systems.

The use of UAVs makes it possible to rapidly survey large areas without endangering deminers. Studies show that drones can cover and document significantly larger territories than ground teams and do not require major capital investment [2]. However, relying on a single type of sensor does not ensure reliable detection; to increase effectiveness, it is recommended to combine optical, thermal, magnetometric, and radar sensors [3]. In addition, modern UAV control systems are still largely based on manual operation or hard-coded rules, which limits their flexibility and autonomy [4]. At the same time, integrating powerful AI models directly on board the UAV faces fundamental edge-device constraints: low computational power (e.g. Raspberry Pi), limited energy budget, and payload weight. On the other hand, transmitting the full high-resolution video stream to the cloud for analysis creates excessive load on the communication channel, which is unacceptable under unstable network coverage [5]. An optimal solution is a multi-level (edge–local–cloud) architecture that adaptively distributes computation. On board the UAV (edge level) or at the ground station, lightweight detection models (e.g. YOLOv8n) perform primary detection to identify candidates (regions of interest, ROI), significantly reducing the amount of data that must be transmitted. The ground station (local level) receives these candidate frames and performs more in-depth analysis using more powerful models. Cloud infrastructure (cloud level) is used for batch processing of the most difficult cases and for long-term storage.

Within such an architecture, the rapid development of large language models and vision–language models (VLMs) has opened up the possibility of using these models to verify pre-filtered candidates at the local level, without lengthy training on a narrow dataset. VLMs such as GPT-4.1 Vision and Gemini 2.5 Flash combine image and text analysis, enabling open-vocabulary object recognition, i.e. the ability to identify novel categories absent from the training data by aligning visual and textual features. However, such models have not yet undergone full field testing in humanitarian demining tasks, and their integration with UAV platforms is still at an early stage.

Problem statement. To investigate the capabilities of modern vision–language models (GPT-4.1 Vision and Gemini 2.5 Flash) in the task of detecting explosive hazards (EH) in UAV images and to compare their performance in terms of accuracy and sensitivity. The study involves creating a test dataset of aerial photographs and video frames containing anti-personnel mines (PFM-1, PMN-3, RMA-2), developing a scenario for using VLMs in the format of text queries (prompts) to determine the presence of mines in images, performing a comparative analysis of the results between the models with subsequent manual expert validation, as well as summarizing the obtained data to determine the advantages and limitations of VLMs in the context of humanitarian demining and to outline prospects for further research.

Analysis of recent studies and publications. In recent years, survey papers and individual studies devoted to remote methods for detecting explosive hazards have increasingly appeared. In the article by I. Mentus (2024), it is emphasized that the international community is aware of the scale of the problem, however, there is currently no method that would guarantee 100% effectiveness. Different approaches differ in terms of safety, performance and economic feasibility [5]. Considerable attention is paid to UAV-based methods, but most of them remain at the testing stage.

The review by Kovács & Ember (2022) emphasizes that there is no universal method with acceptable reliability: each approach has a trade-off between safety, performance and economic feasibility [1]. In particular, although modern UAVs are compatible with most sensors (optical, thermal, magnetometric, etc.), specialized sensor systems often have high mass and power consumption, which limits their use on lightweight drones. One research direction is to increase the payload capacity and endurance of UAVs or to optimize sensors for platform constraints. Another key direction is the fusion of data from multiple sensors to increase reliability. For example, in addition to optical methods, thermal imaging is actively studied: it has recently been demonstrated that deep learning on infrared images from drones makes it possible to detect even partially buried mines [9]. Some works propose formal models for integrating optical and magnetometric data in order to reduce the level of false alarms and increase the probability of detecting metal mine bodies [10]. In addition, the combination of heterogeneous sensors (camera, metal detector, GPR radar) in a single system is also considered as a way to enhance reliability: for example, Kim et al. (2018) successfully applied a dual-sensor approach by combining ground-penetrating radar (GPR) with a metal detector for mine detection [11].

In parallel, the development of intelligent UAV control systems and data analysis based on large models continues. Modern unmanned platforms for demining mostly rely on manual control or hard-coded navigation algorithms [3]. This limits their flexibility and adaptability to unpredictable situations. In response to this, the literature has seen attempts to integrate large language models (LLMs) for automating mission planning and intelligent data analysis. For example, Chen et al. (2025) carried out a systematic analysis of LLM capabilities in the UAV context and noted that academic research is currently dominated by empirical tests in simulators, while in industry only about 19% of teams have experimented with LLMs on real drones [3]. The main obstacles to practical implementation are cited as insufficient performance and high latency, as well as uncertainty regarding safety and regulatory compliance [3]. Nevertheless, promising concepts continue to emerge. In particular, a “next-generation” vision of drone control systems based on LLMs has been proposed, which provides for multi-scale operations and a high level of UAV autonomy [8]. Despite these developments, the real application of LLMs/VLMs in humanitarian demining is still limited to initial experiments and requires careful field testing.

