INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 1, PP. 1-8
Manuscript received December 02, 2025; revised January, 2026. doi: 10.24425/ijet.2026.157894

Mitigating the impact of thermal reflections in object detection using Vision-Language Models

Radosław Feiglewicz, and Andrzej Kos

***Abstract*—Thermal imaging is increasingly employed for navigation in challenging conditions such as dense smoke or fog. However, the limited availability of thermal images compared to RGB data makes training deep learning models, such as Convolutional Neural Networks (CNNs), significantly more difficult and often yields unsatisfactory results. Vision-Language Models (VLMs), due to their ability to perform tasks without extensive retraining or with only a small number of training samples, hold the potential to overcome current limitations in thermal imaging applications. This paper introduces a method leveraging VLMs to reduce the impact of reflections in thermal images on object detection accuracy, with a particular focus on human detection. The proposed approach improves the F1-score from 0.83 to 0.97 on a dedicated evaluation dataset, outperforming a baseline solution based solely on the widely used YOLOv11 model. Furthermore, we investigate the effects of quantization on various open-source VLMs, analyzing their performance, processing speed, and memory requirements.**

***Keywords*—Detection accuracy; Convolutional Neural Networks; Vision-Language Models**

I. Introduction

THERMAL imaging has become an increasingly important sensing technology, particularly in scenarios where other sensors fail. One notable application is in firefighting operations, where dense smoke severely limits visibility. Thermal cameras allow firefighters to navigate burning buildings more effectively, facilitating faster access to victims [1]. Similarly, search-and-rescue robots are often equipped with thermal cameras to conduct reconnaissance in hazardous environments and support emergency operations [2]. Another practical use case is in the automotive domain, where thermal cameras enhance visibility in dense fog, enabling the detection of pedestrians or animals on the road and thereby improving traffic safety [3].

With the rapid advancement of computer vision, machine learning - especially deep learning approaches such as Convolutional Neural Networks (CNNs) - has become the dominant methodology [4]. However, training CNNs typically requires large-scale annotated datasets. In thermal imaging, the number of publicly available datasets is significantly smaller compared to those in the RGB domain, making effective training a substantial challenge. To overcome this limitation, researchers have attempted to synthesize thermal data from RGB images using Generative Adversarial Networks (GANs) [5]-[9]. Yet, such augmentation often provides only marginal improvements in performance compared to models trained on real thermal imagery.

The introduction of the attention mechanism [10] marked a turning point in artificial intelligence, leading to the emergence of Large Language Models (LLMs). These models, powered by attention, have revolutionized the field by enabling AI to perform complex tasks such as solving advanced mathematical problems, analyzing medical documentation, or generating software code [11]. However, LLMs are inherently limited to text processing, which restricts their applicability in multimodal real-world scenarios involving signals such as images or audio. To address this limitation, multimodal LLMs have been developed, particularly Vision-Language Models (VLMs) [12], which combine the strengths of LLMs with Vision Transformers (ViTs) [13]. By leveraging vast numbers of parameters and pretraining on large, diverse datasets, VLMs can generalize effectively to domains where they have not been explicitly trained. Recent studies [14] demonstrate that VLMs achieve promising results in thermal image analysis even in zero-shot settings. This ability highlights their potential for widespread application in thermal imaging tasks.

In this paper, a novel VLM-based method for thermal image preprocessing is introduced, by which the negative influence of reflections in thermal imaging on object detection performance is substantially mitigated. Reflections constitute a particularly challenging phenomenon in thermal imagery, as they often lead to false positives and hinder the reliable identification of humans and other critical objects in safety-related scenarios. To address this issue, a dedicated dataset was constructed to enable both the training and systematic evaluation of Vision-Language Model architectures under conditions where reflective artifacts are present. Within this study, a range of open-source VLMs with different parameter scales was examined, allowing the relationship between model size, generalization capability, and detection performance to be explored in detail. Furthermore, the effects of model quantization were investigated, with a focus on identifying trade-offs between accuracy, memory requirements, and processing efficiency. By combining these analyses, a comprehensive evaluation framework was established, highlighting both the opportunities and practical constraints

Authors are with AGH University of Krakow, Krakow, Poland (e-mail:feiglewicz@agh.edu.pl, kos@agh.edu.pl).

associated with deploying VLMs for thermal imaging applications in real-world, resource-constrained environments.

II. TECHNICAL BACKGROUND

A. Fundamentals of Thermal Radiation

Every object with a temperature above 0 K emits thermal radiation. According to the Stefan–Boltzmann law [15], the radiative power of a body is expressed as:

$$P = \varepsilon\sigma T^4, \quad (1)$$

where:

ε – emissivity of the object,

σ – Stefan–Boltzmann constant,

T – absolute temperature of the object.

Thus, the power emitted by a body is directly proportional to its emissivity (ε) and proportional to the fourth power of its temperature. A thermal imaging camera records this radiation within the infrared spectrum and reconstructs it into an image. Due to the strong dependence on the fourth power of temperature, humans and other objects whose temperature differs from the environment can be readily identified.

