Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do Vision Language Models Detect Wildfire Smoke? A Controlled Event-Disjoint Evaluation Against Classical Background Subtraction

Mohamed Jabli

jeblimedamine@gmail.com

University of Sousse

Mehrez Abdellaoui

University of Sousse

Ai Douik

University of Sousse

Research Article

Keywords: Smoke Detection, PTZ cameras, VLM, Background Subtraction, False-alarm analysis

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11597821/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Do Vision Language Models Detect Wildfire Smoke? A Controlled Event-Disjoint Evaluation Against Classical Background Subtraction

Jabli Mohamed Amine[1] , Mehrez Abdellaoui[2] , Ai Douik[3]

[1,2,3]ENISO – National Engineering School of Sousse, University of Sousse, Sousse, Tunisia

Email: jeblimedamine@gmail.com, mehrez.abdellaoui@eniso.u-sousse.tn, Alidouik@gmail.com

Corresponding Author: Jabli Mohamed Amine, ENISO, National Engineering School of Sousse, University of Sousse, Tunisia. Email: jeblimedamine@gmail.com

Abstract

Fixed-view cameras record smoke as it develops, but faint plumes can resemble clouds or other scene changes. This study compares vision–language models (VLMs), a compact convolutional network, and classical background-difference methods on an event-disjoint subset of HPWREN FIgLib. The subset contains 18,073 labeled frames from 233 sequences; the test partition contains 35 events. Against the tested methods, a fire-free camera reference achieved a frame-level precision–recall area (PR-AUC) of 0.979. VLM drift conditions reached 0.784 and 0.779. Few-shot linear probes did not exceed the corresponding zero-shot scores, and MobileNetV3-Small remained near the test-set prevalence baseline. Requiring agreement between the classical and VLM detector families reduced pre-plume false alarms from 60 to 22. Smoke-frame recall also fell, from 0.878 to 0.561. In a model-assisted audit, Qwen3-VL-8B assigned faint-smoke labels to two of four classical–MobileCLIP agreement episodes, two of 14 classical-only episodes, and two of 30 controls. This work was conducted in an exploratory capacity where 52 % of the frames are analyzed and classified uncertain, there is no human processes during this study, the results validate that camera split specific background data as a robust baseline.

Index Terms: Smoke Detection, PTZ cameras, VLM, Background Subtraction, False-alarm analysis.

1 Introduction

Fixed Cameras are used for monitoring extensive areas, these cameras are capable of capturing a smoke or plumes prior to the visibility of flames, clouds and haze as an atmospheric conditions may generate visual patterns analogous to smoke, these variations in nocturnal lighting contribute to ambiguous signatures . For consequence the risk of false alarms represents a significant challenge. [1,3]. The Figlib and SmokeyNet datasets offer a fixed PTZ camera view wildfire along a spatiotemporal detection. [4] (Figure 1)

FireCLIP evaluates multimodal prompt tuning with few-shot and regional-transfer protocols Wu et al. [5]. SmokeBench evaluates large multimodal models on smoke classification, localization, and detection Qi et al. [6]. EARL models temporal smoke onset from fixed-camera video and uses fire-free camera history Muksimova and Umirzakova [7]. This paper addresses a related but narrower question: how selected semantic and classical baselines compare when evaluated on the same event disjoint FIgLib test partition.

We compare zero-shot and few-shot VLM conditions with a compact supervised CNN and reference-based background subtraction. The models use the same frozen event partition.

We also test whether requiring agreement between detector families. This changes the balance between false alarms if pre-plume and on smoke-frame recall. Then alarms are reviewed by our multimodal modal (VLM). Furthermore, these frames are treated as model outputs.

We compare the camera's fire free reference with zero-shot and with few-shot VLMs, so the last block is finetuned with the model MobileNetV3-Small, then we apply a frame differencing.

Traditional video fire detection has a long and known history [1, 8], the main contribution in this work is matching evaluation standards to explore the alarm events , the test database includes 35 events from one camera network taken from FigLib database , which is related to Southern Californi , and the labels are following the original source time relative to smoke detection , there is no human annotation , rather then it we use Qwen3-VL-8B VLM as AI assisted audit. So, the finding describes the performance specific to AI assisted model.

2 Contributions

This paper introduces three contributions. First, it reports a matched evaluation of selected zero-shot and few-shot VLM conditions, a compact CNN, and classical background methods on one frozen, event-disjoint FIgLib split. The comparison shows how these tested systems rank on the same 35 event test partition; it does not establish a general ranking across wildfire datasets.

Second, the paper quantifies the trade-off produced by an agreement rule. The preplume false alarm count decreases from 60 to 22, while smoke-frame recall declines from 0.878 to 0.561. An audit of 67 alarm and control episodes records Qwen3-VL-8B's cause assignments. The audit is a small, model-assisted analysis. It does not provide human-verified alarm causes.

Third, the study reports a camera-history ablation and measured compute for the evaluated conditions. The raw cold-start comparison tests between one and 40 fire-free reference frames, while separate event-level results distinguish alarms from timely detections.

3 Related Work

3.1 Fixed-view wildfire datasets and event transfer

Dewangan et al. introduced FIgLib, a collection of fixed-view images with plume-relative timing, and presented SmokeyNet for real-time smoke detection Dewangan et al. [4]. Also, it uses a subset of the same source. The unit of generalization is the fire camera event; frames from one event share scene, camera, weather, and illumination conditions. The random division of frames can result a highly similar images in both training dataset and test dataset which can inflating estimates of the model transferability. In 2026 a new research EARL paper learns how scene behave using videos without fires and check the false alarms detected and timing [7]. Because of the use of camera history and measure time to detection, it is a very close temporal reference .But our experiment takes a different path by comparing selected conditions on one specific set of events then we examine a sample of the causes of these detected alarms . For consequence this results do not be direct compared with EARL architecture study because we focus only on alarms detections

3.2 Classical video smoke detection

The classical video systems are based on motion detection, they also use color and texture or even temporal flicker, backgrounds models are essential too. A recent review organizes video fire and smoke methods by scene and application conditions [2]. Töreyin et al. combined visual cues with wavelet analysis [1], their work is based on real time fire and flame detection, another contribution combines visual sensor system motion and the appearance features with a CNN. Then they reported that the system reduced false positives relative to a fully DL baseline on its dataset [3]. These approaches use motion or appearance cues that vary in their sensitivity to camera movement and scene change. They provide established comparators for a fixed-camera setting, where a fire-free reference can be stored for each view.