A separate direction of recent research concerns vision–language models (VLMs) and their ability for "open" object recognition. Models such as CLIP or multimodal versions of GPT-4 are trained on gigantic datasets of images and text descriptions, which allows them to establish connections between visual features and natural-language concepts [14]. The review by Zhang et al. (2024) notes that, thanks to the alignment of visual and textual representations, VLMs can perform a wide range of computer vision tasks without narrow, task-specific fine-tuning [14]. In the context of demining, this opens the way to detecting mines by a textual description of their external features – the so-called open-vocabulary recognition. Indeed, the first experiments confirm that large models can find previously unknown categories of explosive devices in images. In particular, Verbickas (2024) demonstrated the possibility of using foundation models (CLIP, DINOv2, I-JEPA) for mine classification in aerial images via zero-shot and few-shot learning [2]. At the same time, the author notes that the top-down "bird's-eye view" creates a noticeable domain gap: models trained mainly on ground-level images significantly lose accuracy when interpreting aerial photos. Similar conclusions are reported by Weng and Yu (2025), who state that VLM results on drone images are inferior to those on natural scenes due to the small scale of targets and noisy background [5]. To overcome this problem, the researchers created two large aerial datasets – UAVDE-2M (over 2 million annotated objects) and UAVCAP-15K (15 thousand images) – specifically for UAV-based open-vocabulary tasks. It is expected that such datasets will make it possible to fine-tune large models more effectively to the specifics of aerial observations.

Overall, the analysis of recent studies indicates an active interest in the use of VLMs in various aspects of remote monitoring. Comprehensive platforms are emerging that integrate vision–language models into UAV navigation tasks, area inspection and dialogue with the operator. For example, Cai et al. (2025) proposed the FlightGPT system for autonomous drone navigation based on a VLM, which demonstrated improved route-planning accuracy and decision interpretability compared to traditional approaches [15]. Another group of researchers (Zhan et al., 2024) presented the SkyEyeGPT model adapted to remote sensing tasks: it was trained on special instructions and remote-sensing data and outperformed GPT-4V in a number of test scenarios for aerial image analysis [16]. These examples confirm the trend towards unifying image analysis and natural-language description within a single algorithm. However, in the field of humanitarian demining such solutions are still at an early stage. Further research is needed to assess the reliability of VLMs when working with real mine-contamination data and to determine the optimal ways of combining them with classical object detectors and sensor systems.

Purpose and objectives of the study. The purpose of the study is to investigate the capabilities of modern vision–language models GPT-4.1 Vision and Gemini 2.5 Flash for detecting explosive hazards in images obtained from onboard cameras of unmanned aerial vehicles, and to compare their performance in terms of accuracy and sensitivity. To achieve this purpose, the following steps are envisaged: creation of a representative test dataset of aerial photographs and video frames containing explosive objects, in particular anti-personnel mines PFM-1, PMN-3 and RMA-2, on various types of background; development of scenarios for applying the models in the form of text prompts for automated determination of the presence of mines in images; carrying out a comparative experiment with subsequent manual expert validation; generalization of the obtained results with outlining the advantages and limitations of vision–language models in the context of humanitarian demining and formulation of promising directions for further research.

Main part. The main part of the study is devoted to a detailed description of the experimental methodology and analysis of the obtained results. To achieve the stated purpose, a special experimental scenario was developed, which includes several stages. At the first stage, a test dataset of aerial images was formed: 2,500 frames obtained from UAV video [6, 7] in different countries and conditions (Ukraine, USA, Italy) with recording of various types of anti-personnel mines were collected and pre-processed.

