Advancing Early Wildfire Detection: Integration of Vision Language Models with UAV Remote Sensing for Enhanced Situational Awareness

Leon Seidel * , Simon Gehringer , Tobias Raczok , Sven-Nicolas Ivens , Bernd Eckardt , Martin Maerz

Posted Date: 19 March 2025

doi: 10.20944/preprints202503.1391.v1

Keywords: Wildfire detection; Vision Language Models; UAV remote sensing; Early fire detection; Situational awareness; Forest fire data; Edge AI

Article

Advancing Early Wildfire Detection: Integration of Vision Language Models with UAV Remote Sensing for Enhanced Situational Awareness

Leon Seidel [1,*], Simon Gehringer [1], Tobias Raczok [2], Sven-Nicolas Ivens [2], Bernd Eckardt [1] and Martin Maerz [3]

[1] Fraunhofer Institute for Integrated Systems and Device Technology (IISB), Schottkystraße 10, 91058 Erlangen, Germany
[2] Fraunhofer Institute for Integrated Circuits (IIS), Am Wolfsmantel 33, 91058 Erlangen, Germany
[3] Institute of Power Electronics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Fürther Straße 250, 90429 Nuremberg, Germany
* Correspondence: leon.seidel@fau.de

Abstract: Early wildfire detection is critical for effective suppression efforts, necessitating rapid alerts and precise localization. While computer vision techniques offer reliable fire detection, they often lack contextual understanding. This paper addresses this limitation by utilising Vision Language Models (VLMs) to generate structured scene descriptions from Unmanned Aerial Vehicle (UAV) imagery. UAV-based remote sensing provides diverse perspectives of potential wildfires, and state-of-the-art VLMs enable rapid and detailed scene characterization. We evaluate both cloud-based (OpenAI, Google DeepMind) and open-weight, locally deployed VLMs on a novel evaluation dataset specifically curated for forest fire scene understanding. Our results demonstrate that relatively compact, fine-tuned VLMs can provide rich contextual information, including forest type, fire state, and fire type. Specifically, our best-performing model, ForestFireVLM-7B (fine-tuned from Qwen2-5-VL-7B), achieves a 76.6% average accuracy across all categories, surpassing the strongest closed-weight baseline (Gemini 2.0 Pro at 65.5%). Furthermore, zero-shot evaluation on the publicly available FIgLib dataset demonstrates state-of-the-art smoke detection accuracy using VLMs. Our findings highlight the potential of fine-tuned, open-weight VLMs for enhanced wildfire situational awareness via detailed scene interpretation.

Keywords: Wildfire detection; Vision Language Models; UAV remote sensing; Early fire detection; Situational awareness; Forest fire data; Edge AI

1. Introduction

Wildfires are burning large areas yearly and are responsible for substantial life losses and enormous CO_2 emissions. They are expected to increase in likelihood and severity in the near future due to factors like climate change [1]. In Germany, while precautionary measures are established for residential areas and industrial plants through demand planning and pre-assessment of potential hazards, a comparable approach for forest fires is lacking. Consequently, emergency services often face unclear hazard situations upon alarm activation. The ever-changing vitality of vegetation further complicates the hazard potential due to seasonal influences, pest infestations, or droughts. Moreover, forest fires are highly dynamic situations that can engage a large number of emergency personnel. Therefore, early and comprehensive information gathering is essential to enable timely reactions [2]. In addition to static techniques like watchtowers or wireless sensor networks, UAV-based automatic remote sensing platforms have been developed [3]. Traditionally, these platforms have utilized deep learning models such as Convolutional Neural Networks and Long Short-Term Memory networks to detect smoke through image classification tasks [4]. Recent breakthroughs in large language modeling and combining LLMs with vision capabilities extend what is possible with computer vision [5,6]. Specifically, VLMs enable detailed captioning of images combined with advanced reasoning,

 doi:10.20944/preprints202503.1391.v1

thereby enhancing scene understanding. VLMs are increasingly used in real-world applications such as autonomous driving or robotics [7]. Moreover, multiple works discuss using VLMs for general remote sensing tasks [8,9].

While Wei and Kulkarni [10] demonstrated the potential of using VLMs for wildfire detection, this technology can be significantly enhanced to provide fire brigades with a better understanding of the situation and to structure outputs for seamless integration into user interfaces like websites. These situation descriptions could contain detailed information about the fire and its environment, formatted simply and understandably.

2. Materials and Methods

2.1. Wildfires

2.1.1. Wildfire Detection Methods

While in densely populated areas, most wildfires are reported by the public through visitors of the forest or people that live nearby, studies have shown that in remote locations, about 30% to 70% of fires are not detected for a long time [11]. Therefore, specialized methods like towers staffing human watchers have been established. Although these trained personnel perform very well in detecting even small fires at an early stage, cost concerns have driven the development of more automated systems [12]. One of these is the use of camera-based watchtowers. This technology utilizes visible light or infrared cameras in combination with machine learning algorithms to detect smoke or fire glow. Through triangulation, a reasonable estimation of the fire location can be calculated. However, the calculation is prone to errors, especially in hilly or mountainous terrain [12,13]. Satellites are also used in forest fire detection, which can monitor huge areas and be used for other purposes. The use of onboard AI has helped to combat false alarms, such as sun glints or water bodies. However, satellites are very dependent on their orbit times, making detection times very long in most cases. Furthermore, they cannot provide a live picture of the situation. Clouds or smoke can also hinder fire detection, even with IR sensors. Another factor is that, while prices in the space industry have come down in the last century, they still pose a significant investment [14]. Wireless Sensor Networks (WSNs) utilize small, self-sufficient gas, humidity, or temperature sensors. These sensors are placed strategically every 100 - 300 m and communicate via low-power wireless networks or 4G / LTE to form a mesh. This technology is particularly effective in detecting small ground-level fires. However, due to the amount of needed sensors and the work required to place them, it is neither financially nor logistically feasible to place them in large areas [13,15]. In the last years, the use of UAVs in battling forest fires has increased. They can roughly be split into three categories. The first is mainly consumer or hobby-grade drones that are deployed for a short-distance lookout or observation to navigate the fire brigade to a previously detected hot spot or provide a live feed of the ongoing extinguishing actions. These drones usually have a relatively short flight time and are steered by the personnel on the ground. The second category holds specially built long-range, fixed-wing drones that automatically patrol preset areas and alert firefighters if a fire is detected. And lastly, some research has also been done on fighting the fires with the help of drone swarms [14,16].

2.1.2. Wildfire Description

Our approach is to enhance the information available to the fire brigades after the smoke detection. Therefore we define several categories that should be answered by the forest fire VLM in the following section. Forest and vegetation fires are complex events that are influenced by a variety of factors. In addition to the classic elements of the fire triangle (oxygen, energy, fuel), environmental factors play a critical role. These are summarized in the so-called fire behavior triangle of topography, weather, and fuel properties [17]. The topography of the location and the resulting fire behavior can be derived from geo data. Important meteorological parameters such as solar radiation, wind direction, and wind speed can be determined from local weather station data. A central aspect is still the nature of the fuel [18]. To better assess this, one of the labels used, for example, rates the vitality of the surrounding trees

 doi:10.20944/preprints202503.1391.v1

in the categories "vital", "moderately vital", "declining", "dead", "cannot be determined" and "no forest fire visible" [19]. Trees that are dead or infested by pests, in particular, have a low resistance and can accelerate the spread of fire. Observations in Germany also show that forest fires spread particularly quickly in spring before flowering, as the trees are not yet fully supplied with water and have dried out over the winter. This information is an essential factor in predicting the further spread of the fire [20]. For further operational planning, it is also important to assess the potential risk to people or infrastructure. For this purpose, specific questions are asked, such as: "Are people visible near the forest fire?" or "Is infrastructure visible near the forest fire?" [19]. Another decisive criterion is the inspection of the fire. As many fire brigades in Germany work voluntarily, false alarms repeatedly lead to volunteers being torn from their private or professional lives [21]. Possible false alarms can be caused, for example, by controlled burning of green cuttings, fog, or swirling dust. To ensure that a forest fire is recognized reliably, the following questions are therefore asked: "Can smoke from a forest fire be seen in the picture?", "Can flames from a forest fire be seen in the picture?", "Can it be confirmed that this is an uncontrolled forest fire?". In addition to fire detection, dynamic factors are also relevant in order to assess the development and challenges of firefighting. The following parameters are recorded for this purpose: "What state is the forest fire currently in?", "How big is the fire?", "How intense is the fire?", "What type of fire is it?", "Are there multiple sources of fire?" [17]. The full list of questions and corresponding answer options can be seen in Table 1. The combination of these parameters enables a comprehensive assessment of the situation and supports the emergency services in making well-founded decisions. The early and automated analysis of relevant influencing variables can significantly increase the effectiveness of firefighting.