The resulting score is a fixed-reference pixel difference measure. Strong results suggest that using stable camera history can hold significant information's within a specific database, however when moving camera it demand separate evaluation, so changes in scene require another testing , also the geographic settings must be tested also separately .

3.3 Vision–language models and low-label adaptation

The CLIP is acting like a bridge, it is an AI system, it learns representations from and image caption, it supports zero shot classification using text prompts [9]. The MoblieCLIP system focus on faster inference because it uses a compact design and a reinforced training [10]. This study uses the two families; a compact VLM and also a large CLIP encoder. Another model can serve as a baseline which is the MobileNetV3-Small, this is a supervised convolutional model [11]. The comparison spans three types; first, the pretrained text to image scorer and second a small text to image model, then a supervised visual classifier.

The prompt learning technique and the few shot adaptation are the area of interest; the CoOp learns based on continuous prompts [12] improvement is done on numerous downstream tasks. The FigLib applies also multimodal prompt finetuned to forest fire data [5] and it evaluate a few shot budgets and a regional transfer capability. All these models are at the base of evaluating prompts, so they update the representation directly, in contrast our few shot conditions fit a balanced regression. For consequence all these protocols test different approaches based on adaptation strategies, the performance of these works depends on the relationship between the training and the testing sets.

In this study MLLMs prompts using SmokeBench is used [6], it includes 5046 positive smoke and 1000 negative, it covers classification, localization and detection. This approach shares the sources collection for the image benchmark, but they addresses different evaluation metrics, they review short video sequence using a verifier then they use a detector for alarms.

4 Materials and Methods

4.1 Dataset and label construction

In this study, we use the HPWREN FigLib database as our source, our experiment samples 223 sequences of videos from a larger collection of 525 sequences, the collected database confirm a 10.03 GB of data of images, we sampled images at approximately one-minute intervals, each sequence is identified by its specific fire and camera identifiers. The full duration of a sequence data is about 40 min before and after plume, we select our subset using seed 42 for the reason to ensure no incomplete sequence dropped. Image dimensions in the source range from 1600 × 1200 to 3072 × 2048 pixels. The data identified 75 exact duplicate frames, so we retrain all them within their original sequences as a part of the camera record. We

have 233 events, we use 232 containing both labels classes, each event contributes between 40 et 158 frames, so a mean of 77.6 frames for each event. The frame level metrics in this research is weight events based on the labelled frames, we provide a sperate summary at the event level, also we adhered to the filename of plume relative offset convention for our labeling rule. The frames at offset of 10 seconds are labeled by us. In total we exclude 254 ambiguous frames from 18073 labelled frames ;18000 frames are 9057 for smoke and 9016 for no-smoke. (Figure 2)

The prevalence is the processed database is 0.501 in the full dataset and 0.506 in testing dataset, The offset label does not provide a verified ignition time or a human level judgement, we split the database in training, validation and testing sets: 163/35/35 splits. In this partition no fire camera sequence is used. The public GitHub repository of this work includes the splits.csv file and his summary. also, it includes all necessary files for data preparation. (Table 1)

Table 1: Summary of the frozen experimental protocol.

Component	Protocol
Data	We selected 233 out of 525 available FIgLib sequences using random seed 42 and a strict 10-GiB data limit; from this subset, we curated 18,073 labeled frames.
Labels	Smoke classified at offsets ≥ +10 s; no-smoke classified at offsets ≤ [threshold] s; 254 ambiguous frames excluded.
Split	163/35/35 split of fire camera events assigned to training, validation, and test sets; sequence membership fixed prior to model evaluation.
Models	CLIP ViT-B/32 (224-pixel input), MobileCLIP-S0 (256-pixel input), and ImageNet-initialized MobileNetV3-Small with last-block fine-tuning.
Low-label settings	Three smoke and three no-smoke prompts are evaluated. Alongside logistic regression probes using 1, 5, 10, and 25 examples per class, repeated across five fixed seeds.
Classical baselines	Fire-free reference at 256x192, exponential moving background.
Temporal alarm	95 % threshold; an alarm requires k consecutive positive frames.
Cause audit	67 episodes (37 alarms and 30 controls), using Qwen3-VL-8B at temperature zero
Primary reporting	Metrics include Frame PR-AUC and ROC-AUC, event detection also counts and time of detection, along with pre-plume false alarm rates and smoke-frame recall.



Figure 1: Examples of subtle smoke and pre-plume confusers in the study data. Cloud patterns, low-contrast haze, and dark scenes complicate single-frame recognition. The supplied figure labels the examples using plume-relative offsets.

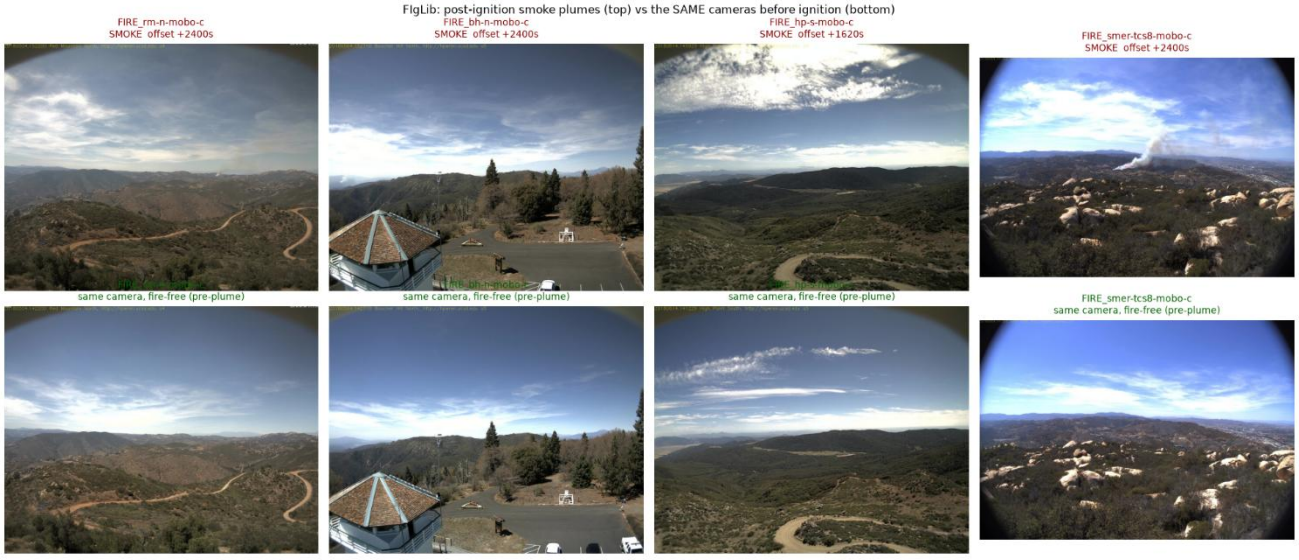


Figure 2: FIgLib subset. Upper panels show positive smoke frames after visible plume appearance. The lower panels show the same camera views before plume appearance..