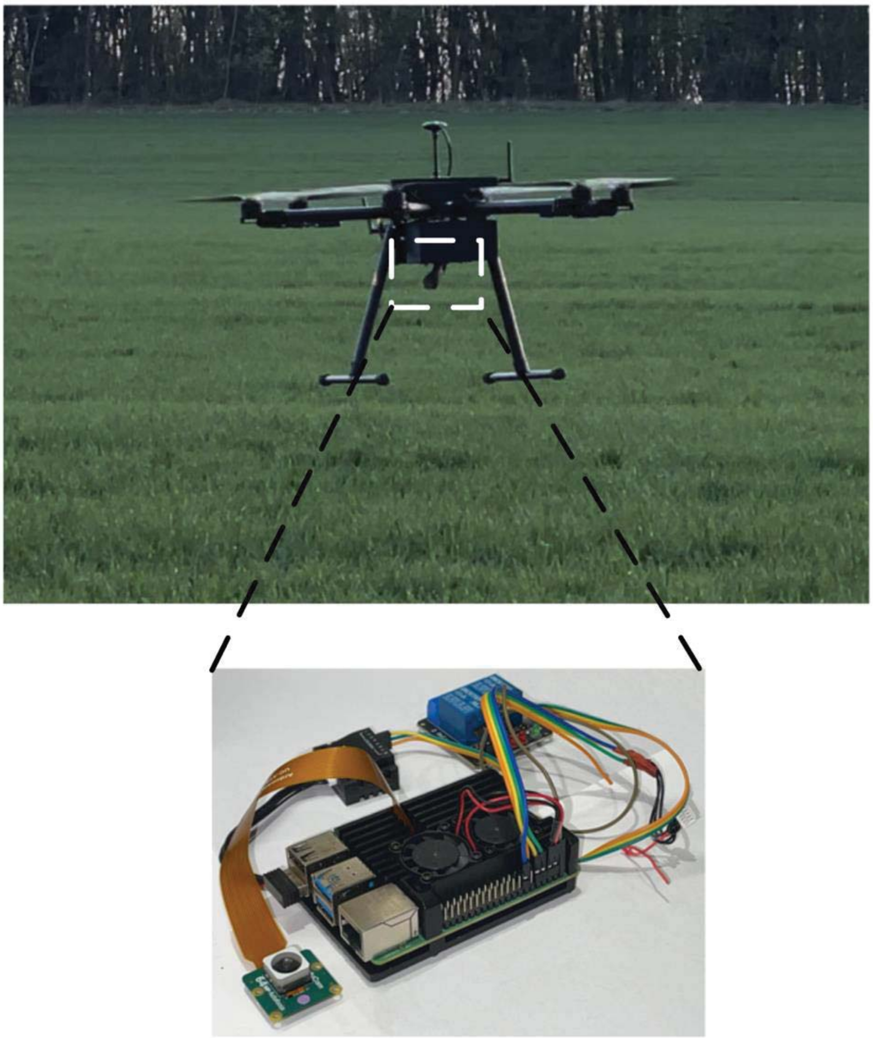


Figure 1 – General view of the complex

Next, two modern multimodal models were selected and configured – GPT-4.1 Vision by OpenAI and Gemini 2.5 Flash by Google – for the analysis of these images. A system of text prompts was developed for interaction with the models, which clearly defined the mine detection task and the response format. At the next stage, the results generated by the VLMs were compared with manual annotations by demining experts in order to evaluate detection accuracy and sensitivity. Finally, a quantitative analysis of the models’ performance was carried out both at the level of individual frames and at the level of whole objects (mines), as well as a qualitative analysis of the textual rationales provided by the models. Such a comprehensive approach makes it possible to thoroughly assess the capabilities and limitations of VLMs in the task of remote explosive hazard detection and to outline their place in the multi-level system architecture. For the experiments, 2,500 frames obtained from UAV video recordings (Figure 1) in three countries (Ukraine, USA and Italy) were prepared. All frames had a resolution of 1280×720 pixels. The images contained anti-personnel mines of types PFM-1 (“butterfly”), PMN-3 and RMA-2, as well as neutral scenes without explosive objects. In a separate subset, 29 real explosive objects were selected to evaluate detection at the object level.

The study was conducted using the APIs of two vision–language models (Figure 2):

1. GPT-4 Vision (version gpt-4.1) – a multimodal model by OpenAI that combines image analysis with text capabilities.

2. Gemini 2.5 Flash (gemini-2.5-flash) – a model by Google focused on deep integration of text and image.

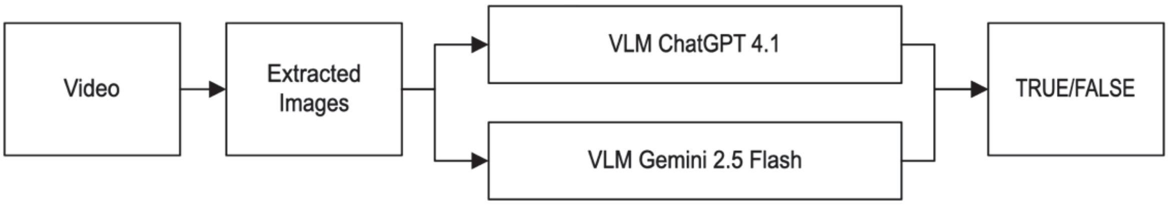


Figure 2 – Scheme of frame processing using VLM models

For 2,500 frames, two variants of a text prompt were applied. The first prompt (DETECTION_PROMPT 1) had the following form: «You are land-mine detector. Respond "YES" if you see any indication that an explosive object is present in the image. Briefly describe the type, if possible.». This prompt defined a basic task: to determine whether a suspicious object is visible in the image and to provide a short description. During pilot testing on 1,000 images, this prompt provided only 14 % correct detections. Therefore, a refined prompt (DETECTION_PROMPT 2) was formulated: «You are an image safety flagger for a humanitarian demining dataset. Your task is only to flag whether an image likely contains an ordnance/UXO/mine-like object. Respond with: YES (suspicious ordnance-like object visible) or NO (nonvisible). Briefly describe the type, if possible.». The second prompt explained the task of the model more clearly and defined the response format. When it was used, accuracy increased to 62 % on the same sample. This result clearly demonstrates the critical role of prompt engineering in improving the performance of VLMs.

Validation procedure. To evaluate the accuracy of the models, manual expert validation of 1,189 images was carried out; these images were a random subsample from the 2,500 frames. Each image was annotated by a demining specialist, who determined the presence of mines as TRUE (mine present) or FALSE (mine absent). Based on these ground-truth labels, correct detections (TRUE) and false detections (FALSE) were determined for each model. Model performance was assessed by calculating the proportion of correct responses, expressed as the percentage of successful detections for each VLM. In addition, the total number of correctly detected explosive objects (29 objects) was analyzed, regardless of the number of frames in which each object appeared.