 doi:10.20944/preprints202503.1391.v1

Table 1. Fields, questions, and answer options for the structured output generation

Field	Question	Options
Smoke	Is smoke from a forest fire visible in the image?	Yes, No
Flames	Are flames from a forest fire visible in the image?	Yes, No
Uncontrolled	Can you confirm that this is an uncontrolled forest fire?	Yes, Closer investigation required, No forest fire visible
Fire State	What is the current state of the forest fire?	Ignition Phase, Growth Phase, Fully Developed Phase, Decay Phase, Cannot be determined, No forest fire visible
Fire Type	What type of fire is it?	Ground Fire, Surface Fire, Crown Fire, Cannot be determined, No forest fire visible
Fire Intensity	What is the intensity of the fire?	Low, Moderate, High, Cannot be determined, No forest fire visible
Fire Size	What is the size of the fire?	Small, Medium, Large, Cannot be determined, No forest fire visible
Fire Hotspots	Does the forest fire have multiple hotspots?	Multiple hotspots, One hotspot, Cannot be determined, No forest fire visible
Infrastructure Nearby	Is there infrastructure visible near the forest fire?	Yes, No, Cannot be determined, No forest fire visible
People Nearby	Are there people visible near the forest fire?	Yes, No, Cannot be determined, No forest fire visible
Tree Vitality	Describe the vitality of the trees around the fire.	Vital, Moderate Vitality, Declining, Dead, Cannot be determined, No forest fire visible

2.2. Evolonic

Evolonic is a student research team at Friedrich-Alexander-University Erlangen in close cooperation with the Fraunhofer IISB. The authors are currently team members or have been in the past.

2.2.1. NF4 UAV

An electric VTOL (Vertical Takeoff and Landing) was previously designed and built by Evolonic. The drone features a separate lift and thrust powertrain as shown in Figure 1. It combines the advantages of a rotary-wing drone, such as the lack of a need for a runway and the ability to stop midair, with the long-range and efficient flight of a fixed-wing aircraft [22]. As the powertrains for the two flight modes are separate, this also serves as an added layer of redundancy. The drone has a wingspan of just under three meters and a cruising speed of 65 km/h, which leads to a range of about 100 km per flight until it has to be recharged. Charging is done with a mobile base station that can be deployed to strategic hot spots. The UAV is equipped with a front-facing camera and an onboard companion computer in order to detect smoke. It also has the ability to send data, pictures, and control commands via LTE or a direct radio-link.

 doi:10.20944/preprints202503.1391.v1

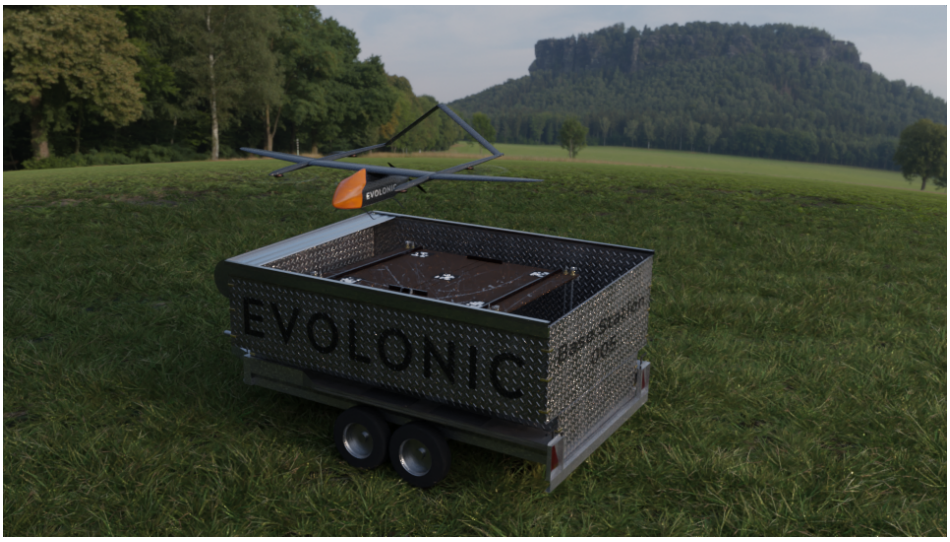
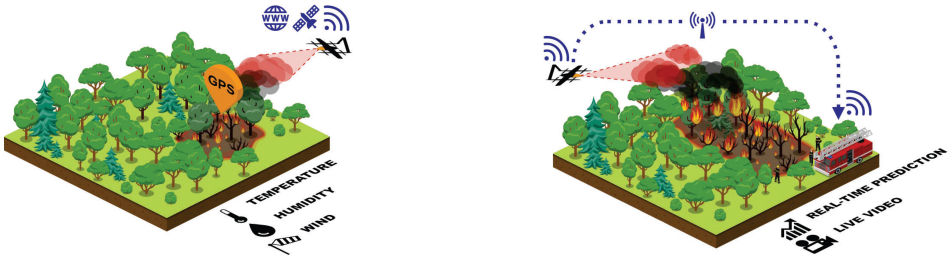


Figure 1. Rendering of the eVTOL drone that is used for this project with the basestation

2.2.2. Automatic operations

The operation of this system centers around the automatic surveillance of predefined forest areas with these UAVs. Depending on the forest fire danger index [11] the system will schedule more or fewer flights per day. The UAV will automatically take off from the base station and follow its preplanned patrol route. If no smoke is detected the drone will fly back to the base station and recharge for the next flight. If smoke is detected by the onboard AI an alert is automatically sent to the fire brigade dispatch center (Figure 2 (a)). The drone will then fly towards the detected hotspot, thus increasing the location accuracy. To support the extinguishing efforts, the UAV will then continue to orbit overhead and provide a live feed from the situation (Figure 2 (b)).



(**a**) (**b**)

Figure 2. Isometric illustrations of **(a)** the drone detecting a wildfire and **(b)** the drone monitoring the situation and providing an overview of the situation for fire brigades.

2.2.3. Software Architecture

A ROS2-based architecture is implemented, containing a node for smoke detection using YOLOv8 instance segmentation. This detection model was annotated and trained by Evolonic from previous forest fires, internet images and data from fire brigades across Germany. Positive detections are sent to a web backend via MQTT together with environmental data, geo-referencing, and a time stamp. A web application, made available to dispatch centers and fire brigades, visualizes this data with a live stream and maps.

2.2.4. Test Flights

The system has been evaluated in two separate field studies. The first of these was a durability test of the drone, which took place in southern Germany in the summer of 2023. This test was done to ensure the long-term durability of the UAV. Over a span of about 30 days almost 50 flights have been

 doi:10.20944/preprints202503.1391.v1

conducted, from which no major system failure or material fatigue were observed. Only one servo for one of the ailerons was damaged during transportation, but this was detected during the preflight inspection. During this test, the stability of the connection system was also evaluated. Although the LTE-connectivity can be difficult, especially in sparsely populated areas, the connection at the cruising altitude of 120 m above ground was proven to be very stable. While the overall durability test was a complete success, the detection pipeline could not be verified due to a lack of forest fires during that time at this location. Therefore, a second test was planned. Here the goal was to measure the time and maximum distance needed to successfully detect a wildfire. A consortium of different research teams from the German Federal Institute for Material Research and Testing (BAM), the Otto von Guericke University Magdeburg, and OneSeven set up a patch of forest to be deliberately set on fire to study various parameters such as spreading speed, new extinguishing foams or the detection with long-range UAVs. The smoke detection system trained from previous data was able to detect smoke on images captured from this event.

2.3. VLMs

2.3.1. Architectures

Vision-Language Models (VLMs) integrate a Large Language Model (LLM) with the ability to process visual data in the form of images or videos. Various architectures have been explored to achieve this integration, with two predominant approaches emerging in state-of-the-art VLMs: fully auto-regressive architecture and cross-attention architecture. The first architecture comprises a pre-trained vision encoder and a multimodal adapter [23]. The vision encoder processes visual inputs, while the multimodal adapter maps the outputs of the encoder to the input format required by the LLM.