4.2 Models and scoring conditions

4.2.1 Zero-shot and few-shot VLMs

We evaluated CLIP ViT-B/32 and MobileCLIP-S0. MobileCLIP-S0 used its 256-pixel image preprocessing and has approximately 53.8 million parameters across image and text towers; CLIP ViT-B/32 has approximately 151 million parameters and used 224-pixel inputs. MobileNetV3-Small has approximately 2.5 million parameters and was initialized with ImageNet-1k weights before last-block fine-tuning. For zero-shot classification, each VLM produces image and text embeddings. A set of three smoke prompts and three no-smoke prompts was fixed in advance. The image–text similarities were combined into a scalar score. The preregistered zero threshold was used for the primary thresholded classification results; validation-tuned thresholds were also recorded.

The valuation used the ECE calibration error for zero shot conditions, for the few shot adaptation; image embedding stay frozen while regression is trained on 1 and 5 and 10 or 25 examples off frames per class. So, the result is reported the mean and the standard deviation across the seeds. All these comparison tests a small supervised CNN instead of a complex architecture or a hyperparameters search.

4.2.2 Classical reference and frame-difference conditions

The classic baseline denoted "bg_pre" compare the test image against the fire free camera, the images are converted grayscale, then decoded and reduced in resolution, then images are compared using pixelwise background difference. The score is normalized using camera history and the temporal transformations. The method uses one reference for the cold start analysis and up to 40 fire-free frames for the full-history condition. The frame-difference baseline compares frames without a fixed reference. An exponential moving average technique background baseline is also tested. These methods use temporal information in a different way from the VLMs. The background subtraction measures change relative to a camera scene. VLM scores measure semantic using prompts. Their score scales are therefore not directly comparable. We compare each method by test-set ranking metrics and by a detection protocol with thresholds selected on validation data.

4.2.3 Camera-drift scores and alarm thresholds

For a model score s_t at frame t, equation of the camera drift transformation is detailed below

$$z_t = \frac{s_t - \mu_{\mathrm{ref}}}{\sigma_{\mathrm{ref}} + \epsilon},$$

where μ_{ref} and σ_{ref} are score mean and standard deviation from fire-free camera history and ϵ prevents division by zero. Optional temporal smoothing was selected from a preregistered configuration set on validation data. The selected configuration was then fixed for test evaluation.

Alarm threshold τ is 0.8423 of validation fire-free scores, corresponding to 5 % frame false-positive target on the validation data. The alarm event required a run of k consecutive frames above the threshold. Temporal window and persistence settings were chosen on validation data under the documented selection rule.

4.3 Evaluation metrics

The evaluation metrics used the PR-AUC, ROC-AUC, F1, precision, and recall. The first one depends on the class prevalence, so no skill baseline equal the smoke prevalence (0.506). also, the PR-AUC for the test is set to 0.506. The event level reports the list of detected events, the median alarm relative to the plume detection, the test set has 2686 frames where 1328 no smoke and 1358 are smoke. The complementary analysis provides a pre-plume false alarm and the smoke frame recall at the selected points. The pre-plume alarm is measured relative to the appearance of the plume. The negative alarms indicate the pre-plume alarms.

The experiment notes state that event-level bootstrap intervals were calculated with 2,000 resamples, with events as the sampling units; the supplied summary omits interval values for some results. Point estimates are descriptive, and no statistical significance is inferred. Frames are not treated as independent event replications.

4.4 Model-assisted alarm-cause audit

The study explores 67 episodes: 37 with detector alarms and 30 without it. The alarm was grouped by the detector membership, the samples included 14 classical only episodes ;4 episodes used combined classical and MobileCLIP-S0 detectors. Then three episodes used all three detectors, 4 episodes on the S0 detector. Then 7 episodes only on CLIP. Finally, two episodes combining the S0 and the CLIP detectors

Controls were selected with fixed seed 7. Qwen3-VL-8B reviewed episode filmstrips with a preregistered prompt at temperature zero. It assigned labels to 174 frames. The model ran locally through an OpenAI-compatible endpoint; images were resized to 640×480.

The cause codebook includes faint smoke, cloud, night-light artifact, illumination change, other artifact, and uncertain. The episode summary reports five categories; the frame summary includes all six. These are Qwen3-VL-8B assignments. No human reviewer adjudicated the filmstrips, so the audit can identify cases for review but cannot estimate true cause rates.

5 Results

5.1 Frame-level comparison

Reference-based background subtraction achieved a test PR-AUC of 0.979, compared with 0.784 for MobileCLIP-S0 with camera-drift scoring and 0.779 for CLIP ViT-B/32 with camera-drift scoring. Zero-shot CLIP reached 0.656 and zero-shot MobileCLIP-S0 reached 0.581. The few-shot probes reached 0.576 for CLIP at 25 examples per class and 0.537 for S0 at 25 examples per class. MobileNetV3-Small remained close to the prevalence baseline, with a best reported PR-AUC of 0.511 at 10 examples per class. Frame differencing without a reference reached 0.505. (Figure 3)

At the reported budgets, the few-shot probes remained below their corresponding zero-shot results. This finding is specific to the tested linear-probe protocol and event-disjoint test events. The zero-shot thresholded output also illustrates why ranking and classification metrics should be separated: at the preregistered zero threshold, both VLMs predicted smoke for every test frame, yielding recall 1.0 but precision equal to prevalence. (Figure 4; Figure 5)

At validation-tuned thresholds, CLIP achieved full smoke-frame recall in 21 of 27 smoke events, and MobileCLIP-S0 did so in 23 of 27. For the eight no-smoke events, neither model had an event with zero false-positive frames. This event-level screen applies a stricter criterion than frame-level recall: one missed smoke frame prevents a full-recall designation, while one false-positive frame prevents a zero-FPR

designation. These counts show that frame ranking and event-level behavior provide different views of the same test set.