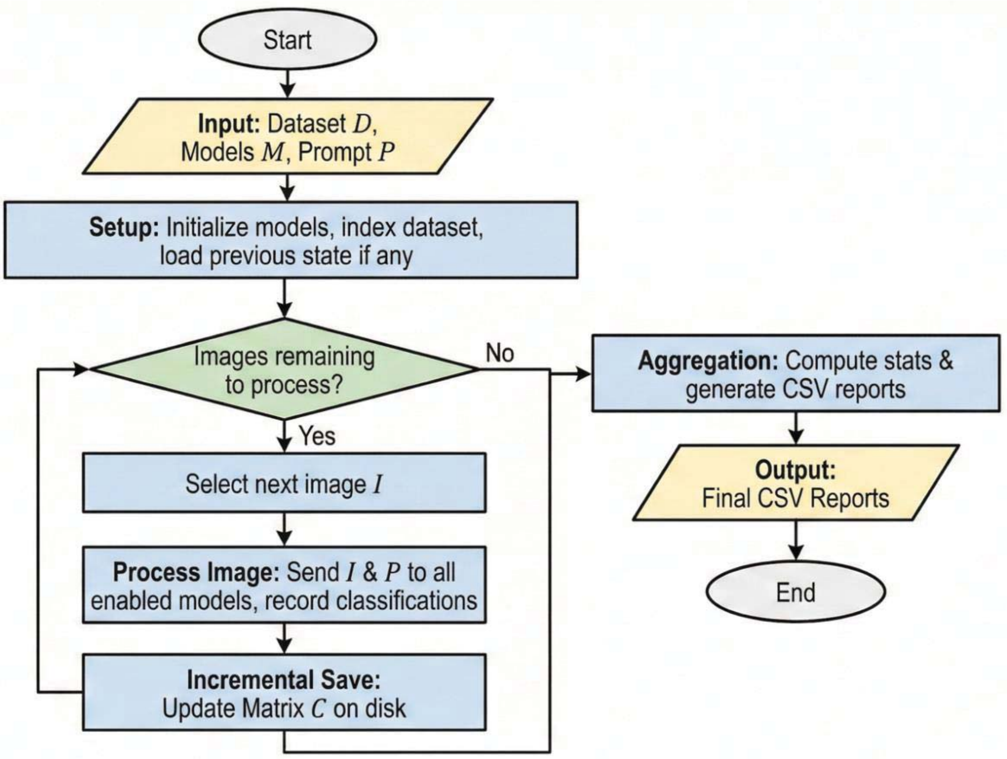
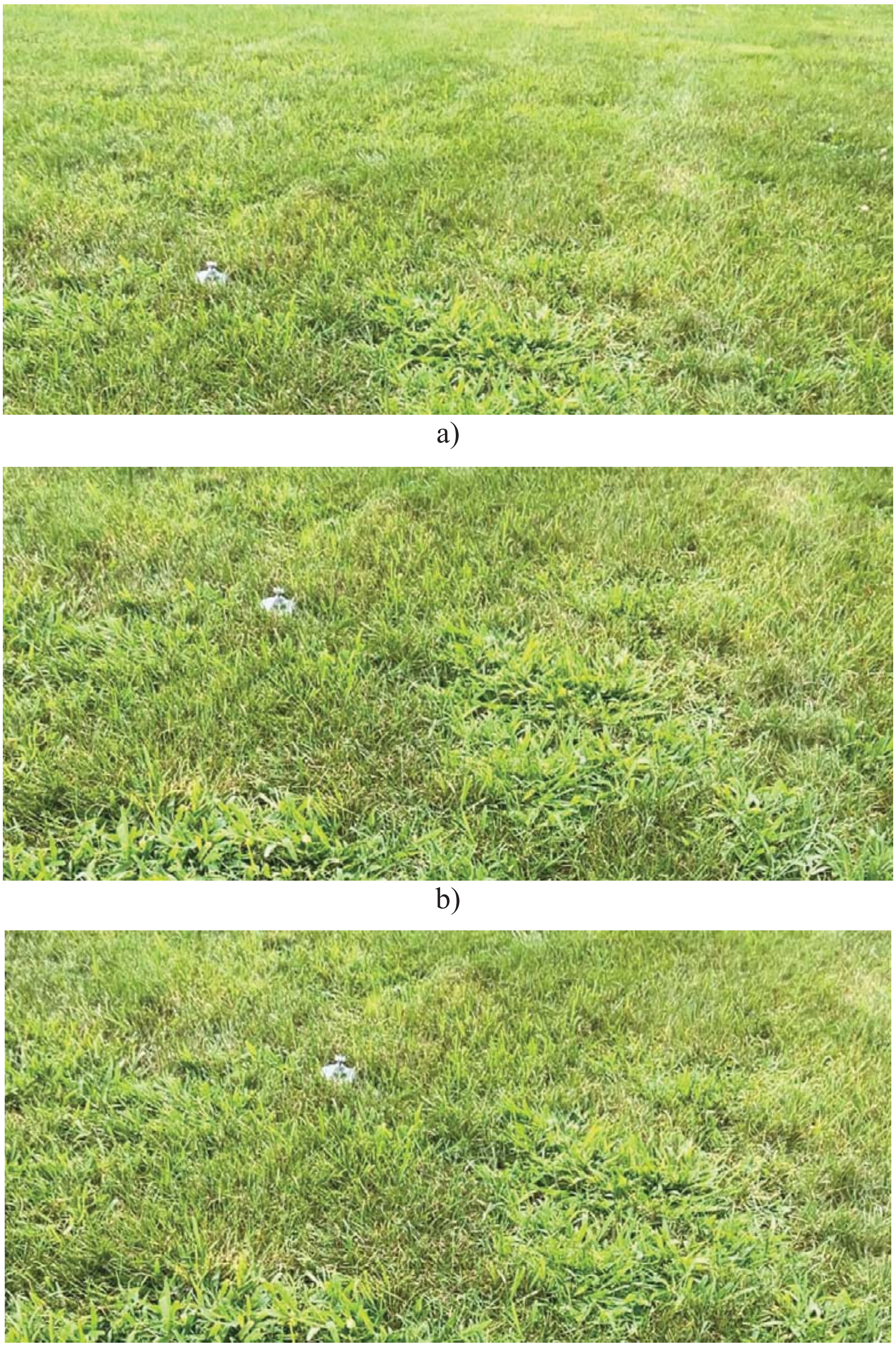