2.3.2. State of the Art Models

We initiated our investigation by evaluating both closed-source and open-source Vision-Language Models (VLMs) on our designated evaluation dataset. Closed-source VLMs serve as benchmarks for the current state-of-the-art, while fine-tunable open-source models are of primary interest to our study. To select suitable VLMs for extracting detailed and structured information from forest fire images, we consider several key factors, such as the computational cost of running a model, either through API credits or on local hardware, results in common benchmarks, and the availability of established fine-tuning frameworks or associated code. Our goal is to identify open-source models that can operate within the 24 GB VRAM capacity of an available NVIDIA RTX 3090 GPU, with an ideal target of running on Evolonic's 16 GB NVIDIA Jetson Orin NX. Assuming 16-bit weights, this constraint allows for models up to approximately 10 billion total parameters; however, quantization techniques could potentially accommodate larger models. Various benchmarks exist for comparing VLMs, ranging from Optical Character Recognition (OCR) to mathematical reasoning and evaluating tendencies of the model towards hallucinations [7]. The OpenVLM Leaderboard on Huggingface aggregates 31 such benchmarks, executable via the VLMEvalkit [24]. For compatibility with fine-tuning frameworks, we prioritize models that are compatible with LLaMA-Factory [25] or Unsloth [26], as well as custom training code provided by the model developers. Based on these criteria, we select Google DeepMind's Gemini 2.0 models and OpenAI's GPT-4o family as closed-source references. For open-weight models, we choose the Qwen2.5-VL family, which includes models with 3B, 7B, and 72B parameters [6]. The InternVL 2.5 series of models [27] performs comparably to or slightly worse than Qwen2.5-VL and is currently unsupported by the aforementioned fine-tuning frameworks. All of these models demonstrate strong performance on general benchmarks such as MMMU [28], and MME-RealWorld [29], which more closely resembles our forest fire description task.

2.3.3. Prompting and Structured Outputs

Instead of allowing the VLM to freely generate outputs, we use structured outputs to only allow a given JSON pattern as an answer. This JSON scheme contains the 11 answer categories derived in

 doi:10.20944/preprints202503.1391.v1

Section 2.1.2, each allowing an enumeration of 2 to 6 possible text answers to choose from. All of these fields are enforced to be present in the output of the LLM, without allowing any additional fields. While binary classification is only used with the fields *Smoke* and *Flames*, the other fields also allow looser answers. Most answers accept "Cannot be determined" in case the situation cannot be fully resolved from the image. In the case of the field *Uncontrolled*, which answers if the forest fire is out of control or if it might be a controlled burning, the answer "Closer investigation required" is possible. The model is also prompted to answer subsequent fields with "No forest fire visible" in case the binary classification is negative.

2.3.4. Datasets

We are working with three different datasets to train and evaluate our approach. The first two are new datasets for the specific task of forest fire description. The dataset *ForestFireInsights-Eval* consists of 301 images, while the training set *ForestFireInsights-Train* consists of 1196 images. The imagery is sourced from four distinct origins: Evolonic drone footage, drone footage obtained from various fire brigades across Germany, a publicly accessible dataset from the University of Split [30,31], and Internet videos. Evolonic's contributions include images captured during an actual wildfire in Tennenloher Forst near Erlangen, Germany (Figure 3), a controlled smoke test in Erlangen, and a forest fire simulation near Calvörde, Germany. The evaluation dataset is publicly available and exclusively features images sourced from Evolonic and the University of Split. In contrast, the training dataset incorporates frames extracted from internet videos and confidential images provided by fire brigades, many of which were shared under non-redistribution agreements. Annotations were created by the authors containing the fields explained in Section 2.1.2. These keys are then put in a predefined JSON answer scheme and converted to a single text string. For evaluation purposes, the original JSON-like Python dictionary is also saved in the dataset. The annotation workflow included creating preliminary suggestions with an off-the-shelf VLM before manually editing all fields in a web-hosted Argilla environment.

Figure 3. Sample image from ForestFireInsights-Eval: Real forest fire near Erlangen, captured by Evolonic using a DJI Mavic drone

The third dataset used in this work is the test set of the FIgLib dataset [4], consisting of 4880 images from stationary wildfire cameras in California. These were not manually annotated by us but contain a timestamp relative to the outbreak of a forest fire visible in the images. Only the smoke

 doi:10.20944/preprints202503.1391.v1

detection was therefore validated on this dataset, allowing a comparison with other state-of-the-art methods. We developed a script to convert the images to a format compatible with our evaluation code and publish the converted dataset.

It is ensured that the training dataset does not contain any images from fire events that overlap with those used in either one of these evaluation datasets. The distribution of annotations across all categories for all three datasets is documented in Appendix A.

2.3.5. Training

Finetuning VLMs on custom datasets mostly relies on 3 different approaches: Full fine-tuning, Low-Rank Adaption (LoRA) [32] and Quantized Low-Rank Adaption (QLoRA) [33]. While the first finetunes all model weights, the latter only trains on a small subset, typically less than 1 %. Instead of relying on FP16 values as with full finetuning and LoRA, QLoRA utilizes a 4-bit quantization of the base model [33]. This leads to a further reduction in required training memory compared to LoRA, which is already multiple times more memory-efficient than full finetuning. Both LoRA and QLoRA trainings produce adapters that can be merged with the base model afterwards. We are using the LLaMA-Factory framework [25] for finetuning Qwen2.5-VL models with the LoRA method. Different learning rates, batch sizes, and numbers of epochs were tested and evaluated, with the best results shown in Table 2.

Table 2. Hyperparameters and GPUs chosen for our models using the LLaMA-Factory framework

Setting	ForestFireVLM-3B	ForestFireVLM-7B
Learning rate	0.00005	0.0002
Epochs	2	2
Batch Size	1	1
Gradient Accumulation	8 Steps	16 Steps
GPU	NVIDIA RTX 3090	NVIDIA A100 80GB

Figure 4 shows the loss curves for the best training runs of the 3B and 7B models finetuned from Qwen2.5-VL, captured with Weights & Biases.

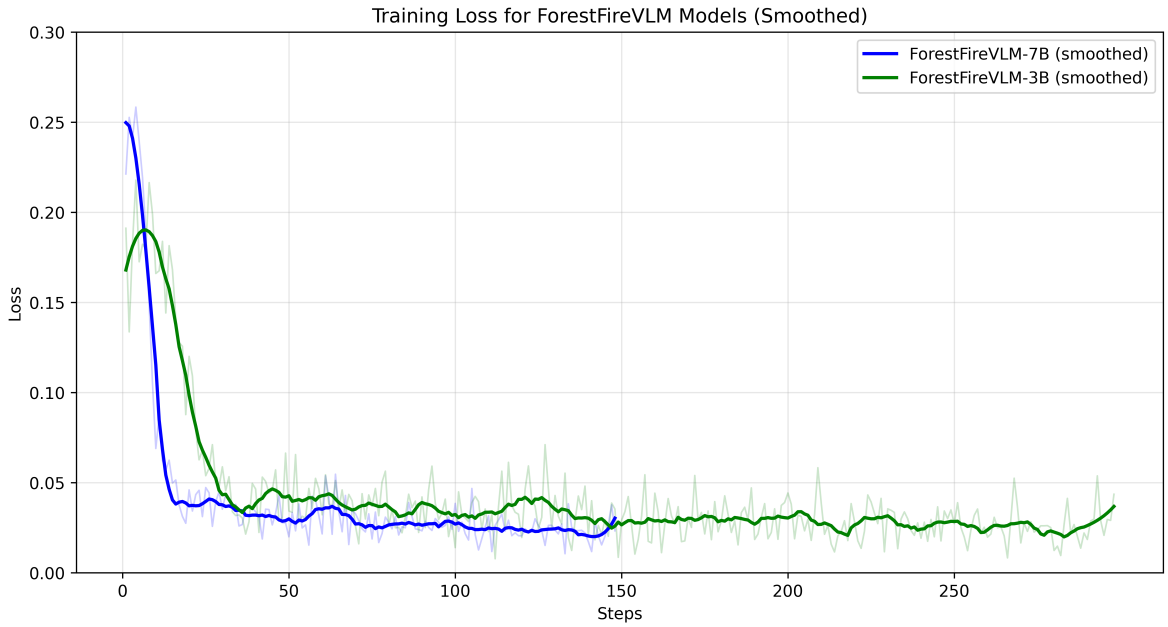


Figure 4. Smoothed training losses for the Qwen2.5-VL finetunes trained on our dataset

 doi:10.20944/preprints202503.1391.v1

2.3.6. Evaluation

Evaluation is performed using a script made publicly available in our repository. First, text predictions from the model are generated on the samples of the evaluation dataset. The script offers two possible backends for this: the Google Gemini API and an OpenAI-compatible endpoint, which can also be used for local inference with compatible frameworks. We are using vLLM [34] for this task, which allows running Qwen2.5-VL models and supports guided decoding backends for generating structured outputs. After all predictions are generated, we validate the structured output, convert the text to a Python dictionary, and compare all keys to the human-annotated groundings from the evaluation dataset. For keys that contain an enumeration of possible values, only the accuracy of correct values is computed, while for the smoke detection also precision, recall, and F_1 score. The metrics are computed using the following equations, also used by other reference works [4,10]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

The separation of prediction generation also allows adding other inference backends without changing the actual evaluation code. All evaluations across inference methods and providers were made with a temperature setting of 0.0, which should improve the reproducibility of our results.