The emissivity of a material ranges from 0 to 1 and primarily depends on its physical and chemical nature [16]. For instance, a polished metallic surface exhibits low emissivity, whereas a roughened and oxidized metallic surface has a high emissivity. In temperature measurement applications, this property often leads to inaccurate readings of the absolute temperature of objects. However, in applications such as object detection and navigation, the material-dependent emissivity enables discrimination between objects, even if they share the same ambient temperature (e.g., within a room).

Owing to these characteristics of thermal radiation, together with the fact that modern thermal cameras now provide high spatial resolution, high sensitivity, and are increasingly affordable, they are being widely adopted for navigation and object detection tasks.

Thermal imaging cameras are increasingly employed in firefighting operations, particularly during fires in which dense smoke hinders movement inside buildings and complicates search-and-rescue activities. A typical thermal camera operates within the long-wavelength infrared (LWIR) band, i.e., between 8 and 15 μm. The diameter of smoke particles typically ranges from 0.01 μm to 1 μm. Since smoke particles are significantly smaller than the wavelengths detected by thermal cameras, the scattering of thermal radiation by smoke is negligible [17]. This effect is illustrated in Figure 1.

Figure 1 presents images of the same scene captured with a visible-light camera (a) and a thermal camera (b). In the visible-light image, dense smoke obscures the people inside, rendering them invisible. In contrast, the thermal image clearly reveals the individuals, confirming the theoretical considerations discussed above.

a)

b)

Fig. 1. Images of the same scene recorded in the presence of dense smoke: (a) visible-light camera, where people are obscured, and (b) thermal imaging camera, where people can be clearly identified [18].

B. YOLO framework

The You Only Look Once (YOLO) is a one-stage, single-shot object detection framework that processes an entire image with a single forward pass of a convolutional neural network (CNN) to predict object locations and classes. Unlike two-stage detectors that first propose candidate regions and then classify them, the YOLO formulates detection as a single regression problem from image pixels to bounding-box coordinates and class probabilities, which enables very fast, real-time inference.

The schematic in Figure 2 summarizes the typical YOLO pipeline: an input image is fed into a backbone network that extracts hierarchical feature maps; these features are optionally refined and fused by a neck module to provide multi-scale contextual information; finally, the detection head produces dense predictions consisting of bounding-box coordinates and class probabilities. The head outputs are interpreted to form final detections (boxes with associated class labels) after non-maximum suppression. YOLO's single-pass design gives it strong runtime performance, making it well suited for applications that require low latency (e.g., robotics, video analytics, and real-time monitoring). Although different YOLO versions introduce architectural variations and improvements, they all preserve this fundamental backbone–neck–head structure.

C. Performance metrics for object detection

Performance metrics are a key component for evaluating the accuracy and efficiency of artificial intelligence models used in object detection. One of the fundamental metrics is the Intersection over Union (IoU), defined as:

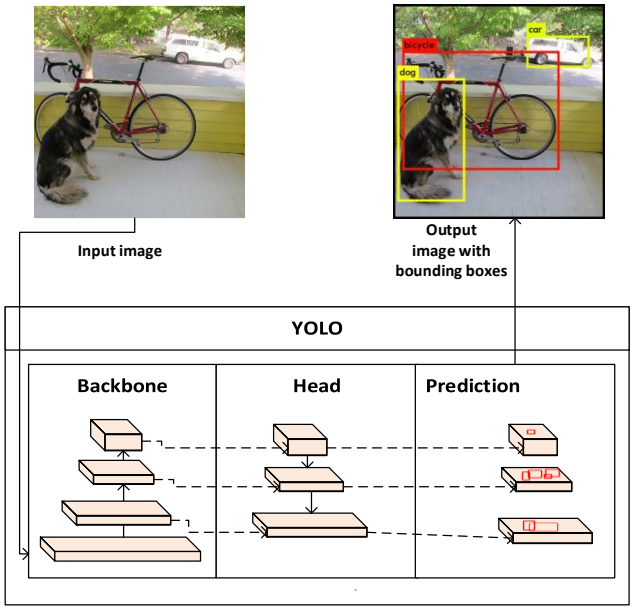

Fig. 2. Schematic representation of the YOLO object detection framework: the input image is processed by the backbone network, features are passed through the head, and the final prediction yields bounding boxes and class labels [19]-[20].

$$IoU = \frac{Area\ of\ Intersecion}{Area\ of\ Union}. \quad (2)$$

IoU measures the degree of overlap between a predicted bounding box and the ground truth bounding box. It provides a numerical value that quantifies how well the model's prediction aligns with the actual object location. Based on IoU and a predefined threshold (in this article set to 50% in all experiments), each prediction can be classified into one of the following categories:

- **True Positive (TP):** the model correctly identifies an object, and the IoU with the ground truth bounding box exceeds the threshold.
- **False Positive (FP):** the model incorrectly predicts an object that does not exist in the ground truth, or the IoU with the ground truth bounding box is below the threshold.
- **False Negative (FN):** the model fails to detect an object that is present in the ground truth.
- **True Negative (TN):** generally not applicable in object detection tasks, since the task typically focuses on the presence and localization of objects rather than explicitly confirming their absence.

From TP, FP, and FN, three widely used evaluation metrics are derived:

$$Precision = \frac{TP}{TP+FP}. \quad (3)$$

Precision quantifies the proportion of correctly identified objects among all detections made by the model. High precision indicates that false detections are rare.

$$Recall = \frac{TP}{TP+FN}. \quad (4)$$

Recall measures the proportion of ground truth objects that are correctly detected by the model. High recall indicates that most objects are successfully found.

$$\text{F1-Score} = \frac{Precision \cdot Recall}{Precision + Recall}. \quad (5)$$

The F1-score is the harmonic mean of precision and recall, providing a balanced metric that is especially useful when an application requires both accurate and comprehensive detection.

Together, these metrics offer a comprehensive assessment of an object detection model, capturing its ability to avoid false alarms (precision), detect as many objects as possible (recall), and balance the two aspects (F1-score) [21].

D. Vision Language Model

Vision-Language Models are a class of multimodal artificial intelligence models designed to jointly process and reason over visual and textual information. Unlike conventional vision-only or text-only models, VLMs integrate both modalities, enabling tasks such as image captioning, visual question answering, or zero-shot object recognition. As illustrated in Figure 3, a VLM can take a visual input together with a textual query and generate an appropriate natural language description, demonstrating its ability to connect visual information with text [22].

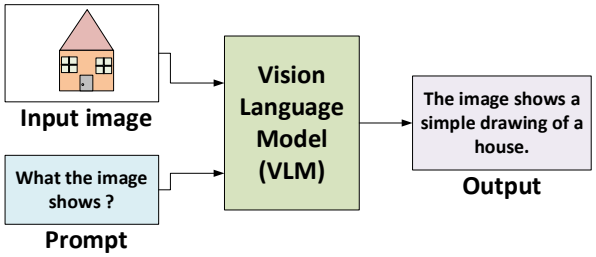

Fig. 3. Illustration of how a Vision Language Model (VLM) works. The model is given a simple drawing of a house together with the question "What the image shows?" and produces the textual description: "The image shows a simple drawing of a house." This demonstrates the model's ability to interpret visual input and provide a meaningful natural language response.

A Vision-Language Model (VLM) typically consists of three main components, as illustrated in Figure 4. The vision encoder is responsible for processing the image and extracting visual features in a numerical form. These features are then mapped into the language space by the projector, which aligns the visual representation with the format understood by the language model. Finally, the large language model (LLM) takes the projected features along with textual input and generates

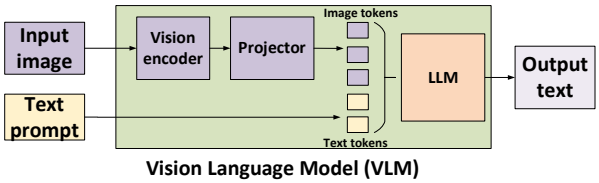

Fig. 4. Overview of a vision-language model architecture. The input image is processed by a vision encoder and a projector to generate image tokens, which are combined with text tokens from the prompt and passed to the large language model (LLM) to produce output text.

a natural language response, enabling tasks such as image captioning or visual question answering [23].
Due to their pretraining on massive and diverse datasets, Vision-Language Models (VLMs) demonstrate strong generalization capabilities, allowing them to effectively solve tasks for which they were never explicitly trained or were only fine-tuned using small, high-quality datasets. Consequently, applying VLMs to the analysis of thermal images may significantly improve the accuracy and robustness of such analyses, particularly given the relative scarcity of thermal image datasets compared to those in the visible spectrum.

E. Quantization

Quantization is one of the most widely used techniques for reducing the size of neural network models. Modern large language models (LLMs) can contain hundreds of billions or even several trillion parameters, resulting in significant computational overhead and high memory requirements - particularly for GPU-based inference, where large amounts of VRAM are needed.

Quantization is a technique that reduces the precision of weight and activation values. Typically, weights and activations are stored using 32-bit floating-point precision (FP32) or 16-bit formats such as FP16 or Brain Float 16 (BF16). Through quantization, these values can be represented using lower-precision formats, such as 8-bit floating-point (FP8) or even 4-bit formats like FP4.

Neural networks are generally robust to quantization error, meaning that compression from 32-bit to 4-bit precision often leads to only a modest degradation in model accuracy. Figure 5 presents a comparison of several numerical formats. The BF16 format shares the same dynamic range as FP32 but with reduced precision, while FP16 offers a smaller range yet slightly better precision than BF16. FP8 formats, depending on the specific variant, typically use 4 exponent bits and 3 mantissa bits (e.g., the E4M3 format). FP4, the smallest among these formats, can represent values approximately in the range of –6.0 to +6.0 [24].

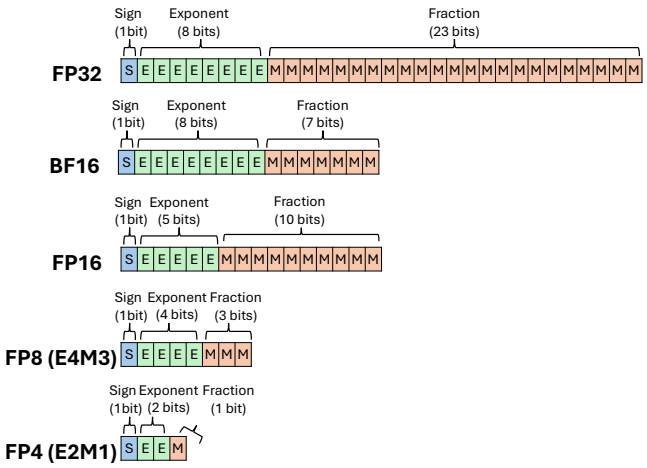

Fig. 5. Bit-level representation of various floating-point formats used in neural network quantization. Each format consists of a sign bit (S), exponent bits (E), and mantissa bits (M). FP32 (IEEE 754 single precision) uses 1 sign bit, 8 exponent bits, and 23 mantissa bits. FP16 and BF16 both use 16 bits in total, with BF16 preserving a wider dynamic range and FP16 offering slightly higher precision. The FP8 format (E4M3) uses 8 bits, providing a balance between range and precision. FP4 is the most compact format, using only 4 bits for extremely low-precision computations.

Due to the limited range and precision of such low-bit representations, scaling is required before performing arithmetic operations. For example, the NVIDIA FP4 (NVFP4) format employs a two-level scaling mechanism: first, a coarse per-tensor scaling factor stored in FP32, followed by fine-grained scaling at the block level, where each 16-element block is scaled using an FP8 (E4M3) factor.

These techniques enable minimal degradation of model accuracy, significant memory savings, and - in hardware supporting low-precision arithmetic - substantial acceleration of inference speed.

III. PROBLEM DESCRIPTION

In thermal imaging, reflection phenomena occur primarily on smooth surfaces such as glass, metal, and even polished concrete [25]. This effect becomes particularly problematic when a thermal camera is used for navigation purposes - for example, by a mobile robot - since reflections can lead to incorrect environmental mapping, thereby hindering or even preventing reliable navigation in indoor environments.

a) b)

Fig. 6. Comparison of the same scene captured using a visible-light camera (a) and a thermal camera (b) processed by the YOLOv11 object detection model. As shown, the reflection from the glass door labeled No. 2 was incorrectly identified as a person, whereas it is actually the reflection of the real person labeled No. 1. The thermal image was captured using a Seek Thermal Nano 300 camera.

Figure 6b illustrates the performance of the YOLOv11 model in detecting humans in a thermal image. Due to a reflection from a glass door, the region enclosed by bounding box No. 2 was incorrectly classified as a real person, while in reality, it represents the reflection of the actual human marked by bounding box No. 1.

In order to assess the performance of thermal image analysis methods in the presence of reflections, the authors developed a publicly accessible dataset [26]. The dataset comprises thermal images acquired mainly in a shopping mall and a single-family house, containing various reflections produced by glass, metallic, and tiled surfaces.

The results obtained using the YOLOv11 model for human detection on the aforementioned dataset are presented in Figure 7 as a confusion matrix. A total of 403 bounding boxes corresponding to actual humans were correctly classified as people (True Positives), while 162 bounding boxes were

incorrectly classified as humans but were in fact reflections (False Positives). No real human objects were missed; therefore, the number of False Negatives is zero. Based on the confusion matrix results, the calculated F1-score is 0.83, which is not satisfactory. To improve this value, the number of False Positives should be reduced without increasing the number of False Negatives.

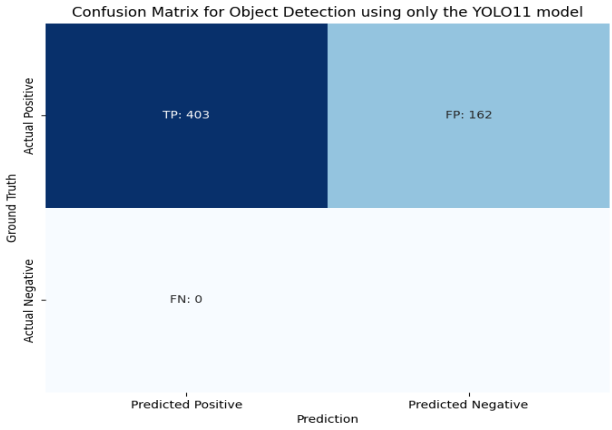

Fig. 7. Confusion matrix illustrating the performance of the YOLOv11 model for human detection on the proposed thermal image dataset. The model correctly detected 403 real human instances (True Positives) and misclassified 162 reflections as humans (False Positives). No actual human objects were missed (False Negatives = 0).

IV. Proposed Method

Since the results obtained using only the YOLOv11 model were not satisfactory, the authors of this paper developed an enhanced method to improve human detection performance in thermal imagery under reflective conditions. To achieve this, an additional processing stage based on Vision-Language Models (VLMs) was introduced. The processing pipeline is illustrated in Figure 8. First, the thermal image is processed by the YOLOv11 model in a standard manner, producing a set of bounding boxes corresponding to detected objects. These bounding boxes are then overlaid on the image and assigned numerical identifiers. The annotated image is subsequently passed to the VLM along with a carefully designed prompt instructing the model to return a list of bounding box numbers that correspond exclusively to reflections, sorted in ascending order. By removing these reflection-related bounding boxes, an image containing only the true human detections is obtained.

To enhance the reliability of this approach, the prompt begins

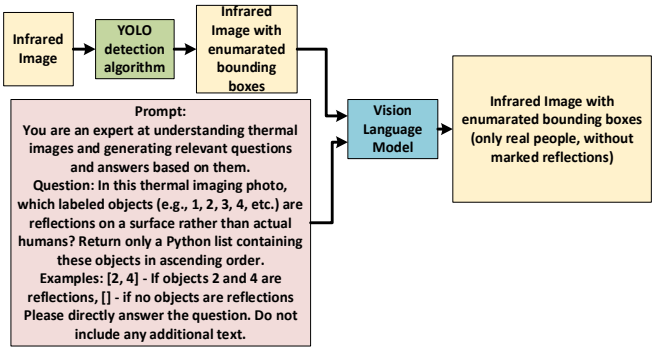

Fig. 8. Processing pipeline for improving human detection in thermal images using a combined YOLOv11 and Vision Language Model (VLM) approach. The input infrared image is first processed by the YOLOv11 detection algorithm, which generates and enumerates bounding boxes around detected objects. The annotated image is then passed to a Vision-Language Model along with a structured prompt instructing it to identify which bounding boxes correspond to reflections.

with a statement indicating that the VLM is an expert in thermal image analysis, followed by the specific task description and several example responses. Finally, the prompt explicitly instructs the model to return the output as a Python list containing only numeric values, without any additional text. This design ensures consistent output formatting and facilitates subsequent automated processing. Empirical observations indicate that such structured prompts yield better results than prompts containing only a single query.

The comparison of detection accuracy depending on the VLM is presented in Table I. The evaluated VLMs were not fine-tuned for reflection-related tasks in thermal imaging. The best performance was achieved by the closed-source Gemini 2.5 Flash model, which obtained an F1-score above 0.97 -

TABLE I
PERFORMANCE COMPARISON OF YOLOV11 AND COMBINED YOLOV11-VLM APPROACHES FOR HUMAN DETECTION IN THERMAL IMAGERY WITHOUT FINE-TUNING.

Model Specs				Accuracy Metrics			Confusion Matrix		
Model name	Open Source?	Number of Parameters	Parameter Data Type	Recall	Precision	F1-score	TP	FP	FN
Without VLM – only YOLOv11	N/A	N/A	N/A	1.0000	0.7133	0.8326	403	162	0
Gemini 2.5 flash	×	N/A	N/A	0.9603	0.9949	**0.9773**	387	2	16
Qwen2-VL-7B	✓	7B	BF16	0.6824	0.8514	0.7576	275	48	128
Gemma-3-4B	✓	4B	BF16	0.3548	0.6976	0.4704	143	62	260
Gemma-3-12B	✓	12B	BF16	0.7122	0.8777	0.7863	287	40	116
InternVL3-8B	✓	8B	BF16	0.8437	0.8924	0.8673	340	41	63

considered an excellent result. Among open-source models, only InternVL-8B outperformed the baseline YOLOv11-only approach, though the improvement was marginal. The remaining models performed worse than YOLOv11 alone in human detection tasks.

The use of closed-source models requires computation to be carried out on external servers, which prevents local deployment. Consequently, applying such models for vehicle navigation support may be infeasible in certain environments due to limited internet access. Moreover, network latency associated with transmitting thermal images to remote servers may also be prohibitive for real-time applications.

To enhance the performance of open-source models, a small dataset was prepared for fine-tuning VLMs specifically on reflection detection in thermal imagery [27]. The dataset was created using thermal images sourced from several publicly available datasets in which reflections were visible [28–29]. To further augment the dataset, in addition to prompts related to identifying reflections, supplementary questions were generated to locate real humans and to associate reflections with their corresponding reflection sources.

For fine-tuning, the popular Low-Rank Adaptation (LoRA) technique was employed [30], which enables efficient training by updating only a small subset of low-rank matrices inserted into the model's weight structure. This significantly reduces the number of trainable parameters, memory usage, and computational cost while preserving the expressive power of the original model. LoRA is particularly effective for large Vision-Language Models, as it avoids modifying the full parameter space and allows training on consumer-grade hardware without compromising performance.

The fine-tuning process was carried out on a desktop PC equipped with an Nvidia GeForce RTX 5090 GPU, which handled all computations. Thanks to LoRA, the hardware requirements remained modest despite the size of the underlying VLM.

The accuracy results obtained after fine-tuning are presented in Table II. All evaluated models improved their F1-score compared with their non-fine-tuned counterparts. The highest performance was achieved by Gemma-3-12B, which exceeded an F1-score of 0.93. Qwen2-VL-7B and InternVL3-8B achieved only slightly lower scores, despite being significantly smaller models, which may translate into faster inference in practical deployments. The smallest model, Gemma-3-4B, reached an F1-score of approximately 0.58, improving from 0.47 without fine-tuning; however, this value remains lower than the performance of methods that do not use VLMs at all. The limited parameter count of such small models restricts their ability to learn and generalize patterns from a relatively small number of fine-tuning examples, which explains their reduced effectiveness.

Since the use of fine-tuning significantly improved the accuracy of open-source VLMs - making them suitable for eliminating the influence of reflections in thermal-image human detection - further optimization was performed to reduce processing time and VRAM consumption. To achieve this, model quantization was applied. Two frameworks were used: Transformers, a universal solution that allows running VLMs on hardware from various vendors, and TensorRT-LLM, a specialized library designed for quantizing and deploying models on NVIDIA GPUs.

The models were quantized to the following formats: Normal Float 4 (NF4), FP8, and NVFP4. Additionally, the effect of reducing the input image size from 512×512 to 256×256 on both processing speed and accuracy was evaluated. The results are presented in Table III.

For the 512×512 input size, the F1-score for all tested quantizations remained above 0.9, and in some cases even exceeded the baseline BF16 model. The best result was obtained using the InternVL3-8B model quantized to NF4. However, NF4 quantization caused a notable increase in inference time - approximately 30% slower than the BF16 baseline. This slowdown occurs because NF4 is not natively supported in hardware, so model parameters must be converted to higher-precision formats (typically FP16 or BF16) during computation.

For this reason, NVFP4 is generally a better choice when supported by the GPU: it offers F1-scores comparable to NF4, similar memory usage, and can be over three times faster in inference. Both NVFP4 and NF4 reduce memory consumption by roughly a factor of two, enabling deployment on devices with limited hardware resources. FP8 provides intermediate performance in terms of both memory footprint and speed, making it a reasonable alternative when NVFP4 is not supported by the GPU architecture.

TABLE II
PERFORMANCE COMPARISON OF YOLOV11-VLM APPROACHES FOR HUMAN DETECTION IN THERMAL IMAGERY WITH FINE-TUNED VLM MODELS.

Model Specs				Accuracy Metrics			Confusion Matrix		
Model name	Open Source?	Number of Parameters	Parameter Data Type	Recall	Precision	F1-score	TP	FP	FN
Qwen2-VL-7B	✓	7B	BF16	0.9330	0.9261	0.9295	376	30	27
Gemma-3-4B	✓	4B	BF16	0.5012	0.7014	0.5847	202	86	201
Gemma-3-12B	✓	12B	BF16	0.9752	0.8932	**0.9324**	393	47	10
InternVL3-8B	✓	8B	BF16	0.9702	0.8947	0.9310	391	46	12

TABLE III
PERFORMANCE OF FINE-TUNED VLMS UNDER DIFFERENT QUANTIZATION SETTINGS.

Model Specs				Results				
Model name	Number of Parameters	Framework	Parameter Data Type	F1-score Input image size [pix^2]		Memory footprint [Gb]	Processing time [ms] Input image size [pix^2]	
				256	512	-	256	512
Qwen2-VL-7B	7B	Transformers	BF16	0.8810	0.9242	16.897	85.2	127
Qwen2-VL-7B	7B	Transformers	NF4	0.8747	0.9212	**7.168**	152	176
Qwen2-VL-7B	7B	TensorRT-LLM	BF16	0.9077	0.9271	17.549	56.1	83.9
Qwen2-VL-7B	7B	TensorRT-LLM	FP8	0.8968	0.9397	11.606	43.7	68.2
Qwen2-VL-7B	7B	TensorRT-LLM	NVFP4	0.9019	0.9121	8.860	**37.2**	49.7
Gemma-3-12B	12B	Transformers	BF16	0.9327	0.9269	24.542	267	265
Gemma-3-12B	12B	Transformers	NF4	0.9156	0.9069	9.229	365	356
InternVL3-8B	8B	Transformers	BF16	0.9300	0.9233	16.077	88.2	91.7
InternVL3-8B	8B	Transformers	NF4	0.9369	**0.9406**	8.245	119	120
InternVL3-8B	8B	TensorRT-LLM	BF16	0.9262	0.9358	16.366	69.8	71.2
InternVL3-8B	8B	TensorRT-LLM	FP8	0.9225	0.9309	10.395	52.1	53.6
InternVL3-8B	8B	TensorRT-LLM	NVFP4	0.9252	0.9332	7.647	37.6	39.4

When reducing the input resolution from 512×512 to 256×256, a decrease in F1-score was observed for most configurations. Notably, for InternVL3-8B and Gemma-3-12B, this reduction in resolution did not provide significant speed improvements. In contrast, Qwen2-VL-7B exhibited a processing-time improvement of approximately 50%. This difference stems from the fact that Qwen2-VL-7B internally splits the input image into smaller patches, whereas the other models always rescale the image to a fixed resolution.

Therefore, if increased inference speed is required, using lower-resolution images is beneficial specifically for Qwen2-VL-7B, albeit at the cost of reduced accuracy.
In Fig. 9, the processing time per thermal image obtained using the Qwen2-VL-7B–based algorithm is compared with the required VRAM for different quantization methods. The Transformers framework provides memory consumption comparable to TensorRT-LLM for both 16-bit and 4-bit quantization schemes. However, TensorRT-LLM achieves substantially lower inference latency, particularly for 4-bit quantization, where NVFP4 is more than three times faster than NF4.

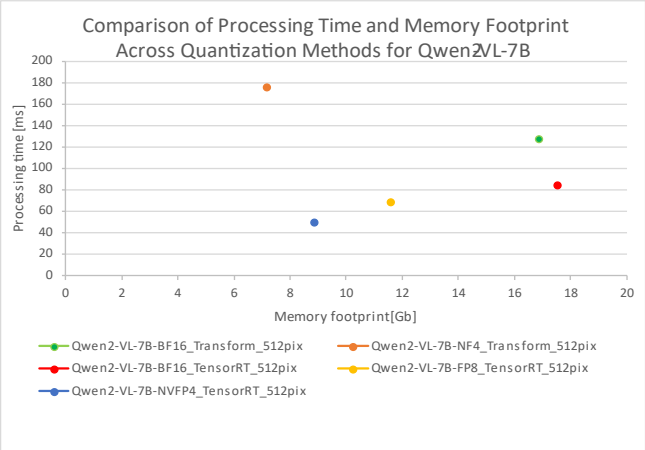

Fig. 9 Processing Time vs. Memory Footprint for Qwen2-VL-7B Under Different Quantization Methods. All results were obtained using 512×512 input images. Quantization to NF4 and FP8 significantly reduces memory usage compared to BF16, while NVFP4 achieves the best trade-off by providing both low memory consumption and the fastest inference among the tested configurations. The increase in processing time for NF4 is attributed to the need for on-the-fly conversion to higher-precision formats due to lack of native hardware support.

CONCLUSION

This work demonstrates that Vision-Language Models can be effectively applied to thermal image analysis, even when only small task-specific training datasets are available. Current closed-source models with hundreds of billions of parameters are capable of achieving high accuracy on thermal imagery without any additional fine-tuning. In contrast, smaller open-source VLMs, containing only a few to several billion parameters, require fine-tuning to reach comparable performance on the target task; however, even limited fine-tuning on a modest dataset is sufficient to close most of the performance gap.

Furthermore, applying quantization significantly reduces memory requirements and can substantially accelerate inference, making VLMs more suitable for deployment in resource-constrained environments. These results highlight the need for continued development of hardware and inference frameworks optimized for running large multimodal models directly on edge devices, such as mobile robots, where on-device thermal image interpretation is required.

REFERENCES

[1] Tsai, P.-F.; Liao, C.-H.; Yuan, S.-M. Using Deep Learning with Thermal Imaging for Human Detection in Heavy Smoke Scenarios. Sensors 2022, 22, 5351. https://doi.org/10.3390/s22145351

[2] P. Fritsche, B. Zeise, P. Hemme and B. Wagner, "Fusion of radar, LiDAR and thermal information for hazard detection in low visibility environments," 2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR), Shanghai, China, 2017, pp. 96-101. https://doi.org/10.1109/SSRR.2017.8088146

[3] P. Tumas, A. Nowosielski and A. Serackis, "Pedestrian Detection in Severe Weather Conditions," in IEEE Access, vol. 8, pp. 62775-62784, 2020, http://doi.org/10.1109/ACCESS.2020.2982539

[4] Raj, R.; Kos, A. An Extensive Study of Convolutional Neural Networks: Applications in Computer Vision for Improved Robotics Perceptions. Sensors 2025, 25, 1033. https://doi.org/10.3390/s25041033

[5] Yun, K., Yu, K., Osborne, J., Eldin, S., Nguyen, L., Huyen, A., & Lu, T. (2019, May). Improved visible to IR image transformation using synthetic data augmentation with cycle-consistent adversarial networks. In Pattern Recognition and Tracking XXX (Vol. 10995, p. 1099502). SPIE. https://doi.org/10.48550/arXiv.1904.11620

[6] Li, Y., Ko, Y. & Lee, W. A Feasibility Study on Translation of RGB Images to Thermal Images: Development of a Machine Learning Algorithm. *SN COMPUT. SCI.* 4, 555 (2023). https://doi.org/10.1007/s42979-023-02040-4

[7] Khan, Md Azim. "Visible to Thermal image Translation for improving visual task in low light conditions." arXiv preprint arXiv:2310.20190 (2023). https://doi.org/10.48550/arXiv.2310.20190

[8] Uddin MS, Kwan C, Li J. MWIRGAN: Unsupervised Visible-to-MWIR Image Translation with Generative Adversarial Network. Electronics. 2023; 12(4):1039. https://doi.org/10.3390/electronics12041039

[9] Mohamed El Mahdi, B., Abdelkrim, N., Abdenour, A., Zohir, I., Wassim, B., & Fethi, D. (2023). A Novel Multispectral Maritime Target classification based on ThermalGAN (RGB-to-Thermal Image Translation). Journal of Experimental & Theoretical Artificial Intelligence, 36(8), 1757–1777. https://doi.org/10.1080/0952813X.2023.2165723

[10] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[11] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv preprint arXiv:2402.06196. https://doi.org/10.48550/arXiv.2402.06196

[12] Zhang, Jingyi, et al. "Vision-language models for vision tasks: A survey." IEEE transactions on pattern analysis and machine intelligence 46.8 (2024): 5625-5644. https://doi.org/10.48550/arXiv.2304.00685

[13] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020). https://doi.org/10.48550/arXiv.2010.11929

[14] Ashqar HI, Alhadidi TI, Elhenawy M, Khanfar NO. Leveraging Multimodal Large Language Models (MLLMs) for Enhanced Object Detection and Scene Understanding in Thermal Images for Autonomous Driving Systems. Automation. 2024; 5(4):508-526. https://doi.org/10.3390/automation5040029

[15] Yang, J., Zhang, X., Zhang, X., Wang, L., Feng, W., & Li, Q. (2021). Beyond the visible: bioinspired infrared adaptive materials. Advanced Materials, 33(14), 2004754. https://doi.org/10.1002/adma.202004754

[16] Marzec, P., & Kos, A. (2021). Thermal navigation for blind people. Bulletin of the Polish Academy of Sciences: Technical Sciences, (1). https://doi.org/10.24425/bpasts.2021.136038

[17] Szajewska, A. (2017). Development of the thermal imaging camera (TIC) technology. Procedia Engineering, 172, 1067-1072. https://doi.org/10.1016/j.proeng.2017.02.164

[18] Liu, J., Fan, X., Huang, Z., Wu, G., Liu, R., Zhong, W., & Luo, Z. (2022). Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5802-5811). https://doi.org/10.48550/arXiv.2203.16220

[19] Elesawy A, Mohammed Abdelkader E, Osman H. A Detailed Comparative Analysis of You Only Look Once-Based Architectures for the Detection of Personal Protective Equipment on Construction Sites. Eng. 2024; 5(1):347-366. https://doi.org/10.3390/eng5010019

[20] Object Detection: The Definitive Guide. [Online]. Available: https://viso.ai/deep-learning/object-detection/. [Accessed: Sept. 15,2025]

[21] Metrics Matter: A Deep Dive into Object Detection Evaluation. [Online]. Available:https://medium.com/@henriquevedoveli/metrics-matter-a-deep-dive-into-object-detection-evaluation-ef01385ec62. [Accessed: Oct. 4,2025]

[22] Alammar, J., & Grootendorst, M. (2024). Hands-on large language models: language understanding and generation. " O'Reilly Media, Inc.".

[23] What Are Vision Language Models. [Online]. Available:https://www.nvidia.com/en-us/glossary/vision-language-models/. [Accessed: Oct. 4,2025]

[24] Introducing NVFP4 for Efficient and Accurate Low-Precision Inference. [Online]. Available: https://developer.nvidia.com/blog/introducing-nvfp4-for-efficient-and-accurate-low-precision-inference/. [Accessed: Oct. 11,2025]

[25] Batchuluun, G., Yoon, H. S., Nguyen, D. T., Pham, T. D., & Park, K. R. (2019). A study on the elimination of thermal reflections. IEEE Access, 7, 174597-174611.

[26] https://huggingface.co/datasets/rfeiglew/ThermalRefl_eval_metrics [Accessed: Oct. 25,2025]

[27] https://huggingface.co/datasets/rfeiglew/ThermalRefl [Accessed: Oct. 30,2025]

[28] Felsberg, Michael, et al. "The thermal infrared visual object tracking VOT-TIR2015 challenge results." Proceedings of the ieee international conference on computer vision workshops. 2015. https://www.cv-foundation.org/openaccess/content_iccv_2015_workshops/w14/papers/Felsberg_The_Thermal_Infrared_ICCV_2015_paper.pdf

[29] Jia, Xinyu, et al. "LLVIP: A visible-infrared paired dataset for low-light vision." Proceedings of the IEEE/CVF international conference on computer vision. 2021. https://doi.org/10.48550/arXiv.2108.10831

[30] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3. https://doi.org/10.48550/arXiv.2106.09685