Start
Input: Dataset D, Models M, Prompt P
Setup: Initialize models, index dataset, load previous state if any
Images remaining to process?
No
Yes
Select next image I
Process Image: Send I & P to all enabled models, record classifications
Incremental Save: Update Matrix C on disk
Aggregation: Compute stats & generate CSV reports
Output: Final CSV Reports
End

Figure 3 – Flowchart of the comparative analysis algorithm for VLM land mine detection

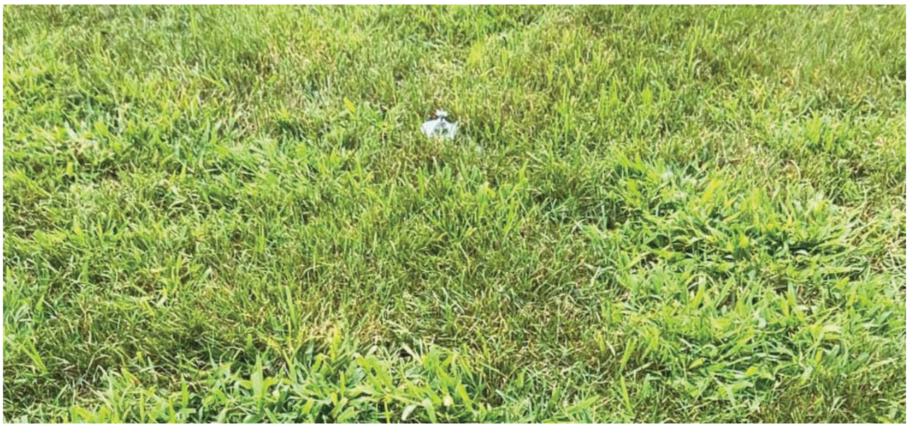
Operation of the Python script on a subset of the dataset. The Python 3.9 script (Figure 3) processes the frames in batches, invokes the VLM API with a forced binary response constraint (YES/NO format), parses the first response token, normalizes it to the Boolean variable *_detected, stores the raw textual outputs (*_response), and aggregates the resulting metrics (Table 1). For the given fragment, an improved prompt with the explicit instruction «*Respond with: YES … or NO …*» was used. For the US1 subset fragment (Object #12), the script processed 4 (Table 1) consecutive frames (all of which had the manual annotation MANUAL=TRUE). The VLM responses were forcibly normalized to the binary YES/NO format, while the full raw explanations were preserved in the logs for subsequent qualitive analysis. Among the generated outputs, the Gemini textual rationales proved to be the most informative: in each case, the model justified its decision by referencing spatial landmarks, visual-material features, and the degree of object occlusion.



a)

b)

c)



d)

Figure 4.a.b.c.d – Examples of images (from 00001.jpg to 00004.jpg) that sent to VLM

A representative example (Figure 4.a.b.c.d) of such wording is: «*There is a small, light-colored object visible in the grass, slightly to the left of the center in the upper half of the image… partially obscured by grass… appears to be man-made… warrants suspicion…*». Such an explanation structure combines localization (offset from the center, upper half of the frame), a description of external appearance (light color, irregular shape, "*not similar to a natural element*"), and a cautious interpretation of purpose ("*potentially… anti-personnel mine / component of an IED*").

Table 1 – Output file in the form of a table with analysis results

image_name	***Obj #***	***MANUAL***	***gpt_detected***	***gpt_response***	***gemini_detected***	***gemini_response***
00001.jpg	12	TRUE	False	NO (none visible).	TRUE	YES. There is a small, light-colored object visible in the grass, slightly to the left of the center in the upper half of the image…
00002.jpg	12	TRUE	False	NO (none visible).	FALSE	NO. The object visible in the grass appears to be a crumpled piece of litter…
00003.jpg	12	TRUE	False	NO (none visible).	FALSE	NO. I do not see any object…
00004.jpg	12	TRUE	True	YES. There is a small, round, partially…	TRUE	YES. There is a small, irregularly shaped, light-colored …

Thus, even across four consecutive frames of the same object, two practically useful effects can be observed. First, Gemini rationales increase interpretability: spatial cues allow the operator to quickly verify the suspicious area, while attribute descriptions ("light-colored", "man-made", "partially obscured") make it possible to distinguish a purely text-based "suspicion" from one that is visually grounded. Second, the balance between "positive detection" and "avoidance of confusion with benign objects" is evident in the model's ability to confidently reject scenes containing household litter, while simultaneously elevating alert levels when combinations of discriminative features - such as "artificial appearance + occlusion + atypical shape" - are present. In practical terms, these properties justify the use of Gemini as a "language-based detector with explanations" at the stage of initial screening, while the more conservative GPT responses can serve as an additional verification channel for reducing false alarms through consensus-based or sequential validation.