The context length depends on the number of tokens required for the text prompt as well as the tokenized image. For Qwen VLMs and their finetunes the number of tokens for our text prompt including the chat template is 526, while the number of image tokens is dependent on the image pixel resolution [6,35]:

$$\text{Image Tokens} = \frac{Heigth \times Width}{(14 \times 14) \times 4} \tag{5}$$

For the evaluations performed with our finetuned VLM on our own evaluation set, we set the context length of the model to 4500 tokens, as in the training. The FIgLib dataset contains images of a higher resolution (4 to 6 Megapixels), requiring a context length of nearly 9000 tokens. Due to the larger image resolution and number of entries the evaluations on this dataset were done with batched inference using a thread pool with up to 64 parallel workers. Wei and Kulkarni [10] obtain better FIgLib results with horizon cropping. As this approach is only applicable to fixed-horizon images, unlike UAV imagery, we do not implement this approach and solely compare to zero-shot evaluation. For cloud-based models we did not evaluate Gemini 2.0 Pro, as only 5 requests per minute were allowed at the time of writing.

2.3.7. Computing and API checkpoints

Training and evaluation processes were conducted on a combination of local and rented cloud GPUs. Evaluation runs typically used an NVIDIA RTX 3090 or an NVIDIA RTX 4060 Ti for the *ForestFireInsights-Eval* dataset, requiring between 16 and 24 GB of VRAM. The parallelized evaluation runs on the larger FIgLib dataset required using an NVIDIA A100 with 80 GB of VRAM. Training runs were done on similar hardware, as noted in Table 2. Closed weight models from Google DeepMind and OpenAI were employed in conjunction with their respective APIs. Table 3 shows the respective checkpoints or versions used for cloud-based API models.

 doi:10.20944/preprints202503.1391.v1

Table 3. Model checkpoints or versions from models used with Google's or OpenAI's API endpoints

Model	Version or checkpoint
Gemini Pro 1.5	Version 002
Gemini Flash 2.0	Version 001
Gemini Flash 2.0 Lite	Version 001
Gemini Pro 2.0	Experimental 2025-02-05
GPT-4o	2024-08-06
GPT-4o mini	2024-07-18

3. Results

3.1. ForestFireInsights-Eval

The evaluation dataset was tested using closed, cloud-based VLMs, open-weight models, and our fine-tuned models. We present the results for all output fields grouped by their coarse category. The first category encompasses binary detection of smoke and flames and estimating whether the fire is uncontrolled or if further investigation is required. The correct percentages for these fields are visualized in Table 4.

Table 4. Performance in smoke and fire detection

Model	Flames (%)	Smoke (%)	Uncontrolled (%)	Average (%)
ForestFireVLM-7B	**98.0**	95.0	**77.1**	**90.0**
ForestFireVLM-3B	96.0	95.4	75.8	89.0
Gemini Pro 1.5	94.7	91.4	65.1	83.7
Gemini Flash 2.0	96.4	95.0	75.4	88.9
Gemini Flash 2.0 Lite	96.7	93.7	58.8	83.1
Gemini Pro 2.0	90.4	**95.7**	53.2	79.7
GPT-4o	95.7	94.7	47.5	79.3
GPT-4o mini	95.0	94.7	43.9	77.9
Qwen2.5-VL-3B	93.4	92.0	39.2	74.9
Qwen2.5-VL-7B	92.7	79.7	32.2	68.2
Qwen2.5-VL-72B	93.7	83.4	34.6	70.5

Most models demonstrate exemplary performance in binary flame and smoke detection, even without additional fine-tuning. Notably, the 3B version of Qwen2.5-VL outperforms its larger variants in smoke detection by a significant margin. The best zero-shot accuracy for smoke detection is achieved by Gemini 2.0 Pro at 95.7%. In the Uncontrolled category, both our fine-tuned models have an advantage over most other models, except for Gemini Flash 2.0. Here, Gemini models generally perform better than both OpenAI and Qwen models. The best overall score is achieved by ForestFireVLM-7B at 90.0%, closely followed by the 3B variant and Gemini Flash 2.0.

Table 5 presents quantitative metrics for fire description and their corresponding scores. Our fine-tuned models achieve superior results across these categories. Specifically, the 7B model excels in fire intensity and size, while the 3B model performs best for fire hotspots. Gemini demonstrates the highest performance among off-the-shelf models, with Gemini 2.0 Pro leading by a significant margin. Notably, Qwen2.5-VL-3B outperforms its larger variants once again, showing a smaller performance gap compared to the 72B model than to the 7B model.

 doi:10.20944/preprints202503.1391.v1

Table 5. Performance in quantitative fire metrics

Model	Fire Hotspots (%)	Fire Intensity (%)	Fire Size (%)	Average (%)
ForestFireVLM-7B	78.1	**74.1**	**81.1**	**77.7**
ForestFireVLM-3B	**82.1**	69.8	80.4	77.4
Gemini Pro 1.5	69.1	48.5	51.5	56.4
Gemini Flash 2.0	58.8	34.2	51.2	48.1
Gemini Flash 2.0 Lite	79.4	47.2	67.4	64.7
Gemini Pro 2.0	76.7	63.1	74.1	71.3
GPT-4o	26.3	21.9	21.6	23.3
GPT-4o mini	35.6	26.6	27.9	30.0
Qwen2.5-VL-3B	55.8	37.9	38.2	44.0
Qwen2.5-VL-7B	9.0	4.3	3.7	5.6
Qwen2.5-VL-72B	28.6	21.9	21.6	24.0

Table 6 displays the performance metrics for qualitative fire description, including fire state, fire type, and their average percentages. Our fine-tuned models, ForestFireVLM-7B and ForestFireVLM-3B, again demonstrate superior performance in both categories. Notably, ForestFireVLM-7B achieves the highest scores across all metrics, with 64.1% for fire state, 71.8% for fire type, and an average of 67.9%. The smaller variant, ForestFireVLM-3B, also performs well with 61.5% for fire state, 68.1% for fire type, and an average of 64.8%. Among the off-the-shelf models, Gemini Pro 2.0 leads with scores of 53.5% for fire state, 66.5% for fire type, and an average of 60.0%, closely followed by Gemini Flash 2.0 Lite with an average of 55.2%. The Qwen2.5-VL models show varying performance, with the 3B and 72B variants outperforming the 7B counterpart.

Table 6. Performance in the qualitative fire description

Model	Fire State (%)	Fire Type (%)	Average (%)
ForestFireVLM-7B	**64.1**	**71.8**	**67.9**
ForestFireVLM-3B	61.5	68.1	64.8
Gemini Pro 1.5	41.5	63.5	52.5
Gemini Flash 2.0	44.5	62.1	53.3
Gemini Flash 2.0 Lite	46.5	63.8	55.2
Gemini Pro 2.0	53.5	66.5	60.0
GPT-4o	29.9	52.2	41.0
GPT-4o mini	31.9	55.8	43.9
Qwen2.5-VL-3B	36.9	38.2	37.5
Qwen2.5-VL-7B	12.0	34.6	23.3
Qwen2.5-VL-72B	28.9	47.8	38.4

Table 7 illustrates the performance metrics for various environmental factors, including infrastructure nearby, people nearby, tree vitality, and their average percentages. Our fine-tuned models, ForestFireVLM-7B and ForestFireVLM-3B, exhibit strong performance across these categories. The 7B model achieves the highest overall average of 67.7%, with notable scores of 66.1% for people nearby and 62.8% for tree vitality. The smaller variant, ForestFireVLM-3B, leads in infrastructure nearby with a score of 74.4% and maintains a competitive average of 64.8%. Among the off-the-shelf models, Gemini Flash 2.0 demonstrates commendable performance with an average of 62.5%, while Gemini Pro 1.5 follows closely with 56.3%. The Qwen2.5-VL models show mixed results, with the 72B variant outperforming its smaller counterparts in terms of average percentage.

 doi:10.20944/preprints202503.1391.v1

Table 7. Performance in environmental fields

Model	Infrastructure Nearby (%)	People Nearby (%)	Tree Vitality (%)	Average (%)
ForestFireVLM-7B	74.1	**66.1**	**62.8**	**67.7**
ForestFireVLM-3B	**74.4**	58.5	61.5	64.8
Gemini Pro 1.5	66.5	57.5	44.9	56.3
Gemini Flash 2.0	70.1	56.5	60.8	62.5
Gemini Flash 2.0 Lite	51.2	37.5	30.6	39.8
Gemini Pro 2.0	60.1	43.2	43.5	48.9
GPT-4o	31.9	52.8	39.9	41.5
GPT-4o mini	33.2	47.5	41.9	40.9
Qwen2.5-VL-3B	56.2	32.6	32.2	40.3
Qwen2.5-VL-7B	51.2	44.9	20.9	39.0
Qwen2.5-VL-72B	62.5	44.2	49.2	51.9

Table 8 presents an overview of the comprehensive performance metrics across all tasks described in the tables above, including detection, fire quantitative description, fire qualitative description, environmental factors, and their overall averages. The overall average here is formulated as the average across all output fields. Our fine-tuned models, ForestFireVLM-7B and ForestFireVLM-3B, clearly outperform other models in most categories. The 7B model achieves an impressive overall average of 76.6%, with scores of 90.0% for detection, 77.7% for fire quantitative description, 67.9% for fire qualitative description, and 67.7% for environmental factors. The smaller variant, ForestFireVLM-3B, also demonstrates strong performance with an overall average of 74.8%. Among the off-the-shelf models, Gemini Pro 2.0 shows commendable results with an overall average of 65.5%, while Gemini Flash 2.0 follows closely with 64.1%. Remarkably, all evaluated families of models are close to their individual peers. The Qwen2.5-VL models exhibit varied performance, with the 7B variant lagging behind its counterparts.