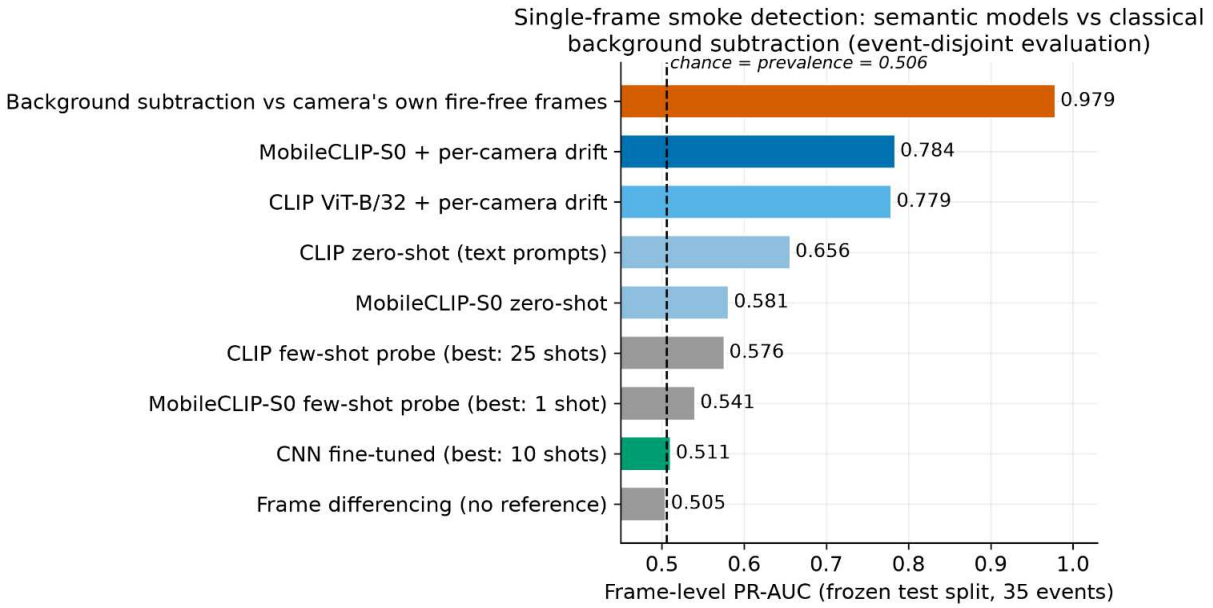


Figure 3: Frame-level PR-AUC on the frozen event-disjoint test split. The dashed reference is test prevalence (0.506). Camera-specific background subtraction has the highest reported PR-AUC among the evaluated conditions.

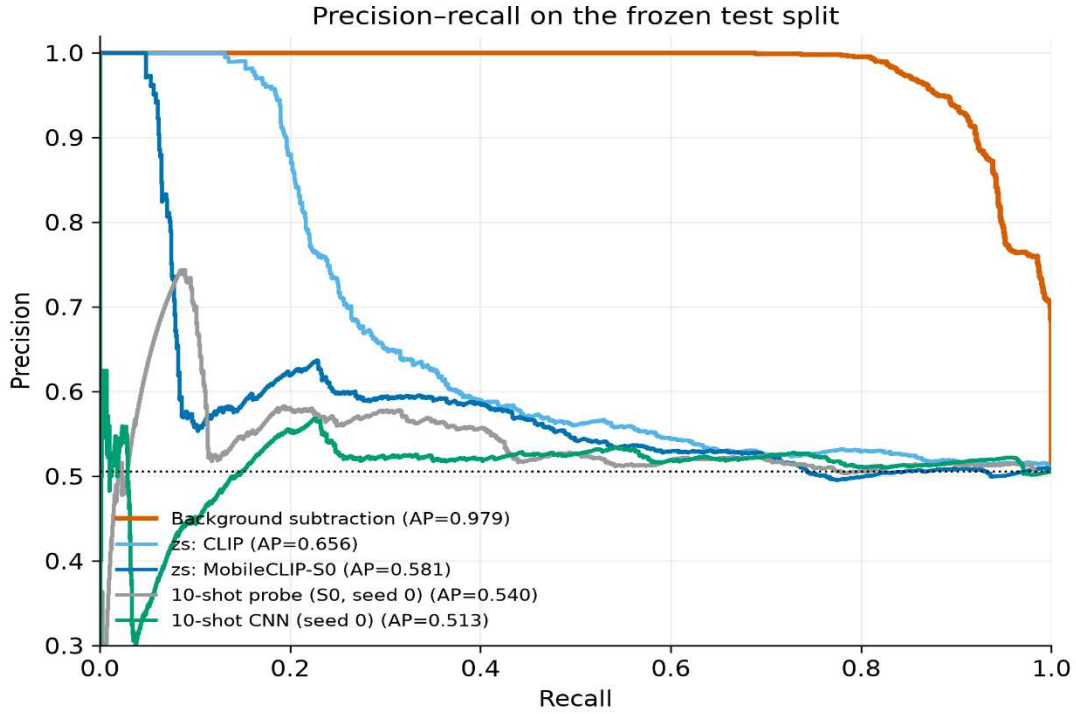


Figure 4: Precision–recall curves on the frozen test split for selected conditions. The horizontal dotted line marks the test prevalence. The single-seed probe and CNN curves shown here are illustrative of their respective 10-shot configurations; aggregate few-shot results are reported in Table 2.

Table 2: Frame-level results on the test split. PR-AUC and ROC-AUC are reported as supplied. Few-shot and CNN rows summarize five fixed seeds where available. Dashes indicate absent or inapplicable metrics.