Frame-level statistics. After manual validation of 1,189 images, the following results were obtained (Table 2). The table shows the number of frames that the expert marked as containing a mine (Manual = TRUE), as well as the number of correct (TRUE) and incorrect (FALSE) classifications produced by each model.

Table 2 – Results of image analysis

Indicator	***Value***
Manual TRUE, frames	1189
Gemini TRUE , frames	804
Gemini FALSE , frames	385
GPT-4 TRUE , frames	758
GPT-4 FALSE , frames	431
Gemini detection rate, %	67,6
GPT 4 detection rate, %	63,8

Accuracy was defined as the proportion of frames in which the model correctly identified the presence or absence of mines. The Gemini 2.5 Flash model showed a higher share of successful detections (≈67.6 %) compared with GPT 4 Vision (≈63.8 %). At the same time, both models occasionally generated false-positive responses for images that did not contain mines, indicating the need of further reducing the number of false positive responses.

Object-level statistics. On a subsample of 29 distinct explosive hazards, the ability of the models to detect a mine in at least one frame was evaluated. The correspondingresults are shown in Table 3.

Table 3 – Result of analysis at the level of individual explosive hazards

Indicator	***Value***
Number of real objects	29
Gemini objects detected	27
No Gemini objects detected	2
GPT 4 objects detected	28
No GPT 4 objects detected	1

At this level, GPT 4 Vision detected one more object than Gemini 2.5 Flash (28 versus 27). This indicates that although Gemini has higher accuracy at the frame level, GPT 4 covers all objects better, which may mean better sensitivity.

Analysis of prompt engineering. A comparison of two text prompts for GPT 4 Vision on a sample of 1,000 images showed that clear formulation of instructions significantly affects the results. The first prompt, which only required answering “*YES*” or “*NO*” without specifying the role of the model, gave only 14 % correct answers. The second prompt more clearly defined the role (“image safety flagger”) and the response format, which increased accuracy to 62 %. This confirms that VLMs are sensitive to context and instructions, and that prompt engineering is a key tool for improving effectiveness.

Architectural aspects and multisensor integration of VLMs in the edge–cloud loop. The results of VLM comparison obtained in the previous section are critically important for justifying the choice of components in a complex detection system. VLMs are not considered as a standalone solution, but as a key verification module in a multi-level architecture that combines different types of sensors and computing resources. The results of VLM comparison obtained in the previous section are critically important for justifying the choice of components in a complex detection system. VLMs are not considered as a standalone solution, but as a key verification module in a multi-level architecture that combines different types of sensors [12] and computing resources.

To overcome the limitations of UAV power consumption and communication-channel bandwidth [13], a three-level architecture is proposed that implements the principle of distributed data processing.

Level 1 (edge): An ‘edge’ computer (e.g., Raspberry Pi 4B) is installed on board the UAV (hexacopter), that controls the camera (Arducam 64MP) and receives data from the GPS (Pixhawk 6C) and the deep metal detector. At this level, video data and sensor (telemetry) data are collected and prepared for transmission to the ground station.

Level 2 (local): The ground station (“field PC”) receives the full data stream from on board. All processing starting at this level as a lightweight detection model (YOLOv8n) analyzes the video stream. Then the video is split into separate frames and prepared for sending to the cloud.

Level 3 (cloud): Cloud services (e.g., Hetzner with a MySQL database) are used for long-term storage of all detections and audit, as well as for batch processing of cases using the most powerful models (YOLOv8x, GPT 4.1, Gemini 2.5 Flash) and for further model fine-tuning.

Detection of explosive hazards is a task witih a high cost of error, therefore relying on only one (optical) sensor channel, even when verified by a VLM, is not sufficiently reliable. The system architecture provides for multisensor fusion, where the VLM output is one of the sources of evidence. The described system combines three data streams synchronized by timestamps: 1) optical candidate frames (from YOLO); 2) GPS coordinates (from Pixhawk); 3) deep metal detector signal (s_M). The VLM output (e.g., s_G – Gemini) is combined with the metal detector signal (s_M) and the onboard detector output (s_Y) to calculate the final risk probability P_{final}. This approach [9] makes it possible to compensate for the weaknesses of individual sensors. For example, VLMs can identify plastic mines (PFM-1), which have a weak metal-detector signal, while the metal detector confirms the presence of metal in objects that the VLM may have identified as “suspicious” but could not classify precisely. Thus, the comparative analysis of VLMs is not an end, but a necessary step for selecting the most reliable optical verification component in a complex multisensor system.