Table 8. Total model performances

Model	Detection (%)	Fire Quantitative (%)	Fire Qualitative (%)	Environmental (%)	Overall (%)
ForestFireVLM-7B	**90.0**	**77.7**	**67.9**	**67.7**	**76.6**
ForestFireVLM-3B	89.0	77.4	64.8	64.8	74.8
Gemini Pro 1.5	83.7	56.4	52.5	56.3	63.1
Gemini Flash 2.0	88.9	48.1	53.3	62.5	64.1
Gemini Flash 2.0 Lite	83.1	64.7	55.2	39.8	61.2
Gemini Pro 2.0	79.7	71.3	60.0	48.9	65.5
GPT-4o	79.3	23.3	41.0	41.5	46.8
GPT-4o mini	77.9	30.0	43.9	40.9	48.5
Qwen2.5-VL-3B	74.9	44.0	37.5	40.3	50.2
Qwen2.5-VL-7B	68.2	5.6	23.3	39.0	35.0
Qwen2.5-VL-72B	70.5	24.0	38.4	51.9	46.9

3.2. Examples

The following images were randomly picked from the evaluation dataset and show the outputs with ForestFireVLM-7B. In Figure 5, an example image from the forest fire in Tennenloher Forst near Erlangen is demonstrated. The human-annotated ground truths were set similarly for most categories but setting the Uncontrolled field to "Closer investigation required", People Nearby to "Cannot be determined" and the Tree Vitality to "Moderate Vitality".

 doi:10.20944/preprints202503.1391.v1

Category	VLM Answer
Smoke	Yes
Flames	No
Uncontrolled	Yes
Fire State	Ignition Phase
Fire Type	Surface Fire
Fire Intensity	Low
Fire Size	Small
Fire Hotspots	One hotspot
Infrastructure Nearby	No
People Nearby	No
Tree Vitality	Vital

Figure 5. Answers from ForestFireVLM-7B for an image of a real forest fire in 2022

An image without any fire is tested in Figure 6, where the ForestFireVLM-7B correctly outputs "No forest fire visible" for all additional categories. These results match our human-made annotations.

Category	VLM Answer
Smoke	No
Flames	No
Uncontrolled	No forest fire visible
Fire State	No forest fire visible
Fire Type	No forest fire visible
Fire Intensity	No forest fire visible
Fire Size	No forest fire visible
Fire Hotspots	No forest fire visible
Infrastructure Nearby	No forest fire visible
People Nearby	No forest fire visible
Tree Vitality	No forest fire visible

Figure 6. ForestFireVLM-7B correctly outputs "No forest fire visible" for all categories when no forest fire is in the image

For the last image, once again, most answers match our annotations, with People Nearby being set to "Yes" and Tree Vitality set to "Moderate Vitality" in the human-made annotations.

Category	VLM Answer
Smoke	Yes
Flames	No
Uncontrolled	Closer investigation required
Fire State	Ignition Phase
Fire Type	Cannot be determined
Fire Intensity	Low
Fire Size	Small
Fire Hotspots	One hotspots
Infrastructure Nearby	No
People Nearby	Cannot be determined
Tree Vitality	Vital

Figure 7. ForestFireVLM-7B captions for a frame from a test with the fire brigade Erlangen in 2022

 doi:10.20944/preprints202503.1391.v1

These examples show both the good results of our fine-tuned model and the ambiguity of our ground truth labels. The decision between setting a definitive answer or "Cannot be determined" can be especially hard. The human-made annotations are, therefore, not perfect and will vary between different annotators and with different knowledge of the imaged forest fire events.

3.3. FIgLib Test Dataset

We evaluate the zero-shot smoke detection capabilities on the test set of the Figlib dataset. Despite training with a context length of 4500 tokens, inferencing with a context length of 9000 tokens works well and improves the model performance significantly compared to the baseline Qwen2.5-VL model. Table 9 shows that our finetuned models outperform other VLM-based zero-shot approaches. Once again, Qwen2.5-VL-3B outperforms its 7B counterpart without further finetuning. After finetuning, the results are similar to our findings with the *ForestFireInsights-Eval* dataset.

Table 9. VLM-based zero-shot smoke detection results on the FigLib dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)
ForestFireVLM-7B	**78.5**	98.6	**57.2**	**72.4**
ForestFireVLM-3B	76.3	98.8	52.6	68.6
Gemini 1.5 Pro	70.0	**100.0**	39.2	56.3
Gemini 2.0 Flash	74.1	96.9	49.1	65.2
Gemini 2.0 Flash Lite	71.5	95.6	44.2	60.5
Qwen2.5-VL-3B	70.4	91.3	44.1	59.5
Qwen2.5-VL-7B	60.3	**100.0**	19.5	32.7
PaliGemma [10]	52.1	**100.0**	3.0	5.7
Phi3 [10]	52.6	**100.0**	4.0	7.6
GPT-4o [10]	74.5	95.2	50.6	66.1
LLaVA 7B [10]	67.5	87.6	39.2	54.1

Lastly, we compare the performance of our fine-tuned VLMs to other approaches from Dewangan et al. [4] and Wei and Kulkarni [10] in Table 10. SmokeyNet with a single frame combines a ResNet34 with a Vision Transformer, while the other variant uses three frames and an additional LSTM (Long Short-Term Memory) network. As the human experts could look at multiple images before deciding and the latter SmokeyNet utilizes three frames at once, they are not directly comparable to our approach. The Horizon Tiling approach with LLaVA 7B splits the image in multiple tiles across a fixed horizon line [10] and is therefore also not directly comparable with ours.

Table 10. Comparision to other methods on the FIgLib dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F_1 score (%)
ForestFireVLM-7B	78.5	98.6	57.2	72.4
ForestFireVLM-3B	76.3	**98.8**	52.6	68.6
Human (average of 3) [4]	78.5	93.5	74.4	**82.8**
SmokeyNet (1 frame) [4]	82.5	88.6	75.2	81.3
SmokeyNet (3 frames) [4]	**83.6**	90.9	**76.1**	**82.8**
LLaVA (Horizon Tiling) [10]	81.4	86.5	73.7	79.6

4. Discussion

4.1. Key Takeaways

We show that general-purpose VLMs can classify multiple key attributes of forest fires. Binary classification of smoke or fire visibility is reliable with all VLMs evaluated while describing the fire and the environment is more challenging. We tested closed, cloud-based VLMs from OpenAI and

 doi:10.20944/preprints202503.1391.v1

Google Deepmind and Qwen2.5-VL open weigths models on our evaluation dataset. Gemini Pro 2.0 scores an average of 65.5% on our benchmark, being the best model without further finetuning. We then finetuned Qwen2.5-VL models on our private training dataset with 1.2K images and improved the evaluation to an average of 76.6%. We also validated the smoke detection performance of our approach with the FIgLib dataset, gaining significantly better results on binary smoke classification than previous VLM-based zero-shot methods.

4.2. Implications and Limitations

One significant limitation lies in the annotation process of the forest fire images. All annotations were carried out by the authors themselves, who are not trained fire brigade professionals. Many annotations remain vague and could benefit from annotators who are more experienced with interpreting forest fires of all stages. The success of our fine-tuned models can be explained by its adaption to the annotator's preferences in some part. Another limitation regarding the FIgLib and University of Split datasets might be their contamination towards the VLM pretraining datasets, as the vision encoders used in modern VLMs are trained on almost all publicly available image data. The strong results on Evolonic's own, previously unpublished, drone footage show robustness against this.

4.3. Future Work

As demonstrated in this work, it is possible to finetune VLMs on a broad range of forest fire-related tasks. With about 1.2K samples, we are using a relatively small instruction finetuning dataset, which could be scaled with more labeled data. While running a 3B parameter 16-bit or 7B parameter quantized VLM is possible with 16 GB of memory, matching the available memory on Evolonic's NVIDIA Orin NX onboard computers, we are using desktop machines for now. While memory usage might not be a limiting factor anymore, the latency of running onboard VLMs remains a challenge that has to be addressed.