Condition	F1	PR-AUC	ROC-AUC	ECE	Brier	Notes
CLIP zero-shot, preregistered threshold	0.672	0.656	0.591	–	–	threshold 0.0
MobileCLIP-S0 zero-shot, preregistered threshold	0.672	0.581	0.548	–	–	threshold 0.0
CLIP probe, 1 shot/class	0.299 ± 0.171	0.543 ± 0.041	0.527	0.014	0.247	5 seeds
CLIP probe, 5 shots/class	0.532 ± 0.028	0.547 ± 0.031	0.534	0.051	0.236	5 seeds
CLIP probe, 10 shots/class	0.540 ± 0.037	0.548 ± 0.037	0.530	0.069	0.223	5 seeds
CLIP probe, 25 shots/class	0.500 ± 0.013	0.576 ± 0.032	0.544	0.114	0.191	5 seeds
S0 probe, 1 shot/class	0.303 ± 0.132	0.541 ± 0.015	0.522	0.026	0.243	5 seeds
S0 probe, 5 shots/class	0.529 ± 0.067	0.528 ± 0.018	0.517	0.074	0.222	5 seeds
S0 probe, 10 shots/class	0.507 ± 0.033	0.514 ± 0.025	0.508	0.125	0.201	5 seeds
S0 probe, 25 shots/class	0.482 ± 0.060	0.537 ± 0.025	0.517	0.228	0.151	5 seeds
MobileNetV3-Small, 1 shot/class	0.335	0.499	0.504	0.367	0.260	last-block fine-tuning
MobileNetV3-Small, 5 shots/class	0.376	0.508	0.516	0.371	0.221	last-block fine-tuning
MobileNetV3-Small, 10 shots/class	0.524	0.511	0.515	0.309	0.131	last-block fine-tuning
MobileNetV3-Small, 25 shots/class	0.514	0.493	0.501	0.195	0.065	last-block fine-tuning

5.2 Temporal detection and camera history

The selected `bg_pre` configuration reached PR-AUC 0.979, detected 34 of 35 test events, and had a median alarm offset of +60 seconds relative to visible plume appearance. Its premature-alarm rate was 0.400. The VLM-drift conditions each detected 25 of 35 events, with PR-AUC values of 0.784 and 0.779. MobileCLIP-S0 had a median alarm offset of -180 seconds; CLIP had a median offset of +240 seconds. Negative offsets mark pre-plume alarms; the offset labels alone do not identify their visual cause. Classical EMA background modeling and frame differencing had lower PR-AUC values of 0.598 and 0.505. (Table 3)

The camera-history ablation compared up to 40 fire-free reference frames. For the raw reference condition, classical PR-AUC was 0.701 with one reference frame and 0.729 at 40 frames; the intervening points varied modestly. The full temporal transformation reached 0.979. This result comes from a separate configuration and must be read apart from the raw-history curve. The VLM drift scores ranged from 0.688 to 0.732 for S0 and 0.714 to 0.725 for CLIP across the plotted reference counts. On these sequences, a semantic prior showed no clear cold-start advantage over a single fire-free frame. This conclusion applies to sequences that begin with fire-free camera history. (Figure 6)

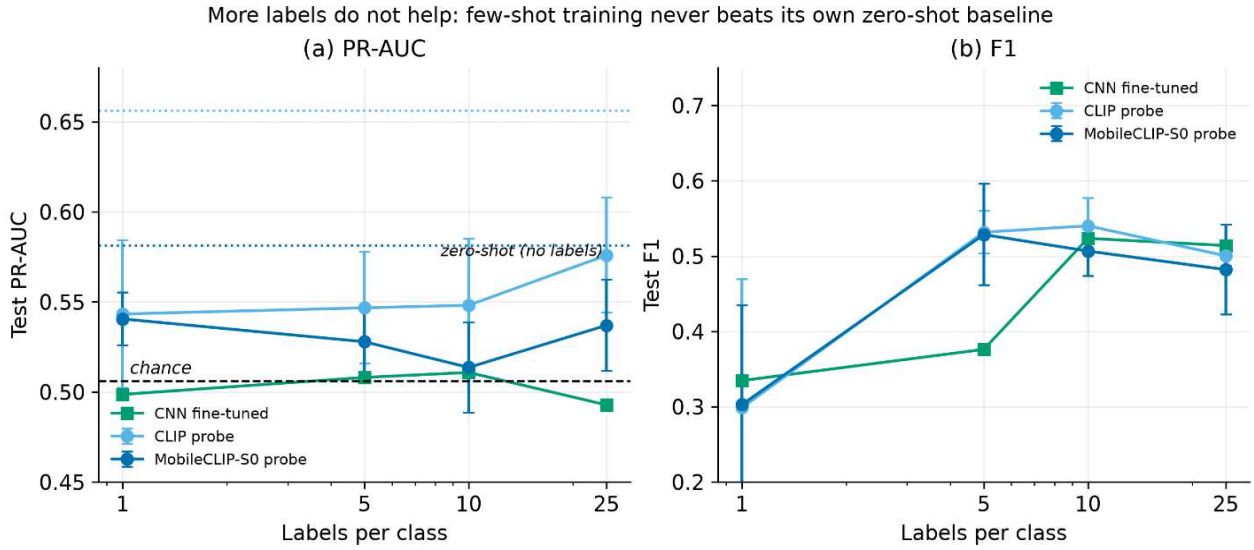


Figure 5: Few-shot performance by label budget. The dotted lines in panel (a) show the corresponding zero-shot PR-AUC values. The points and error bars summarize the supplied five-seed runs.

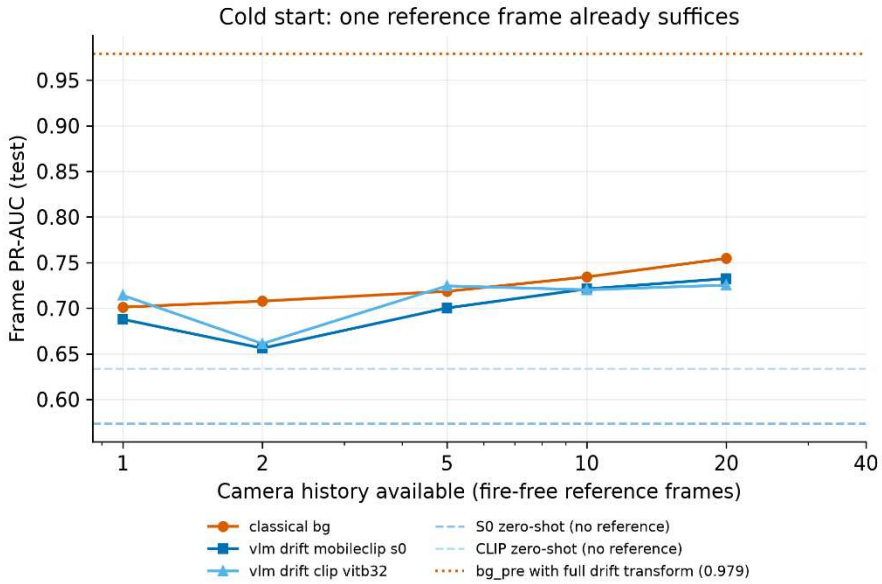


Figure 6: PR-AUC as the number of available fire-free reference frames increases. The solid curves show raw camera-history ablations. The dotted reference denotes the separately reported, fully transformed classical score. These values represent different configurations.

Table 3: Temporal detection results on 35 test events. A negative time-to-detection (TTD) marks an alarm before the visible-plume reference. Independent onset labels are required to verify early smoke detection.

Method	PR-AUC	Median TTD (s)	Detected	Premature rate
Classical `bg_pre`	0.979	+60	34/35	0.400
MobileCLIP-S0 drift	0.784	−180	25/35	0.400
CLIP drift	0.779	+240	25/35	0.314
Classical EMA	0.598	−481	7/35	0.114
Frame differencing	0.505	+600	5/35	0.057

5.3 Alarm-set complementarity

At the preregistered operating points, the classical detector produced 60 pre-plume false alarms among 1,328 pre-plume frames, for a frame-level false-positive rate of 0.045. The union of VLM drift alarms produced 106 false alarms, or 0.080. Requiring both detector families to alarm reduced the count to 22, or 0.017. This is a reduction of 63% relative to the classical count. The overlap contained 22 alarms; relative to the 60 classical alarms, 36.7% were also identified by the VLM family. The supplied notes also report 144 unique alarms across the family-level union.

The reduction in false alarms came with lower smoke-frame recall. Classical detection alone reached 0.878 recall, VLM drift alone 0.613, and the AND rule 0.561. The AND rule prioritizes fewer alarms at the cost of recall. A recall-oriented system requires additional review or escalation logic before applying this rule. (Figure 7)

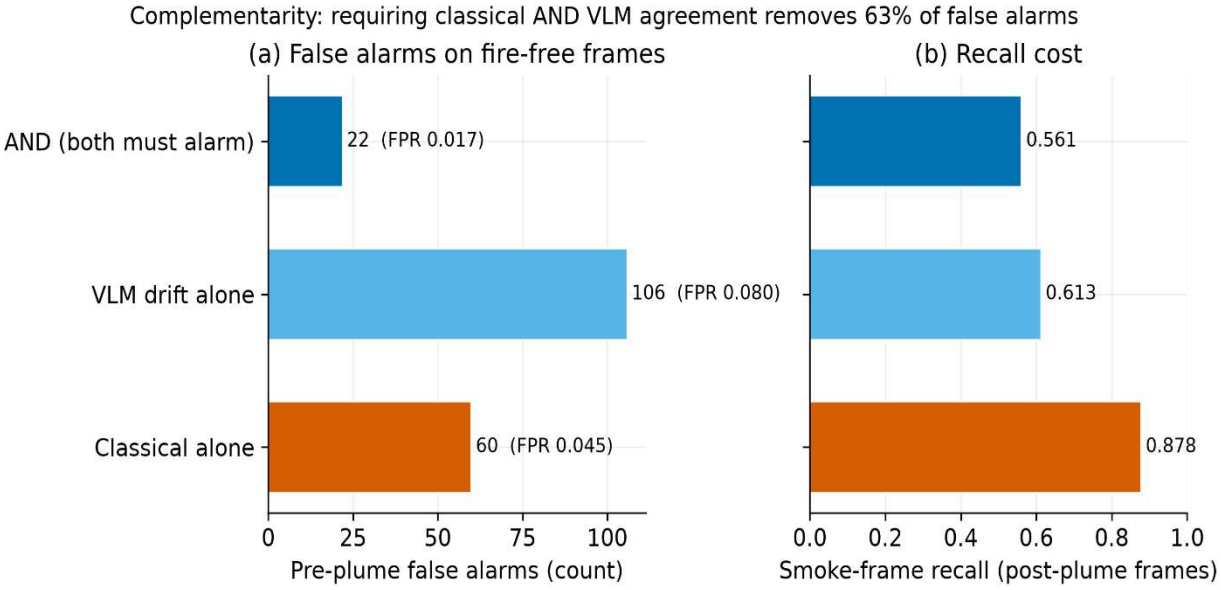


Figure 7: False-alarm and recall trade-off for the classical detector, VLM drift family, and agreement rule. The AND rule reduces pre-plume false alarms, but it also reduces smoke-frame recall.

5.4 MLLM-assigned causes of alarm episodes

The alarm audit suggests that different detector families respond to different scene changes. Among 14 classical-only episodes, the verifier labeled seven as clouds, five as uncertain, and two as faint smoke. For the four classical–S0 episodes, two were labeled faint smoke, one night-light artifact, and one uncertain. In the 30 controls, two were labeled faint smoke, 10 clouds, two other artifacts, and 16 uncertain. (Figure 8; Table 4)

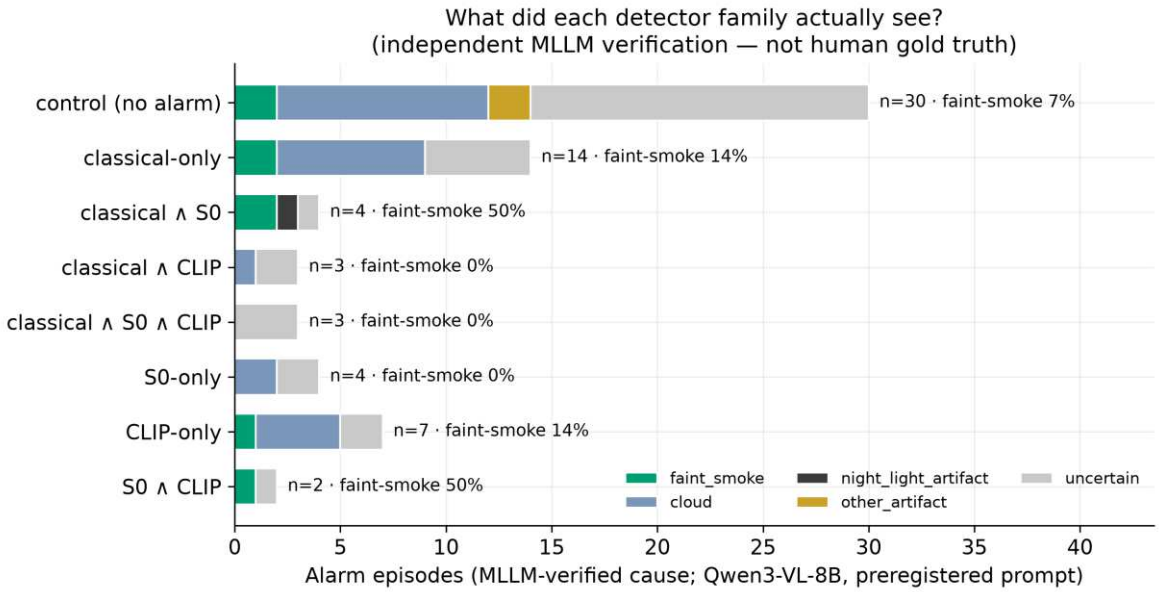


Figure 8: Qwen3-VL-8B cause labels for the 67 audited episodes. Group sizes vary, and 52% of reviewed frames were uncertain.

Table 4: Episode-level verifier labels by detector membership. Each entry is a count of episodes. The model's six frame-level codebook categories are aggregated into the five categories shown in the supplied episode summary.

Verifier label	Classical only (14)	Classical+S0 (4)	Classical+CLIP (3)	All three (3)	S0 only (4)	CLIP only (7)	S0+CLIP (2)	Controls (30)
Faint smoke	2	2	0	0	0	1	1	2
Cloud	7	0	1	0	2	4	0	10
Night-light artifact	0	1	0	0	0	0	0	0
Other artifact	0	0	0	0	0	0	0	2
Uncertain	5	1	2	3	2	2	1	16

5.5 Failure cases and compute

The *bg_pre*-method detects 13 out of 35 test events within 120 seconds, this is without a false early alarm, it flag 11 as premature, 10 as late and 1 misses it. only the 13 timely detections satisfy the criteria, the missed event is also missed by all the other detectors. The hight score of pre-plume frames shown in red

lens artifact at night as show in the next figure .These images support the interpretation of the verifier table but do not validate all alarm causes. (Figure 9)

On the development machine, classical baselines processed all 18,073 frames in about 240 seconds on CPU, at approximately 75 images per second. MobileCLIP-S0 feature extraction required 1,623 seconds, with 427 MiB allocated peak GPU memory and 11.1 images per second. CLIP feature extraction required 1,739 seconds, with 658 MiB peak memory and 10.4 images per second. MobileNetV3-Small training took 3.7–13.4 seconds per run under cached evaluation conditions.

The local VLM model called Qwen3-VL-8B verify the processed of 174 frames in 533 seconds , in average of 3 seconds per frame , these measurement is obtained with RTX 5070 GPU 8 GB VRAM , the environment use PyTorch 2.14.1 and CUDA 13.0 , these results reflect the performance on a single system. (Table 5)

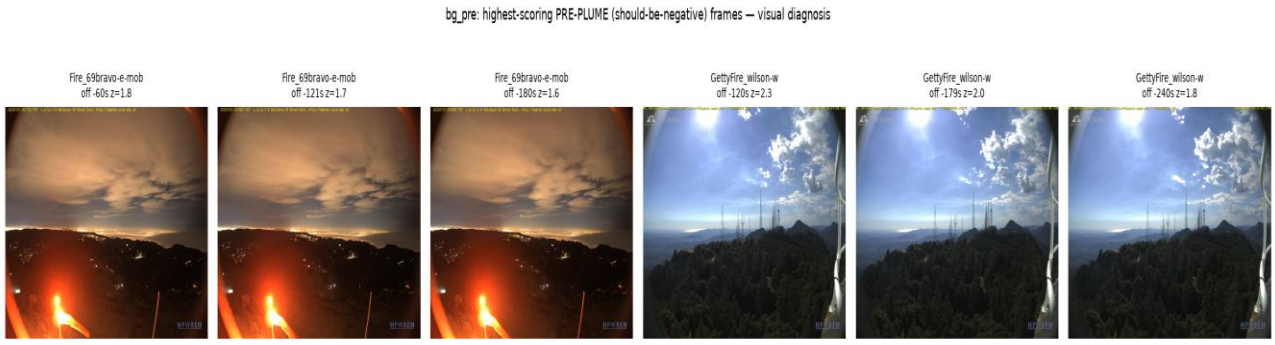


Figure 9: Examples of high-scoring pre-plume frames for the reference-based classical detector. The left sequence contains an orange lamp/lens artifact; the right sequence contains clouds. The figure provides qualitative context for the false-alarm audit.

Table 5: Measured compute on the development system. Values are environment-specific and do not estimate embedded deployment performance.

Component	Wall time	Peak GPU memory
MobileCLIP-S0 features	1,623 s	427 MiB
CLIP ViT-B/32 features	1,739 s	658 MiB
MobileNetV3-Small training	3.7–13.4 s/run	118 MiB
Qwen3-VL-8B verification	533 s	local host
Classical baselines	about 240 s	CPU only

6 Discussion

6.1 Camera history as a strong reference

The best results based on system performance is observed between the camera specific background and the semantic image scoring, the classical method achieve a PR-AUC of 0.979 while the VLM drift has a score between 0.779 and 07.84, the outcome align with the fixed view camera in a nature view sequence. The camera has its own fire-free history, so it provides a direct reference for scene

changes, the VLM must inter where visible texture represents smoke, cloud, haze or unrelated feature. These comparisons have limitations; the images are captured from a stationary camera with fire-free frames proceeding the appearance of plume, so the moving camera can reduce the value of the stored references, even the action of recording after ignition. The raw cold start analysis yielded a PR-AUC of 0.701 with one reference frame. These results do not cover all conditions.

While the classical system registered alarms in 34 of 35 test events, adherence to study conventions revealed that 11 instances were premature and 10 were late, thereby confirming that merely 13 events met the requisite timely detection criteria. Consequently, deployment protocols necessitate an explicit event-level timing objective complemented by a frame-ranking metric to ensure robust operational efficacy.

6.2 Low-label semantic adaptation

The tested liner probes is below the zero shot baselines, this outcome is because of protocol used and the data split , the CoOp and FireCLIP methods employs the prompts for learning strategies , to distinct between the embedding logistic probes [5,12] , the SomkeBench evaluate the large language model for classification and for localization .Our VLM method scores from CLIP style Image Text similarity , while Qwen Model act as a verifier [6].These model families and tasks must be strictly distinguished. For example, the MobileNetV3 Small model baseline hovered near chance levels. It has less performance that may stem from the small sample budget, the cross-event variation, also the fine-tuning scope, or of the optimization choices. The current experiment leaves these causal explanations unresolved.

6.3 Agreement rules and alarm triage

The application of the AND rule successfully reduce the false alarms, these false alarms are related to the plume detection, the efficacity of this rule is contingent based on the relative cost assigned for the detection of false alarms compared with the missed detections. In this condition where detection is missed that carry a Hight penalty, the system or any system must have the ability to access to other independent detectors outputs. An architecture should get all the outputs escalate the conflicting results and send them to manual review, The MLLM method used in this paper provide a preliminary support for this approach where many classical detectors classified frames as clouds or uncertain. EARL [7] is a relevant study that can be a concurrent of the proposed method, the current experiment evaluates selected references, semantic and supervised baseline, however the EARL architecture remains outside comparison, the future work should use shared events, identical definition, human verified causes and also operational time metrics.

7 Limitations

Our study focuses on a specific set of 35 test events. These events are drawn from a single geographic region using a fixed camera source. We ensure that each event was tested independent

within this dataset, also we include one camera sequence with challenging night-light conditions. To see how these methods works in real time situations like rain, mountain, or using old camera, or different times in one day, we should test them also in a huge amount of data in different places. In this study we use the FigLIB database.

We use 233 sequences from 525 limited by 10 Giga Byte of data, it represents video stream, but it based is not representative of all times. Only 35 events used to catch rare failures. The data are labeled based on when the smoke appear, it skips the first 10 seconds after the detection to avoid confusion. The rest of the labels are simple; they are “smoke there” or “no smoke”, so detecting the start of the fire need to has its owl labeled data. At this stage we can’t say for sure whether an alarm going off right after smoke shows up is actually useful for early warning.

As images appear at a rate of approximately one every minute, the offset cutoffs are working with quite a sparse timeline. To get our almost balanced test set, we excluded 254 frames of the ambiguous border. It is worth keeping in mind that the chosen test prevalence of 0.506 is a characteristic of this specific subset of labeled data and does not represent an estimation of actual frequency of smoke occurrence in a live camera feed. For this reason, it is impossible to convert the false positive counts per image into "false alarms per camera-day" without additional information about typical observation timeframes and our definition of event.

The alarm cause was based on a single MLLM verifier. Given the low number of episodes, 91 out of 174 frames became ambiguous. Thus, all the conclusions about because rates are descriptive. At the moment, before drawing firm conclusions on the factors causing false alarms, it would be useful to conduct a blinded human evaluation.

As for the few-shot standard deviations provided, they summarize five repetitions of sampling training examples from a single frozen data split and do not represent variability of the results obtained over different partitions of events. In spite of the fact that we have performed bootstrap confidence intervals for individual events as part of our workflow, the present summary does not include interval estimations for every condition. Thus, point differences described correspond exclusively to this specific test partition.

8 Conclusion

In this paper, we provide the first thorough pilot evaluation of Vision-Language Models (VLMs) for smoke detection in fixed-view setting. Our main contribution consists in showing that camera-specific background subtraction using frozen-embedding logistic probes serves as a strong baseline surpassing all other examined VLMs, few-shot probes, compact CNNs, and even frame-difference baselines for this particular subset. Additionally, our study proposes a new agreement rule where both classical detectors and VLMs need to concur before raising an alarm; this approach managed to decrease the number of pre-

plume false alarms from 60 to 22, thus proving the effectiveness of hybrid methods even at the expense of certain recall decrease (from 0.878 to 0.561).

This paper offers a tentative framework for alarm triaging through employing MLLM-based verifiers to process only those episodes that are flagged by classical models as being unclear or contradictory. It is important to note that our study does not claim to evaluate the effectiveness of MLLMs as a tool for fire safety system because it is limited to an exploratory proof-of-concept based on a small subset of event-disjoint events (35 events) obtained from one geographical location and camera type. At the same time, there is no evidence that VLMs can be used as an operational early-warning solution across various regions. To check their potential effectiveness in the fire safety domain, further investigations are needed.

Data and Code Availability

The analysis code, experiment configuration, frozen split manifest, and summary results are available at https://github.com/Skytech-Ai-Technologie/smokevlm-pilot. The repository README describes the setup and commands for reproducing the analysis. The FIgLib image subset and model weights are not included; users must obtain the data under HPWREN's current access and attribution terms. Generated feature caches and raw per-frame prediction files are also not committed. The repository provides the split manifest and its checksum for reference.

Funding and Conflict of Interest

The authors should provide funding and conflict-of-interest statements before submission.

References

[1] B. U. Töreyin, Y. Dedeoğlu, U. Güdükbay, and A. E. Çetin, "Computer vision based method for real-time fire and flame detection," *Pattern Recognition Letters*, Vol. 27, no. 1, pp. 49–58.

[2] D. Gragnaniello, A. Greco, C. Sansone, and B. Vento, "Fire and smoke detection from videos: A literature review under a novel taxonomy," *Expert Systems with Applications*, Vol. 255, p. 124783.

[3] V. Carletti, A. Greco, A. Saggese, and B. Vento, "A smart visual sensor for smoke detection based on deep neural networks," *Sensors*, Vol. 24, no. 14, p. 4519.

[4] A. Dewangan, Y. Pande, H.-W. Braun, F. Vernon, I. Perez, I. Altintas, et al., "FIgLib & SmokeyNet: Dataset and deep learning model for real-time wildland fire smoke detection," *Remote Sensing*, Vol. 14, no. 4, p. 1007.

[5] S. Wu, Y. Qiao, S. He, J. Zhou, Z. Wang, X. Li, et al., "FireCLIP: Enhancing forest fire detection with multimodal prompt tuning and vision-language understanding," *Fire*, Vol. 8, no. 6, p. 237.

[6] T. Qi, W. Li, and N. Barnes, "SmokeBench: Evaluating multimodal large language models for wildfire smoke detection," In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2026.

[7] S. Muksimova and S. Umirzakova, "Detecting wildfire smoke onset from fixed view camera networks," *Sensors*, Vol. 26, no. 19, p. 6020.

[8] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," In *Proceedings of the 17th international conference on pattern recognition*, Cambridge, UK, 2004.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," In *Proceedings of the 38th international conference on machine learning*, PMLR, 2021. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html.

[10] P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel, "MobileCLIP: Fast image-text models through multi-modal reinforced training," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[11] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, et al., "Searching for MobileNetV3," In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[12] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, Vol. 130, pp. 2337–2348.