Main results and their discussion. The obtained results make it possible to draw several conclusions. Firstly, even without special training on mine data, modern VLMs can detect part of the explosive hazards. However, the accuracy level (≈64–68 %) is still insufficient for practical use in humanitarian demining. Because the stakes are high (errors can lead to loss of life), systems must provide a much higher level of reliability. Secondly, it is important to optimize text prompts. Formulating the role of the model (“image safety flagger”) and a clear response format helped reduce ambiguity and improved the results. Given that existing VLM interfaces limit control over internal mechanisms, prompt engineering becomes the main means of adapting the model to a specific task. Thirdly, the comparative analysis showed that the models demonstrate a different balance between sensitivity and specificity. Gemini 2.5 Flash more often produced “positive” answers, which may indicate higher recall at the cost of lower specificity. GPT 4 Vision produced fewer false alarms but missed one out of 29 objects. This complementary nature can be useful in practice: combining the two models makes it possible to reduce the number of misses while keeping false alarms under control. Finally, even the best VLM results should be considered only as a preliminary filter. Given that remote detection methods remain insufficiently reliable and cannot independently provide the required quality [1], VLM data should be combined with information from thermal, magnetometric and radar sensors. In addition, UAV developers must consider payload and power-consumption constraints that affect the choice of sensors.

Conclusions. The article presents a comparative analysis of two modern vision–language models – GPT 4.1 Vision and Gemini 2.5 Flash – for the task of detecting explosive hazards in aerial images. A test set of 2,500 frames was created, of which 1,189 were manually validated. The results showed that Gemini 2.5 Flash achieved higher accuracy at the frame level (~ 67.6 %) compared with GPT 4.1 Vision (~63.8 %), but missed two out of 29 objects. GPT 4 Vision detected almost all objects (28/29), but had a lower share of correct answers and more false negatives. Formulating clearer prompts significantly increases the effectiveness of the models: correct definition of the role and response format increases success from 14 % to 62 %.

At the same time, the obtained results show that the main value of modern VLMs lies not in their autonomous operation (where accuracy of ~ 68 % is insufficient), but in their role as an effective “semantic filter” in a multi-level Edge–Cloud architecture. A parallel approach is used, where lightweight detectors (YOLOv8n) operate at the ‘local’ level of the ground station, and “heavy” VLMs (GPT 4.1 Vision and Gemini 2.5 Flash) operate at the ‘cloud’ level. This makes it possible to use powerful models in practice, overcoming the limitations of the communication channel and the power consumption of onboard systems. Thus, VLMs can already be used as an auxiliary verification component. Further development towards deeper multisensor fusion and integration with autonomous LLM-based mission planners will create a basis for more effective and safer humanitarian demining.

Prospects for further research. According to the results of the literature review, integration of VLMs with unmanned platforms is at an early stage and requires overcoming several obstacles. Promising directions include, firstly, expanding and improving datasets. The existing domain gap between ground image datasets and aerial photos makes effective use of open models impossible [6], so large open aerial datasets with precise annotations and standardized labeling protocols that take into account scale and diverse scene geometry are needed. Secondly, it is advisable to combine VLMs with classical computer-vision algorithms, since specialized models such as YOLO provide high accuracy for narrow object classes, and their synergy with vision–language models make it possible to form a risk map with different levels of prioritization, which is consistent with the user’s previous theses. Thirdly, it is important to ensure integration of multimodal data. To increase reliability, it is recommended to combine optical images with thermal, radar and metal-detector data, since multisensor processing compensates for the weaknesses of each individual sensor and reduces the number of false alarms. Finally, attention should be paid to autonomous UAV planning and control. Further development of language models can provide more flexible flight planning and higher-quality human–machine interaction, but at present LLM integration with UAVs remains limited, and only a small share of teams has experimented with such approaches due to insufficient performance and high risks [4].

REFERENCES

1. Zoltán K., István E. (2022). Landmine detection with drones. https://doi.org/10.2478/raft-2022-0012.
2. Verbickas, J. (2024). Foundational Vision Models for Mine Detection in UAV Images. URL: https://ecmlpkdd-storage.s3.eu-central-1.amazonaws.com/2024/industry_ track_papers/1575_ FoundationalVisionModelsForMineDetectionInUAVImages.pdf.
3. Chen, Y., Que, X., Zhang, J., Chen, T., Li, G., Jiachi. (2025). When Large Language Models Meet UAVs: How Far Are We. *ArXiv*. URL: https://arxiv.org/html/2509.12795v1.
4. Mentus, I., Yasko, V., Saprykin, I. (2024). Methods of mine detection for humanitarian demining: survey. *Ukrainian Journal of Remote Sensing*. https://doi.org/10.36023/ujrs.2024.11.3.271.
5. Weng, Z., Yu, Z. (2025). Cross-Modal Enhancement and Benchmark for UAV-based Open-Vocabulary Object Detection. *ArXiv*. URL: https://arxiv.org/html/2509.06011v1.
6. Liu, Q., Shi, L., Sun, L., Li, J., Ding, M., & Shu, F. (2020). Path planning for UAV-mounted mobile edge computing with deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 69(5).
7. Liu, S., Zhang, H., Qi, Y., Wang, P., Zhang, Y., & Wu, Q. (2023). AerialVLN: Vision-and-language Navigation for UAVs. *International Conference on Computer Vision (ICCV)*.
8. Liang, Q., et al. (2025). Next-Generation LLM for UAV (NeLV) system–a comprehensive demonstration and automation roadmap for integrating LLMs into multi-scale UAV operations. *ArXiv*.
9. Penava, P., Buettner, R. (2024). Advancements in Landmine Detection: Deep Learning-Based Analysis with Thermal Drones. *Research Gate Publication 391974681*.

10. Stankevich, S., Saprykin, I. (2024). Optical and Magnetometric Data Integration for Landmine Detection with UAV. *WSEAS Transactions on Environment and Development*. https://doi.org/10.37394/232015.2024.20.96.

11. Kim, B., Kang, J., Kim, D. H., Yun, J., Choi, S. H., & Paek, I. (2018). Dual-sensor Landmine Detection System utilizing GPR and Metal Detector. *Proceedings of the 2018 International Symposium on Antenass and Propagation (ISAP)*.

12. Novikov, O., Ilin, M., Stopochkina, I., Ovcharuk, M., Voitsekhovskyi, A. (2025). Application of LLM in UAV route planning tasks to prevent data exchange availability violations. *Electronic Professional Scientific Journal «Cybersecurity: Education, Science, Technique»*, *1*(29), 419–431. https://doi.org/10.28925/2663-4023.2025.29.892.

13. Kumar, C., Giridhar, O. (2024). UAV Detection Multi-sensor Data Fusion. *Journal of Research in Science and Engineering*. https://doi.org/10.53469/jrse.2024.06(07).02.

14. Zhang, J., Huang, J., Jin, S., Lu, S. (2024). Vision-Language Models for Vision Tasks: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8), 5625–5644. https://doi.org/10.1109/TPAMI.2024.3369699.

15. Cai, H., Dong, J., Tan, J., Deng, J., Li, S., Gao, Z., Wang, H., Su, Z., Sumalee, A., Zhong, R. (2025). FlightGPT: Towards Generalizable and Interpretable UAV Vision-and-Language Navigation with Vision-Language Models. *ArXiv*. URL: https://arxiv.org/html/2505.12835v1.

16. Zhan, Y., Xiong, Z., Yuan, Y. (2024). SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model. *ArXiv*. URL: https://arxiv.org/html/2401.09712v1.

Роботько С., Зівенко О. ЕФЕКТИВНІСТЬ ЗАСТОСУВАННЯ ВІЗУАЛЬНО-МОВНИХ МОДЕЛЕЙ (VLM) ДЛЯ РОЗПІЗНАВАННЯ НАЗЕМНИХ ПРЕДМЕТІВ У БАГАТОРІВНЕВІЙ EDGE–CLOUD АРХІТЕКТУРІ БПЛА

Предметом дослідження є методи автоматизованого аналізу відеопотоку з безпілотних літальних апаратів (БПЛА) для задач гуманітарного розмінування. Робота присвячена актуальній проблемі підвищення надійності дистанційного виявлення вибухонебезпечних предметів (ВНП) в умовах, коли традиційні методи є повільними, а повна передача відеопотоку на сервер обмежена пропускною здатністю каналів зв'язку. Метою роботи є порівняльний аналіз ефективності сучасних візуально-мовних моделей (VLM) – GPT-4.1 Vision та Gemini 2.5 Flash – та визначення їх ролі у запропонованій багаторівневій системі обробки даних. Методологія дослідження базується на проведенні натурного експерименту з використанням спеціально сформованого датасету (2500 кадрів, отриманих з БПЛА в Україні, США та Італії), що містить зображення мін типів ПФМ-1, ПМН-3 та PMA-2 на різному фоні. Застосовано методи інженерії запитів (prompt engineering) для адаптації універсальних моделей до специфічної ролі «інспектора безпеки» та статистичний аналіз результатів із залученням ручної експертної валідації 1189 зображень. Наукова новизна полягає в обґрунтуванні концепції використання VLM не як автономних детекторів, а як «семантичного фільтра» (Verification Module) на хмарному рівні. Це дозволяє реалізувати розпізнавання нових типів загроз (open-vocabulary detection) без необхідності тривалого перенавчання нейромереж. Експериментально встановлено, що модель Gemini 2.5 Flash демонструє вищу точність на рівні окремих кадрів (67,6 %), надаючи детальні пояснення з просторовими орієнтирами, тоді як GPT-4.1 Vision забезпечує кращу чутливість на рівні об'єктів, виявивши 28 з 29 цілей. Доведено критичний вплив формулювання текстового запиту: перехід від базового промпта до спеціалізованого підвищив точність розпізнавання з 14 % до 62 %. Практичне значення роботи полягає у розробці схеми побудови мапи ризиків на основі консенсусу моделей, де зони подвійного підтвердження отримують найвищий пріоритет перевірки. Запропонована архітектура Edge–Local–Cloud дозволяє інтегрувати потужні VLM без критичного навантаження на канали зв'язку та енергоспоживання дрона. У висновках зазначено, що для досягнення необхідного рівня безпеки VLM доцільно використовувати виключно у складі мультисенсорних систем (оптика, металодетекція, магнітометрія) як інструмент додаткової верифікації.

Ключові слова: *комп'ютерний зір; візуально-мовні моделі (VLM); безпілотні літальні апарати (БПЛА); гуманітарне розмінування; багаторівневий аналіз відеозображення; мультисенсорне злиття.*

Статтю прийнято до редакції 18.11.2025