5. Conclusion

We present a family of state-of-the-art VLMs fine-tuned specifically for forest fire detection and structured description. Our contributions include:

- ForestFireVLM-7B and ForestFireVLM-3B, fine-tuned versions of their Qwen2.5-VL counterparts, publicly available for research and practical applications.
- A framework for detailed descriptions of forest fires in a structured format.
- Improving the VLM-based detection performance on the FIgLib dataset.
- A dedicated evaluation dataset for structured forest fire descriptions and accompanying code for future research in this domain.

While real-time inference of VLMs poses challenges for time-critical tasks such as forest fire detection, we posit that providing additional information can significantly aid fire brigades in their decision-making processes. Unlike previous approaches, our system provides actionable intelligence in a structured format, ready for integration into real-world emergency response workflows. By offering these resources, we aim to support further advancements in the application of VLMs for wildfire management and response. Our evaluation and inference code with additional data can be found on GitHub, while our models and evaluation datasets are hosted on HuggingFace.

Author Contributions: This section reports the main contributions of each author according to the Contributor Roles Taxonomy (CRediT) (https://www.mdpi.com/data/contributor-role-instruction.pdf). Conceptualization, L.S.; methodology, L.S., S.G. and T.R.; software, L.S.; validation, L.S. and S.G.; formal analysis, L.S.; investigation, L.S.; resources, L.S., S.I.; data curation, L.S. and S.G.; writing—original draft preparation, L.S., S.G. and S.I.; writing—review and editing, L.S., S.G., T.R. and S.I.; visualization, L.S. and S.G.; supervision, B.E and M.M; project administration, B.E and M.M; funding acquisition, B.E and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

 doi:10.20944/preprints202503.1391.v1

Data Availability Statement: The evaluation dataset *ForestFireInsights-Eval* as well as a modified version of the FIgLib test dataset are available on HuggingFace. The training dataset *ForestFireInsights-Train* is not made publicly available due to legal reasons.

Acknowledgments: The authors would like to express their gratitude to Adrian Sauer for his leadership as Project Lead. Special thanks are extended to Lorenz Einberger and Dominik Schuler for their work on the drone design and building, to Leonhard Kluge for his responsibility in the design and construction of the basestation, and to Isabella Hufnagl for her design of the web application. Appreciation is also given to Oliver Grau for his expertise in electronics engineering, and to Simon Grau for improving our smoke detection and dataset. Thanks, are also due to Felix Körwer for his coordination of operations, and to Lara Schindhelm and Sebastian Wiederhold for recording forest fire videos in Erlangen, along with all other Evolonic members who contributed to the project. Thanks also go to the Office for Fire and Civil Protection of the city of Erlangen and all fire brigade members for their collaboration.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
LoRA	Low-Rank Adaption
LSTM	Long Short-Term Memory
OCR	Optical Character Recognition
UAV	Unmanned Aerial Vehicle
VLM	Vision Language Model
VTOL	Vertical Takeoff And Landing
WSN	Wireless Sensor Network

Appendix A

The following figures contain the distributions of human-made annotations for their respective categories across the three datasets considered in this work. As the FIgLib-Test dataset is only used for verification of the smoke detection we only require annotations in the "Smoke" field here.

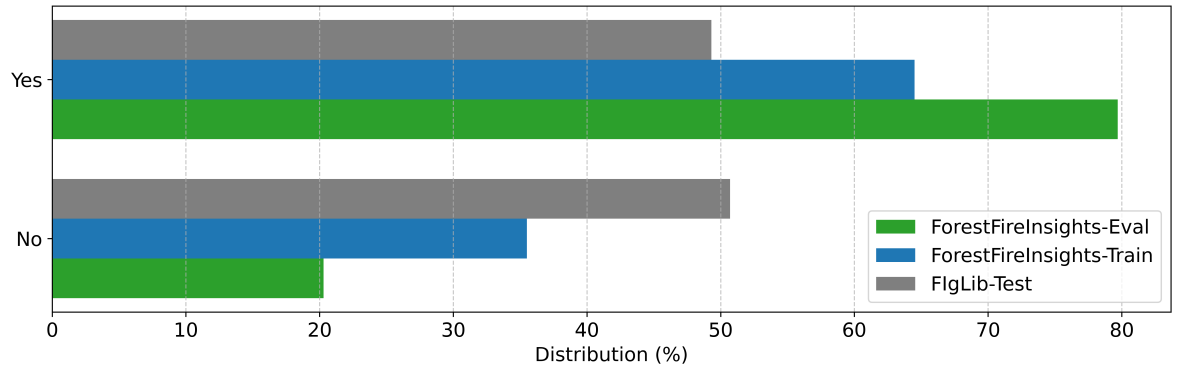


Figure A1. Annotation distribution for "Smoke" in all datasets

 doi:10.20944/preprints202503.1391.v1

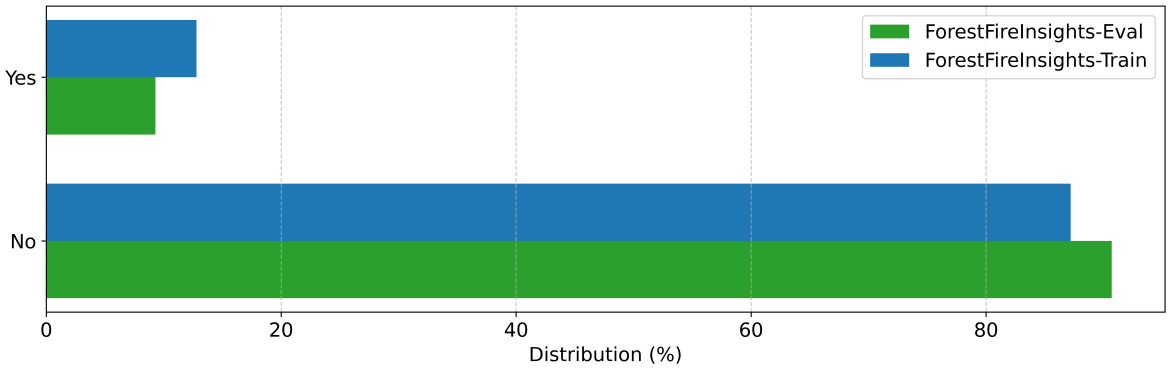


Figure A2. Annotation distribution for "Flames" in all datasets

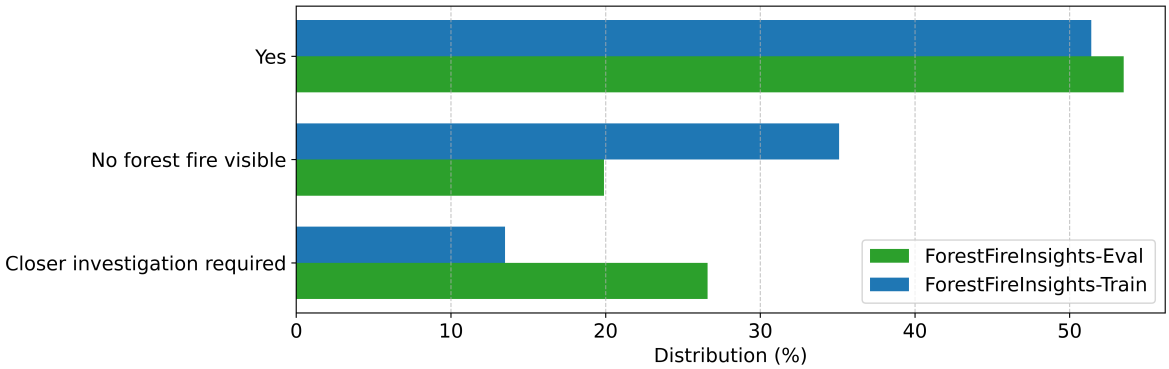


Figure A3. Annotation distribution for "Uncontrolled" in all datasets

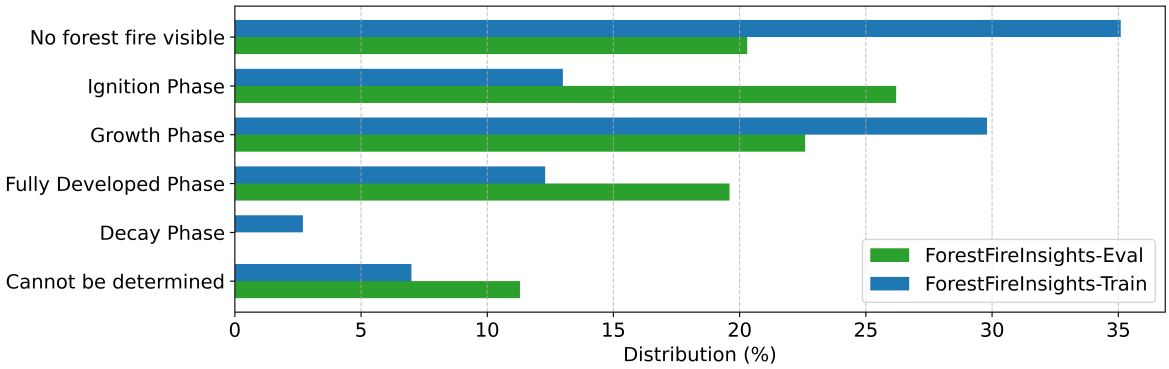


Figure A4. Annotation distribution for "Fire State" in all datasets

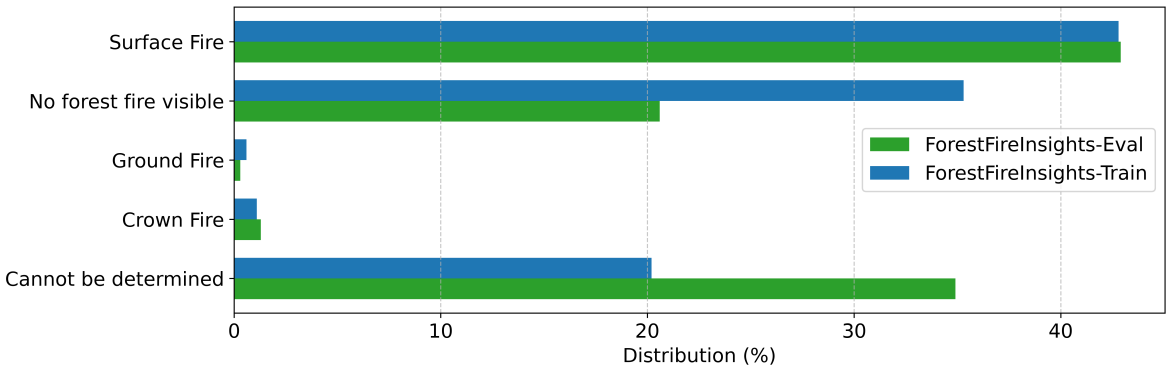


Figure A5. Annotation distribution for "Fire Type" in all datasets

 doi:10.20944/preprints202503.1391.v1

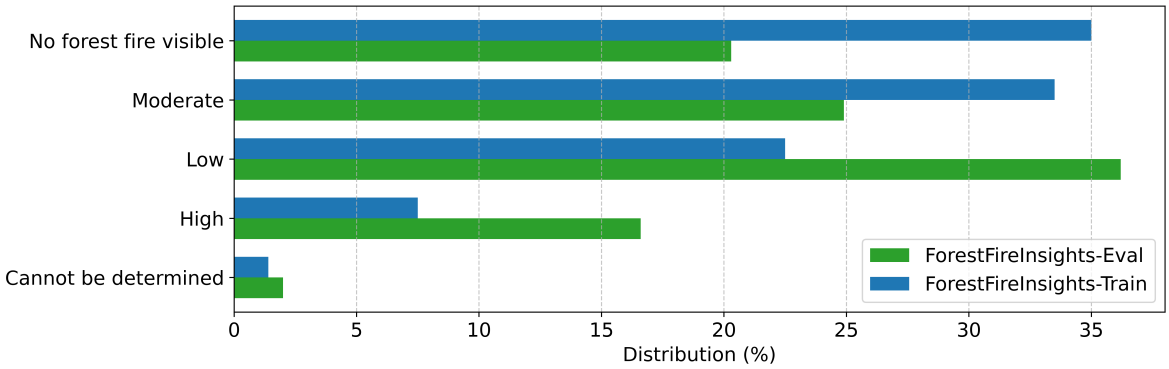


Figure A6. Annotation distribution for "Fire Intensity" in all datasets

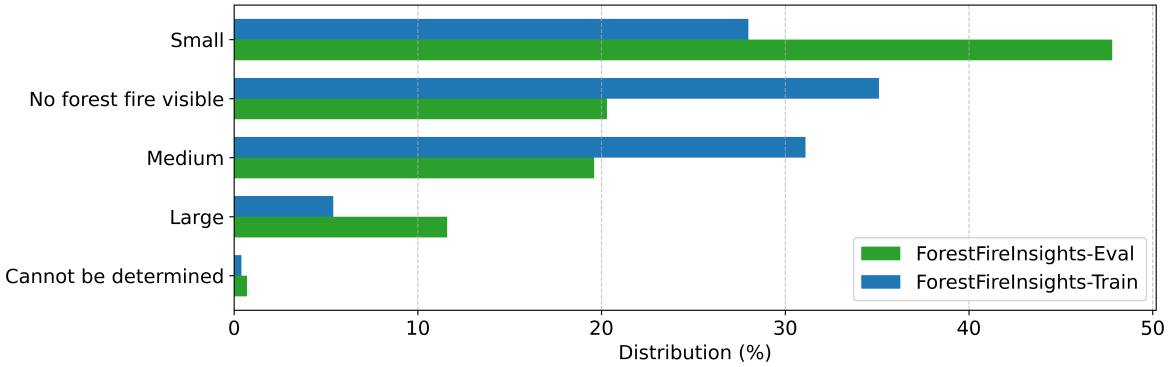


Figure A7. Annotation distribution for "Fire Size" in all datasets

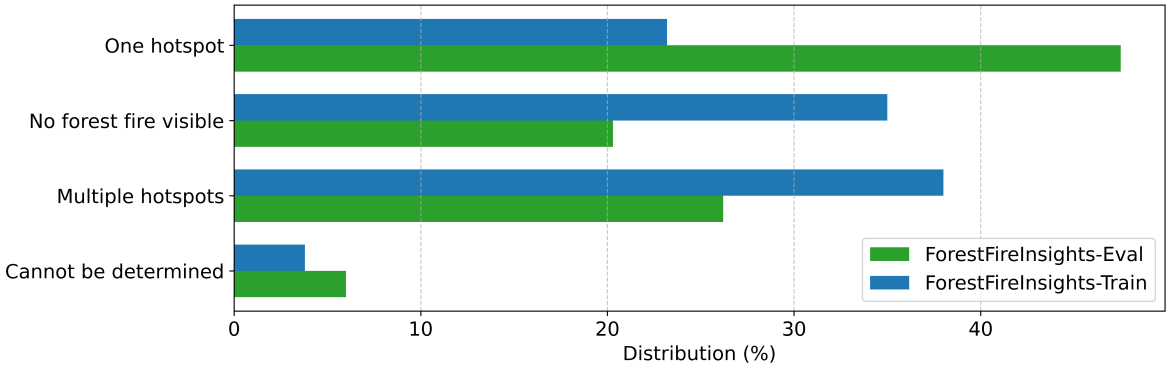


Figure A8. Annotation distribution for "Fire Hotspots" in all datasets

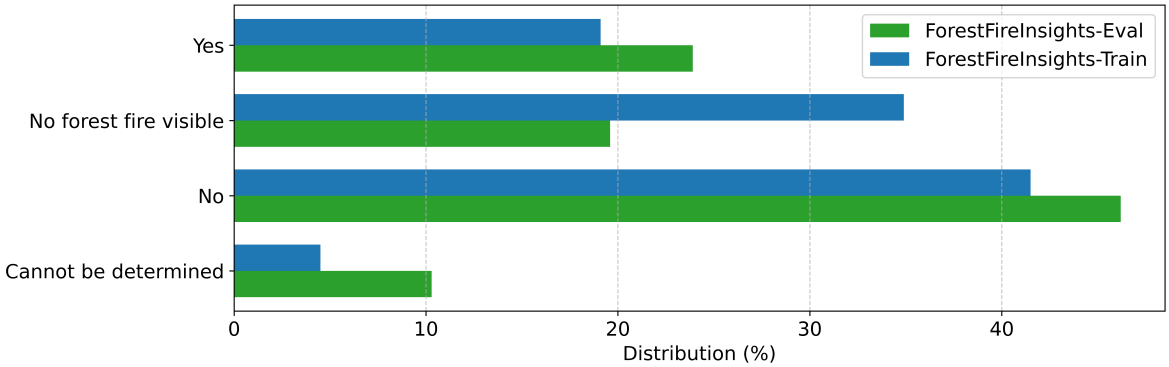


Figure A9. Annotation distribution for "Infrastructure Nearby" in all datasets

 doi:10.20944/preprints202503.1391.v1

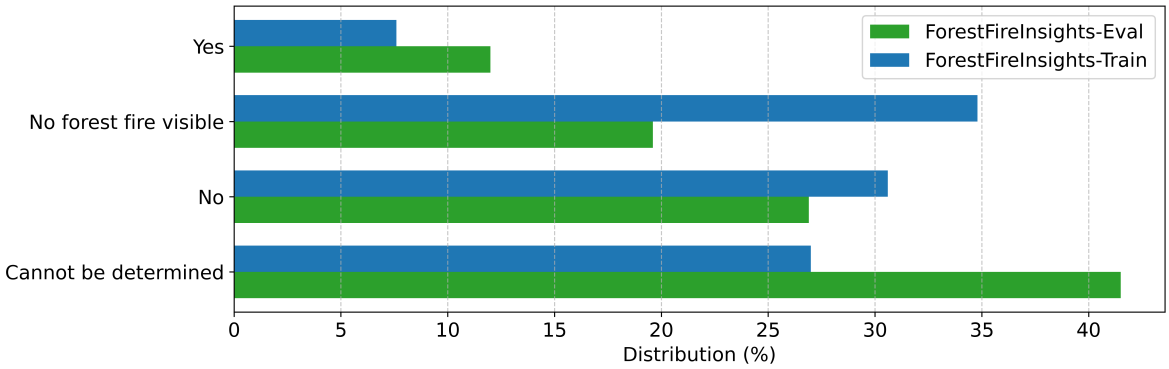


Figure A10. Annotation distribution for "People Nearby" in all datasets

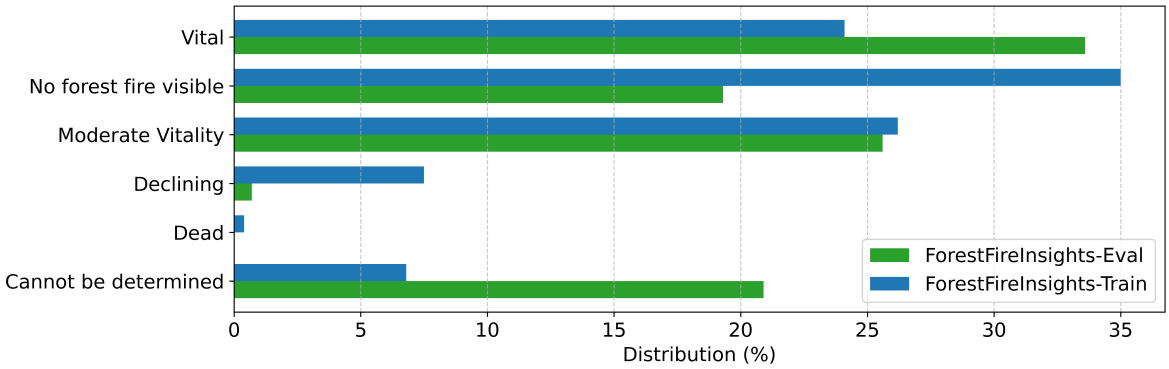


Figure A11. Annotation distribution for "Tree Vitality" in all datasets

References

1. Jones, M.W.; Kelley, D.I.; Burton, C.A.; Di Giuseppe, F.; Barbosa, M.L.F.; Brambleby, E.; Hartley, A.J.; Lombardi, A.; Mataveli, G.; McNorton, J.R.; et al. State of Wildfires 2023–2024. *Earth System Science Data* **2024**, *16*, 3601–3685. https://doi.org/10.5194/essd-16-3601-2024.
2. Kalabokidis, K.; Xanthopoulos, G.; Moore, P.; Caballero, D.; Kallos, G.; Llorens, J.; Roussou, O.; Vasilakos, C. Decision support system for forest fire protection in the Euro-Mediterranean region. *European Journal of Forest Research* **2012**, *131*, 597–608. https://doi.org/10.1007/s10342-011-0534-0.
3. Bouguettaya, A.; Zarzour, H.; Taberkit, A.M.; Kechida, A. A review on early wildfire detection from unmanned aerial vehicles using deep learning-based computer vision algorithms. *Signal Processing* **2022**, *190*, 108309. https://doi.org/10.1016/j.sigpro.2021.108309.
4. Dewangan, A.; Pande, Y.; Braun, H.W.; Vernon, F.; Perez, I.; Altintas, I.; Cottrell, G.W.; Nguyen, M.H. FIgLib & SmokeyNet: Dataset and Deep Learning Model for Real-Time Wildland Fire Smoke Detection. *Remote Sensing* **2022**, *14*, 1007. https://doi.org/10.3390/rs14041007.
5. OpenAI.; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report.
6. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-VL Technical Report.
7. Li, Z.; Wu, X.; Du Hongyang.; Nghiem, H.; Shi, G. Benchmark Evaluations, Applications, and Challenges of Large Vision Language Models: A Survey.
8. Tao, L.; Zhang, H.; Jing, H.; Liu, Y.; Yan, D.; Wei, G.; Xue, X. Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques. *Remote Sensing* **2025**, *17*, 162. https://doi.org/10.3390/rs17010162.
9. Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; Zhu, X.X. Vision-Language Models in Remote Sensing: Current progress and future trends. *IEEE Geoscience and Remote Sensing Magazine* **2024**, *12*, 32–66. https://doi.org/10.1109/MGRS.2024.3383473.
10. Wei, T.; Kulkarni, P. Enhancing the Binary Classification of Wildfire Smoke Through Vision-Language Models. In Proceedings of the 2024 Conference on AI, Science, Engineering, and Technology (AIxSET). IEEE, 2024, pp. 115–118. https://doi.org/10.1109/AIxSET62544.2024.00021.

11. Schneider, D. *Waldbrandfrüherkennung*; W. Kohlhammer GmbH: Stuttgart, 2021. https://doi.org/10.17433/978-3-17-036507-0.
12. Hsieh, R. *Alberta Wildfire Detection Challenge: Operational Demonstration of Six Wildfire Detection Systems*; Vol. Technical Report; TR 2023 n.1, 2023.
13. Alkhatib, A.A.A. A Review on Forest Fire Detection Techniques. *International Journal of Distributed Sensor Networks* **2014**, *10*, 597368. https://doi.org/10.1155/2014/597368.
14. Mohapatra, A.; Trinh, T. Early Wildfire Detection Technologies in Practice—A Review. *Sustainability* **2022**, *14*, 12270. https://doi.org/10.3390/su141912270.
15. Göttlein, A.; Laniewski, R.; Brinkschulte, C.; Schwichtenberg, H. Praxistest eines Waldbrand-Frühwarnsystems. *AFZ der Wald* **2023**.
16. Blais, M.A.; Akhloufi, M.A. Drone Swarm Coordination Using Reinforcement Learning for Efficient Wildfires Fighting. *SN Computer Science* **2024**, *5*. https://doi.org/10.1007/s42979-024-02650-6.
17. Pronto, L.; Held, A. Einführung in das Feuerverhalten, 2021.
18. Saxena, S.; Dubey, R.; Yaghoobian, N. A Model for Predicting Ignition Potential of Complex Fuel in. PhD thesis, Florida State University, Florida, 2023. https://doi.org/10.48550/arXiv.2206.02518.
19. Cimolino, U. Analyse der Einsatzerfahrungen und Entwicklung von Optimierungsmöglichkeiten bei der Bekämpfung von Vegetationsbränden in Deutschland. Dissertation, Universität Wuppertal, Wuppertal, 2014.
20. Patzelt, S.T. Waldbrandprognose und Waldbrandbekämpfung in Deutschland - zukunftsorientierte Strategien und Konzepte unter besonderer Berücksichtigung der Brandbekämpfung aus der Luft. Dissertation, Johannes Gutenberg Universität, Mainz, 2008. https://doi.org/10.25358/openscience-2569.
21. Deutscher Feuerwehrverband. Anzahl der Feuerwehren, 2023.
22. Tielin, M.; Chuanguang, Y.; Wenbiao, G.; Zihan, X.; Qinling, Z.; Xiaoou, Z., Eds. *Proceedings of 2017 IEEE International Conference on Unmanned Systems (ICUS): Oct. 27-29, 2017, Beijing, China*; IEEE: Piscataway, NJ, 2017.
23. Laurençon, H.; Tronchon, L.; Cord, M.; Sanh, V. What matters when building vision-language models?
24. Duan, H.; Fang, X.; Yang, J.; Zhao, X.; Qiao, Y.; Li, M.; Agarwal, A.; Chen, Z.; Chen, L.; Liu, Y.; et al. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models.
25. Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; Ma, Y. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models.
26. Daniel Han, M.H.; Unsloth team. Unsloth, 2023.
27. Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling.
28. Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI.
29. Zhang, Y.F.; Zhang, H.; Tian, H.; Fu, C.; Zhang, S.; Wu, J.; Li, F.; Wang, K.; Wen, Q.; Zhang, Z.; et al. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans?
30. Center for Wildfire Research, University of Split. FESB MLID dataset, 3/13/2025.
31. Krstinić, D.; Stipaničev, D.; Jakovčević, T. *Histogram-based smoke segmentation in forest fire detection system*; 2009.
32. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models.
33. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs.
34. Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C.H.; Gonzalez, J.E.; Zhang, H.; Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention.
35. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution.