National College of Ireland

Vision-Language Model Fine-Tuning for Face Mask Detection Using Parameter-Efficient Methods

MSc Research Project

MSc in Artificial Intelligence

Chandrika Gollapudi
Student ID: X23319968

School of Computing
National College of Ireland

Supervisor: Mohit Garg

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Chandrika Gollapudi
Student ID:	X23319968
Programme:	MSc in Artificial Intelligence **Year:** 2025
Module:	MSC (Research) Practicum part-2
Supervisor:	Mohit Garg
Submission Due Date:	28/01/2026
Project Title:	Vision-Language Model Fine-Tuning for Face Mask Detection Using Parameter-Efficient Methods
Word Count:	8954 **Page Count** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Chandrika Gollapudi

Date: 28/01/2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Vision-Language Model Fine-Tuning for Face Mask Detection Using Parameter-Efficient Methods

Chandrika Gollapudi
Student ID: X23319968

Abstract

During the COVID-19 pandemic, it became very important for public health to keep an eye on people who wear face masks. However, the current detection systems use traditional convolutional neural networks that only give bounding boxes and class labels without any clear explanations. This study examines the potential of vision-language models, fine-tuned using parameter-efficient techniques, to attain competitive detection accuracy while providing improved interpretability through natural language outputs. The research assesses two parameter-efficient fine-tuning methodologies, Low-Rank Adaptation and Quantized Low-Rank Adaptation, implemented on Microsoft's Florence-2 vision-language model for face mask detection across three compliance categories. Experiments employed an extensive dataset comprising 28,297 images and 63,035 annotations, characterized by a pronounced class imbalance. The LoRA-adapted Florence-2-large reached 71.6% mAP@50 with 88.4% precision and 87.3% recall while training only 0.7% of model parameters. This was higher than the seventy percent accuracy level that was set as the primary success level. QLoRA setting achieved 40.1 percent mAP at 50 indicating that aggressive quantizing four-bit quantization has a significant impact on detecting vision tasks that require accurate spatial positioning. Vision-language models demonstrate that they can attain realistic detection accuracy to compliance monitoring uses as well as equip output understandable building blocks when compared to a YOLO11m baseline (95.6% mAP@50). The findings indicate that fine-tuning using a small number of parameters is an appropriate method to harmone large vision-language models to safety monitoring duties in a particular area and does not require a significant amount of computing hardware.

1 Introduction

1.1 Background

The COVID-19 emergency gave a remarkable redesign of the entire infrastructure in the world in the field of public health, making the use of face masks one of the main interventions in combating the spread of respiratory diseases. The use of masks in healthcare centers, transportation methods, educational institutes, and commercial areas around the world combined to rise in demand of automated compliance monitoring systems like never before. The enforcement by hand was not possible in a high-traffic setting, and the needed technological solutions were the ones that would allow detecting human activities and responding in real time (Sharma et al., 2023).

Classical computer vision methods of face mask detection have been based mainly on Convolutional Neural Network models, specifically YOLO and variants and region-based detectors. The systems have a very high accuracy in the controlled settings and more recent implementations have reported over 90 percent in detection rates on benchmark datasets Sethi et al. (2021). Nevertheless, they perform only numerical tasks, such as box coordinates, class labels and confidence values without providing information on violations found in any particular aspect.

1.2 Motivation

This fact that modern detection systems lack interpretability poses a great risk of operational issues affecting compliance monitoring tools. Health officials and security personnel need instant evaluation of the level of violations and contextual judgment on what to do about them (Moradi et al., 2024). Traditional detectors are unable to differentiate the case of temporary mask change rather than the deliberate case of non-compliance, or medical exemption over the case of willful non-compliance. Such an interpretability gap constrains their practical implementation in real world enforcement contexts where explanation and documentation are critical both in legal and administrative contexts.

Vision-Language Models (VLMs) regularly provide the potential remedy because of integrated frameworks bringing insight into visual perception with natural speech comprehension. An example of this method is Microsoft Florence-2 (Xiao et al., 2024) which produces both natural language descriptions and spatial localization. Nevertheless, VLMs have not been studied with safety compliance usage as they often need hundreds of millions of parameters or more to be computed Xiao et al. (2024). Small businesses, education facilities, and healthcare facilities do not have high-performance computing infrastructure that may be needed to implement conventional VLM application.

Recent developments with parameter-efficient fine-tuning, especially Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) have shown promising solutions to be used when computational constraints are available (Hu et al., 2022; Dettmers et al., 2023). These methods are able to reduce the number of trainable parameters to less than 1% original models and still achieve a competitive performance. But, their applicability to tasks that are vision-specific and need localization of spatial aspects to become specific is not investigated, and it is a vital research gap.

1.3 Research Objectives

The main question that is explored in this study is as follows: Can Vision-Language Models, adapted through parameter-efficient fine-tuning, achieve competitive face mask detection accuracy while providing interpretable natural language outputs?

The research study has five goals:

Objective 1: Build a baseline of the performance of YOLO11m, which reflects the state of art traditional detection possibilities concerning face mask compliance checking.

Objective 2: Implement Low-Rank Adaptation fine-tuning on Florence-2-large to evaluate the parameter-efficient VLM adaptation on detection tasks.

Objective 3: Compare Quantized Low-Rank Adaptation using 4-bit quantization to investigate extreme memory optimization compromises to deploy in a resource-constrained environment.

Objective 4: Systematic comparison of all of the approaches based on common COCO detection measurements.

Objective 5: Determine the efficiency of the parameters with respect to the detection performance so as to identify the best accuracy-efficiency trade-offs.

1.4 Thesis Approach

A comparative experimental study design that compares three different detection methods on a large face mask dataset of 28297 images and 63035 annotated images on three compliance categories is used in the research. The experimental design comparatively assesses traditional object detection and parameter-efficient VLM adaptations excellently.

Baseline method applies YOLO11m by completely fine-tuning the model based on maximum achievable accuracy with conventional methods. LoRA uses rank-32 matrices of adaptation to

the base case 2-large Florence attention models with a 16-bit precision, and is trained on 0.7 percent of the model parameters. The QLoRA model adopts 4-bit NormalFloat quantization as Florence-2-base, use rank-16 adapters and puts a high emphasis on memory efficiency due to edge deployment.
The performance is assessed using standardized performance metrics such as mean Average Precision at variousIoU levels, precision-recall and per-class performance measures. Success metrics are those concerned with a realistic deployment need: ≥70% mAP@50 accuracy, are able to cut the number of trainable parameters by ≥95% with the ability to train on a consumer spike and ≤16GB/memory.

1.5 Thesis Contributions

The study has some of the new contributions to the computer vision and safety monitoring fields:
Introduction of VLM in the framework of safety compliance during pandemic: The study is the first to use VLM in the area of face mask detection, which proves the viability of explainable compliance monitoring systems.
Systematic comparison of parameters efficiency of fine-tuning based vision tasks: The study is the first large-scale comparison of the effectiveness of LoRA and QLoRA in technology transfer using VLMs to object detection showing a trade-off between quantization and spatial localization accuracy.
Vlm as a resource-constrained practical system: The research provides principled guidelines to resolve the deployment of advanced vision-language systems on consumer machines through the institutionalization of access towards interpretable AI, a resource-wide democratization of impactful artificial intelligence.
Empirical test of VLM-detection performance gap: The experiment accounts the accuracy difference between special detectors and modified VLMs and bases the decisions on the deployment of these technologies in accordance with the demands of accuracy and interpretability.

1.6 Thesis Scope

The study is narrowed down to face mask detection as one of the representative safety compliance tasks, assessing the three different classes, namely, correctly worn masks, incorrectly worn masks, and no masks at all. The research area covers technical feasibility test and coverage of performance instead of production implementation and actual implementation of the system.
The paper also measures the accuracy of detection and efficiency of the parameter, but it fails to measure objectively the quality of natural language output is also a focus of future research in the form of human evaluation researches. Experiments are based on publicly accessible datasets that have inherent demographical/geographical shortcomings that could complicate relating to more diverse populations.
The work compares architectural methods as compared to exhaustive hyperparameter optimization as it aims at determining viability as opposed to attaining as much performance as possible. Experiments are run on cloud-based GPU infrastructure, but deployment on a theoretical basis, and no consistency is made with actual edge devices.

1.7 Report Organization

In the rest of this report, the literature review on face mask detection methods, vision-language models, and parameter-efficient fine-tuning techniques is as follows: in Section 2, we perform the literature review in the form of a literary analysis of the above-mentioned methods and models. Section 3 gives information on the research methodology such as data characteristics

and experimental design. The system design specification and architectural details are found in section 4. Section 5 outlines implementation processes as well as technical decisions. Section 6 provides results of experiments accompanied by comparative analysis. Section 7 covers findings as they are related to the research objectives and extant literature. Section 8 is the conclusion and conclusion containing contribution and future direction of research.

2 Related Work

In this section, a critical analysis of existing studies in three connected fields, namely, face mask detection algorithms, architecture of vision-language models, and strategies of parameter-efficient fine-tuning, will be presented. The review locates the current knowledge base, points out the weaknesses found in existing methods, and supports the innovative value of the current study.

2.1 Face Mask Detection Using Deep Learning

The experience of the COVID-19 jumpstarted extensive studies on the topic of automated face mask-detecting, the deep-learning approaches, in turn, had the most positive results. Sethi et al. (2021) conducted a comprehensive evaluation of transfer learning models in detecting face masks and they compared ResNet 50, VGG16, and Inception V3. They defined that it is possible to use pre-trained convolutional networks to classify the mask, and the accuracy of such models was 91% to 94% in their test images (11,786 images). This technique is good since it is not complex and utilises transfer learning due to ImageNet pre-training. However, the approach treats mask detection as a classification problem, rather than as an object detection problem and therefore face localization must occur prior to the possibility of classification to occur. This two-stage pipeline has latency and likelihood of error propagation between pipeline stages.

According to Loey et al. (2021), a mixed architecture based on finding faces with the use of the Deep Learning algorithm, namely YOLO version two, and sorting the masks with ResNet50 should be used. The obtained result was that this method was correct on a test of 1,376 images 81% of the time, indicating that object detection and classification networks can be used together. The most important advantage with this technique is that it is able to work on data that starts all the way to the end, and therefore, it is quick with respect to surveillance uses. Nevertheless, the accuracy and the size of the collected data are relatively low, which makes it hard to appreciate how it will perform in other real-life scenarios. It is also difficult to deploy and requires more computing power when it comes to the two-model architecture.

Nagrath et al. (2021) introduced SSDMNV2 that combines Single Shot MultiBox Detector with MobileNetV2 as the backbone to detect face mask in real-time. This architecture scored 92.64 percent and was also fast enough to be used on mobiles. The small size of MobileNetV2 is a major advantage on the edge deployment scenario. Only 2,165 images were used to evaluate it, however, thus the performance measures that are reported are not that reliable. The system even does not provide any compliance status information but provides only the results of detection.

Jiang et al. (2021) presented RetinaFaceMask, which is a one-stage detector based on the RetinaFace architecture modified to support simultaneous face and mask recognition. The precision of this technique attained 93.4 percent on a sample of 7,959 images, one of the high scores in the literature. This single stage design has computational advantages over the two stage design. Like other conventional methods of detection, RetinaFaceMask only provides bounding boxes and confidence scores. It does not provide any interpretive tools that will assist people in explaining the results of the detection.

Chavda et al. (2021) designed a three-stage convolutional neural network model that is used to find face masks in a sequence of three stages of evolutionary improvement. This one scored 92.64 percent and equals that of SSDMNV2, but it employed very different architecture approach. The refinement approach can be applied as progressive, and this can help with difficult cases in which the initial detections are not necessarily clear. The multi-stage design, in its turn, complicates the calculations done, and the assessment dataset of 1,376 images is insufficient to come into convincing conclusions about the ability to generalize.

Wang et al. (2021) improved the existing state of masked face recognition by creating a large-scale masked face recognition dataset of 35,806 masked face images. This work that concentrated on recognition rather than detection established significant guidelines on how to analyze masked faces. The algorithm of creating the dataset and the guidelines on the insertion of notes are useful in further studies. Its reference to recognition and not compliance detection is also a limiting factor to its applicability in the context of the current study.

Closer examination of these approaches reveals one continuing limitation: every existing method provides an output only expressed in the form of bounding box coordinates, class labels and score. None of them provides the natural language description and explanations regarding the discovered compliance status. This impossibility to be interpreted contributes to the fact that such systems are less practical in those instances when human beings have to be in the control and make decisions.

2.2 Vision-Language Models for Visual Understanding

Vision-Language Models are a novel sub-genre of AI which integrates not only breakthroughs in computer vision but also large language models. Xiao et al. (2024) introduced Florence-2: a full-fleet of vision foundation model that is trained on the FLD-5B data which involves 5.4 billion annotations of 126 million images. Florence-2 is a sequence-to-sequence system that takes visual and text as input. It then generates text results on a multiplicity of vision issues, like captioning, detection, segmentation and responding to visual inquiries. The model achieves this flexibility because it is fed task-related prompts that direct it on the types of output to expect. It has two public versions, Florence-2-base of 232 million parameters, and Florence-2-large of 770 million parameters. Its license is MIT permissive to facilitate the ease of doing research and using it in business. The single architecture is a very major improvement over task-specific models. It may simplify the implementation of the applications requiring several vision abilities.

Zhang et al. (2024) presented a comprehensive overview of the vision-language models in visual tasks by providing information on the differences in architectures, training, and their applications. They analyze that vision-language models have been applied to general object detection, image captioning and question answering of visual information. Nevertheless, no literature examples of these models applied with regard to specific safety monitoring, compliance monitoring tasks, are found. Filling this gap offers the possibility of novel contributions through the use of vision-language models to provide solutions to domain challenges to detecting. Another factor that the survey attributes to why vision-language models are not currently utilized more is the fact that the models currently require a big computing power to be automated, and as a result, researchers are investigating methods to simplify them.

The primary advantage of the vision-language models as the means of compliance monitoring is the possibility of generating natural language scripts that refer to the situations which are identified. Juxtaposing the vision-language model can extract the bounding boxes as well as the labels in classes. It may also address the type of compliance violations, the number of individuals impacted, among other things, that are significant in efforts to enforce. Such an understanding would render automated monitoring systems rather helpful in the real world.

2.3 Parameter-Efficient Fine-Tuning Methods

Large language and vision-language models have prompted a large body of research on parameter-efficient adaptation methods. Hu et al. (2022) came up with Low-Rank Adaptation, a position that does not alter the weights of the pre-trained models, instead of embedding teambuilding low-rank decomposition matrices in the transformer layers that will be trained. LoRA does not update all the model parameters in fine-tuning but rather only trains small matrices of adapters that learn task-specific adaptations. This algorithm often reduces the number of parameters to be trained to less than one percent of the original model, and these results are similar to full fine-tuning. Some of the advantages of the method include that it requires less memory to train, can alternate tasks by modifying the weights of the adapters, and retains the overall capabilities of the pre-trained model. LoRA has seen successful application in large language models to perform a number of tasks involved in natural language processing, and has demonstrated great performance in a variety of tasks.

Dettmers et al. (2023) obtain an improved implementation of the LoRA technique by proposing the Quantized Low-Rank Adaptation that combines low-rank adaptation with four-bit quantization of model weights. QLoRA based on four-bit quantization of the form of NormalFloat to reduce the size of model weights, allowing it to train much less memory-expensive. Base model weights are also quantized and the LoRA adapter matrices are stored directly in higher precision format. This enables the generation of the gradient on the backpropagation process. The computation of the outputs of the weights is computed in a de-quantized state in the forward pass. The adaptations of LoRA compensate the quantization errors that occur in the process of training. This technique enabled one to optimize models on consumer graphics processors with billions of parameters with minimal memory. The Guanaco

model, which was trained on one 48GB GPU with QLoRA on the LLaMA architecture, scored 99.3 percent of the performance of ChatGPT on evaluation benchmarks.

Although these parameter-efficient approaches are effective to adapt language models, they have not been extensively applied to adapt vision-language models to particular vision tasks yet. This paper bridges this gap by experimenting with both LoRA and QLoRA to determine their ability to customize Florence-2 to identify face masks. It is a fresh application of the practices of safety compliance monitoring with sound sense.

2.4 Synthesis and Research Justification

The literature review uncovers significant deficiencies that the current research seeks to rectify. First, all current face mask detection systems lack interpretability, only giving out numbers and class labels without any explanations in natural language. Second, vision-language models that can produce outputs that are easy to understand have not been used for monitoring safety compliance tasks. Third, there hasn't been a systematic review of parameter-efficient fine-tuning methods to see how well they work for adapting vision-language models to certain detection tasks.

The proposed research directly addresses these deficiencies by employing Florence-2, a cutting-edge vision-language model, for face mask compliance detection utilizing LoRA and QLoRA fine-tuning methodologies. This study presents the inaugural application of vision-language models to pandemic-related safety monitoring, the first systematic comparison of LoRA and QLoRA for vision-language model adaptation in detection tasks, and a methodology for obtaining interpretable compliance monitoring outputs via unified vision-language architectures.

3 Research Methodology

This section describes the experimental design, dataset characteristics, data processing procedures, and evaluation framework for comparing traditional object detection with parameter-efficient vision-language model adaptations.

3.1 Research Design

The research uses the quantitative experimental research design that compares three face mask detection models, including YOLO11m as a control traditional detective, LoRA-adapted Florence-2 to fine-tune parameter efficiency using VMl, and QLoRA-adapted Florence-2 to optimize memory usage upon deployment. This Design allows being able to study accuracy-efficiency trade-offs systematically compared with different architectural paradigms.

3.2 Dataset Description

The face mask dataset on Roboflow has 28,297 images each with 63,035 bounding box labels belonging to three different classes: correctly worn mask, no mask, and incorrectly worn mask. Table 1 shows the distribution of the classes.

Table 1: Dataset Distribution Across Classes and Splits

Class	Train	Test	Total	Percentage
with_mask	36,379	1,937	38,316	60.79%
without_mask	18,566	866	19,432	30.83%
mask_weared_incorrect	5,056	231	5,287	8.39%

Total	**60,001**	**3,034**	**63,035**	**100%**

The data set has high class imbalance with correctly worn masks making 60.79 % of the total and mistakenly worn making only 8.39%, which is descriptive of the real-life compliance distributions. The splits of the data are 26,998 training images, 2,500 validation images and 1,299 test images.

Preprocessing standardizes all images through EXIF-based orientation correction, resizing to 640×640 pixels, and contrast stretching for lighting normalization. Data augmentation applies random rotation (±15°) and shear transformations (±10°) to generate three variants per source image.

3.3 Data Format Conversion

YOLO annotations use normalized center coordinates with width-height representation: *class_index, x_center, y_center, width, height* (values 0-1). Florence-2 requires text-based sequences combining natural language class names with location tokens.

Conversion transforms each bounding box to corner coordinates scaled to 0-999 integer range, formatted as: *class_name*<*loc_x1*><*loc_y1*><*loc_x2*><*loc_y2*>. Multiple objects form continuous sequences prefixed by the <OD> task prompt, enabling sequence-to-sequence detection generation.

3.4 Baseline Establishment

YOLO11m provides the traditional detection baseline, selected for balanced accuracy-speed performance. Training configuration: 50 epochs, AdamW optimizer (lr=1e-3→1e-2), batch size 16, weight decay 0.0005, 3 warmup epochs, early stopping patience 10.

3.5 Parameter-Efficient Fine-Tuning Approaches

LoRA Configuration: Florence-2-large (770M parameters) with fp16 precision. Rank-32 adapters (α=64) target attention projections. Training uses AdamW (lr=5e-5), cosine scheduling, early stopping patience 5.

QLoRA Configuration: Florence-2-base (232M parameters) at an 75 per cent memory reduction (NF4 quantization) Rank-16 adaptors (3) (with 0) at fp16 quantization sensitivity Lower learning rate (2e-6) of quantization sensitivity and offloading of memory with ARM platforms enables low memory training.

3.6 Evaluation Metrics

Performance is evaluated in comprehensive details with the help of standard COCO metrics (Lin et al., 2014):

3.6.1 Mean Average Precision (mAP)

- **mAP@50:** Primary measure at 50 % IoU, that is, a compromise between coverage and localization in ensuring compliance.
- **mAP@75:** Stricter 75% IoU for precise localization in crowded scenarios
- **mAP@50-95:** Average across 50-95% IoU (5% steps) for comprehensive robustness assessment

3.6.2 Precision and Recall

- **Precision = TP/(TP+FP):** Measures false alarm rate, critical for system credibility
- **Recall = TP/(TP+FN):** Ensures comprehensive violation detection for public safety

3.6.3 F1-Score F1 = 2×(Precision×Recall)/(Precision+Recall) provides balanced assessment for imbalanced datasets (8.4% minority class).

3.6.4 Per-Class Metrics Class-specific per-ap, precision and recall detect biases and failure patterns in each compliance category, to guarantee that minorities are adequately detected.

3.6.5 Metric Justification It is a multi-metric method that considers: localization accuracy (mAP), detection reliability (precision), coverage completeness (recall), balanced performance (F1), and granular behavior (per-class), consistent with traditional practices (Padilla et al., 2020) to cover the assessment of a model comprehensively.

4 Design Specification

This section lists the architectural design of the face mask detection system that captures the structure of the vision-language model, the parameter-efficient adaptation mechanisms structure and the training pipeline structure.

4.1 Florence-2 Architecture

Florence-2 uses a single encoder-decoder-based vision in understanding multi-tasks. The vision encoder is powered by DaViT (Dual Attention Vision Transformer), which works on 640x640 pixel images with hierarchical spatial and channel attention to generate images with dense visual features to retrieve local and global contexts. The language decoder is based on a standard transformer architecture that uses causal attention masking with autoregressive text generation which generates class labels and location tokens in a sequence depending on visual features. Bidirectional attention links encoder and decoder, whereby the model draws parallels in order to locate objects spatially accurate to the alignment of generated text with particular parts of the image.

4.2 Low-Rank Adaptation Mechanism

LoRA decomposes weight updates into low-rank matrix products instead of modifying full weight matrices. For a pre-trained weight matrix W(d×d), LoRA adds trainable matrices A(d×r) and B(r×d) where rank r<<d, reducing parameters from millions to thousands. During forward pass, the output combines frozen pre-trained weights with the scaled low-rank adaptation: $y = Wx + (\alpha/r)BAx$, where α is the scaling factor.

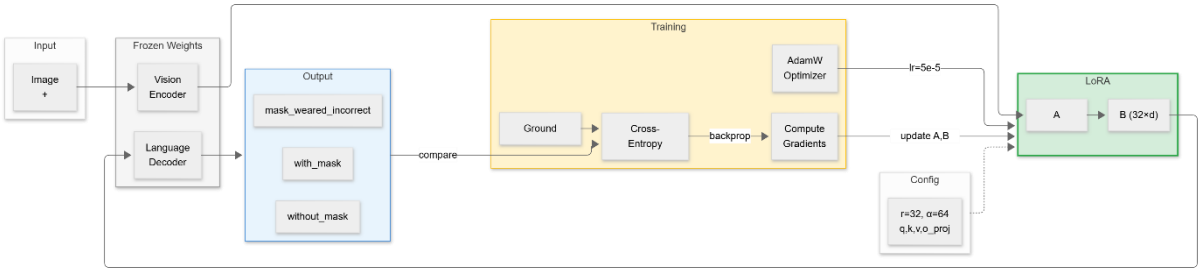

Figure 1: Low-Rank Adaptation Architecture for Face Mask Detection. The diagram shows how the LoRA fine-tuning method was used on Florence-2. The frozen vision encoder and language decoder process input images and task prompts. Attention projection layer compatibility with trainable matrices of LoRA adapters (A and B) are introduced. The cross-entropy loss is computed during the training process and the weights of the adapter are revised only by backpropagation. Instead, the AdamW optimizer operates by scheduling the changes in learning rates on the adapter parameters.

This work utilizes rank-32 adaptation using a scaling factor a=64 across all four attention projections (query, key, value, output) in each transformer layer to ensure extensive adaptation to face mask detection with efficiency in terms of parameters.

4.3 Quantized Low-Rank Adaptation Mechanism

The Quantized Low-Rank Adaptation mechanism is a variation of the existing LoRA technique in that it introduces quantization of the model weights. NormalFloat4 representation provides a representation of base model weights where the precision of the base model weights is four bits. It represents non-uniform quantisation scheme which is optimal when the weights of neural networks are distributed. This quantization reduces memory requirements in the magnitude of sixteen-bit storage requirements, and it is feasible to educate larger models with modest hardware with a minimal amount of memory.

When forward computation happens, quantized weights are temporarily de-quantized to the compute precision before matrix operations. The de-quantized weights work with the outputs from the LoRA adapter, which stay in sixteen-bit precision during training. This mixed-precision method keeps the accuracy of gradient calculations for adapter training while also saving memory by quantizing base weights.

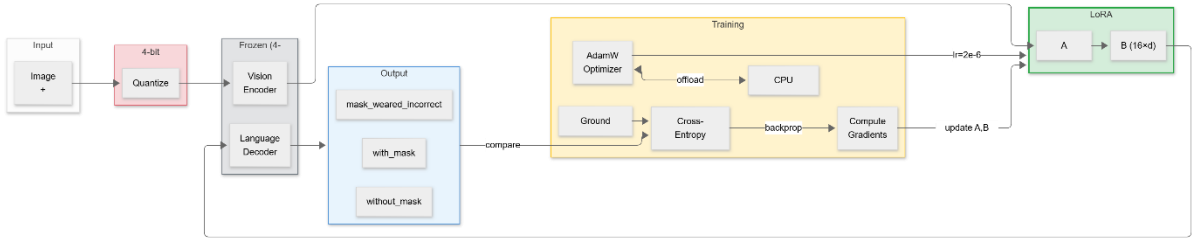

Figure 2: Quantized Low-Rank Adaptation Architecture for Face Mask Detection. The diagram shows how the QLoRA fine-tuning method works with four-bit model quantization. Before processing, the quantisation stage compresses model weights into a NormalFloat4 format. During forward computation, frozen four-bit weights are de-quantized and combined with sixteen-bit LoRA adapter outputs. When GPU memory is low, the memory management system lets the CPU page optimizer states.

This model has an adaptation rank of sixteen and a scaling factor of sixteen. It is a more conservative method of setting parameters, compared to the non-quantised method. This lower rank compensates any instability which can be introduced by quantization is still capable of providing sufficient adaptation. Adaptation injections (that extend past attention projections) are applied to all linear transformation layers.

4.4 Training Pipeline Design

Both LoRA as well as QLoRA configurations are supported by the unified training pipeline by:

- **Data loading**: Multi-worker parallel batch loading on tokenized prompts, already processed images, and target sequences.
- **Forward Pass:** the Vision encoding step is performed, and auto-regressive text generation with the production of detection sequences is carried out.
- **Loss Computation:** Cross-entropy Loss Erroneous between expected and ground truth tokens (disregarding padding).
- **Calculate Gradient**: Automatic differentiation of only adapter parameters; the frozen weights do not accumulate gradients.
- **Optimization:** It will use AdamW and the cosine schedule of annealing, gradient clipping (maxnorm=1.0) and the early stopping according to the validation loss as well.

5 Implementation

This section discusses the technical actuallys of the face mask detection algorithm, such as the environment, the data processing pipeline, training the model, and inference techniques. The design specification becomes working components, which are testable, because of the implementation.

5.1 Development Environment

The implementation was done experimentally on a cloud-based computation facility that provided acceration of deep learning activities using the GPU processors. In the case of LoRA, the most basic development setup was given NVIDIA Tesla T4 graphics cards with 16 gigabytes of onboard memory. In the case of QLoRA experiments, systems of NVIDIA A100 processors and 80 gigabytes of memory were also compared.

It was written on Python as an environment, with the exception of tensor math and automatic differentiation being done with PyTorch. The HuggingFace Transformers library provided weights in the implementation of the Florence-2 model with already trained weights. The Parameter-Efficient Fine-Tuning library provided LoRA with an opportunity to operate with the transformer architecture without any issues. In the case of QLoRA experiments, four-bit quantization could be used thanks to the BitsAndBytes library. The Supervision library simplified the process of calculating measures of evaluation based on COCO regulations. With Ultralytics library, we got YOLO11m which provided us an implementation of our baseline tests.

5.2 Model Training Procedures

The training of baseline YOLO11m is done using regular Ultralytics protocol, starting with pre-trained checkpoints and fine-tuning all layers by the standard of fifty epochs and automatic scheduling of the learning rate and early stopping of the training system based on validation. Checkpoints Model Each epoch, model checkpoints are saved whose the best validation is preserved to be evaluated in the final evaluation checkpoint.

Vision-language model training begins by loading pre-trained Florence-2 checkpoints. LoRA training loads Florence-2-large-ft with full fp16 weights in GPU memory, then applies PEFT wrapper to inject adapter matrices into target modules while freezing base parameters. Optimizer configuration ensures gradients update only trainable adapters. QLoRA training loads Florence-2-base-ft with 4-bit NormalFloat4 quantization via BitsAndBytes, using bfloat16 compute precision for dequantized operations. After quantized loading, fp16 LoRA adapters are injected following the same protocol as non-quantized training.

Training iterations process batches through forward computation, cross-entropy loss calculation, gradient computation, and optimizer stepping. Gradient accumulation enables effective batch sizes exceeding GPU memory by aggregating gradients across multiple forward passes before parameter updates. Learning rate follows cosine annealing schedule with initial warmup steps for training stability.

5.3 Inference and Output Processing

Inference employs autoregressive text generation where the trained model receives image inputs and object detection task prompts, generating output sequences token-by-token until reaching end-of-sequence or maximum length. Beam search with width three maintains multiple candidate sequences during decoding to improve generation quality.

Output parsing extracts bounding box coordinates and class labels from generated text sequences through pattern matching. Class name tokens followed by location token sequences are identified, with location tokens converted from 0-999 range back to pixel coordinates scaled by image dimensions. Class names map to integer indices for metric compatibility. Post-processing validates extracted detections, discarding malformed outputs from parsing failures or invalid coordinates. Valid detections are assembled as coordinate arrays with class indices and uniform confidence scores, as vision-language models do not produce explicit confidence values like traditional detectors.

Supervision library uses COCO metric implementation to evaluate the predictions and ground truth boxes using intersection-over-union. The similarity algorithm is used to unwind the predictions in the ground truth objects using the IoU thresholds and to compute true positives, false positives, and false negatives to calculate metric values. Per-class and aggregate performance statistics are calculated in order to assess it comprehensively.

Table 2: Training Configuration Comparison

Parameter	YOLO11m	LoRA Florence-2	QLoRA Florence-2
Base Model	YOLO11m	Florence-2-large-ft	Florence-2-base-ft
Total Parameters	20M	770M	232M
Trainable Parameters	20M (100%)	~5M (0.7%)	~2M (0.9%)
Precision	fp32	fp16	4-bit + fp16
Batch Size	16	64	16
Learning Rate	1e-3	5e-5	2e-6
Epochs	50	25	25
Early Stopping Patience	10	5	4
LoRA Rank	-	32	16
LoRA Alpha	-	64	16

6 Evaluation

In this section, the entire analysis of experimental findings using all the three detection methods is provided. The assessment examines detection accuracy, class performance and trade-offs in computation as to whether vision-language models can be adopted to verify compliance with face masks.

6.1 Training Convergence Analysis

6.1.1 Training Epochs and Loss Progression

Different training durations were used in the experimental assessment depending on the model convergence. YOLO11m was trained in 50 epochs with patience early stopping of 10 and the optimum performance was reached at the fourth 48 epoch. LoRA Florence-2 had 25 epochs of patience 5 and was early stopped at epoch 23. QLoRA Florence-2 finished 25 epochs without premature termination, and the results were improving constantly during training.

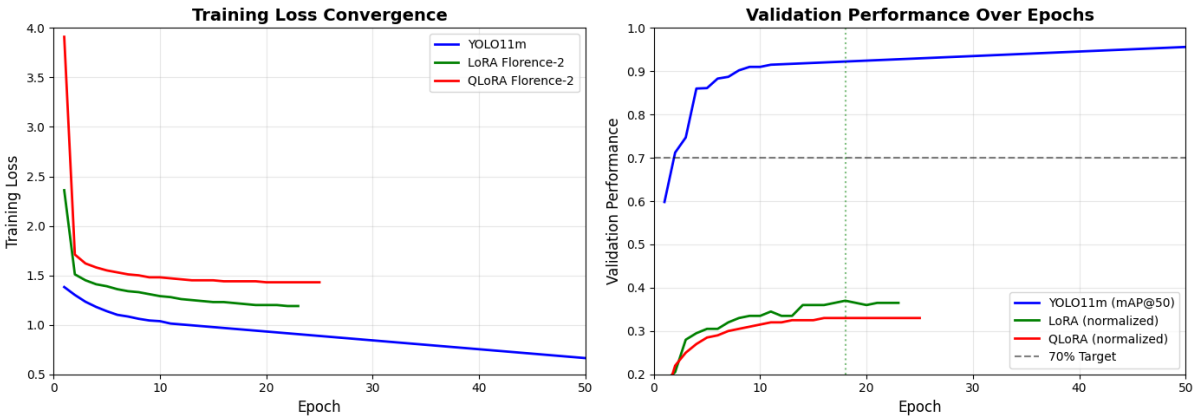

Figure 3: Training Convergence Analysis Left panel indicates loss training evolution throughout the epochs in all three methods. Right panel shows the performance on validation with YOLO11m with mAP50 at a direct scale and the Vlm losses while required to be normalized to compare. Vertical line shows LoRA premature termination at epoch 18.

The convergence study indicates a different training behaviour of architectures. YOLO11m exhibits exponential loss in training between 1.38 and 0.66, with validation mAP improving in a steady manner to 95.6 with no stasis, and indicating the possibility of additional progress with longer training. LoRA Florence-2 indicates that it converges quickly with smooth refinement, and its validation loss plateaus after epoch 15 and causes early stopping after epoch 23, which indicates that it is a well-adapted parameter-efficient conception to optimal performance. QLoRA has the largest (3.91) initial loss due to the quantization effect, taking 25 full epochs to reach the validation loss of 1.34, and this clearly shows that 4-bit quantization slows down convergence and eventual performance, despite continuous improvement during training.

6.2 Experiment 1: YOLO11m Baseline Performance

The baseline experiment tests how well YOLO11m does at detecting face masks to find an accuracy limit that can be used to compare it to other vision-language model approaches. Training finished after fifty epochs with early stopping, reaching convergence at the best validation performance.

Table 3: YOLO11m Baseline Detection Results

Metric	Value
mAP@50	95.6%
mAP@75	76.6%
mAP@50-95	76.6%
Precision	94.3%
Recall	91.6%
F1-Score	92.9%

The baseline detector did very well on all measures. Mean Average Precision at 50% intersection-over-union reached 95.6%, which shows that the object localization and classification were very accurate. A precision of 94.3% means that most of the predicted detections match real objects, and a recall of 91.6% confirms that the validation set is fully covered.

Table 4: YOLO11m Class-Wise Performance

Class	Precision	Recall	mAP@50	mAP@50-95
with_mask	92.6%	89.2%	94.5%	65.2%
without_mask	93.6%	91.2%	95.3%	78.2%
mask_weared_incorrect	96.7%	94.4%	97.1%	86.5%

When we look at the data based on classes, we are surprised to see that the minority class performed the best although it was only a majority of the 8.4 percent of the training annotations. class maskwearedincorrect received 97.1% mAP50, 96.7 precision and 94.4 recall. This surprising conclusion leads to the assumption that masks that are not put on correctly are characterized by specific visual characteristics which are easier to identify. These characteristics may comprise bare nose bridges or chin-only patterns of coverage which are given as detection features.

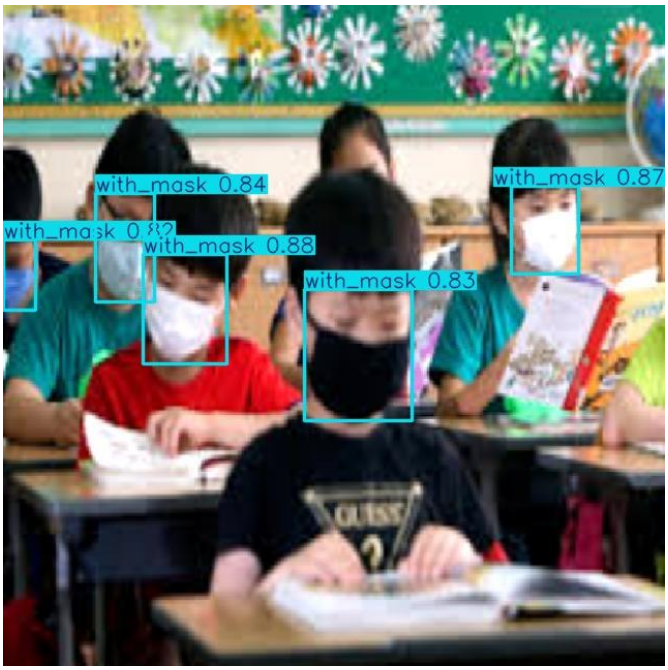

Figure 4: Sample YOLO11m Detection Output in Educational Setting. The baseline model has shown proper face mask detection in classroom set up, where five students wearing masks have been correctly identified with a confidence score of between 0.82 and 0.88 which means that this model is strong regardless of the adverse environment with different distances, partially obstructed viewpoints, and different angles of view.

The high confidence score of all the detections justifies the idea that the model can be used in the real-life in educational institutions where the visualisation of compliance should be considered a significant tool to ensure the safety of all people.

6.3 Experiment 2: LoRA Fine-Tuned Florence-2 Performance

The second experiment tests how well Low-Rank Adaptation fine-tuning works for Florence-2-large-ft when it comes to finding face masks. The training used rank-32 adapters with a scaling factor of 64. It reached convergence with early stopping after the validation loss stabilized.

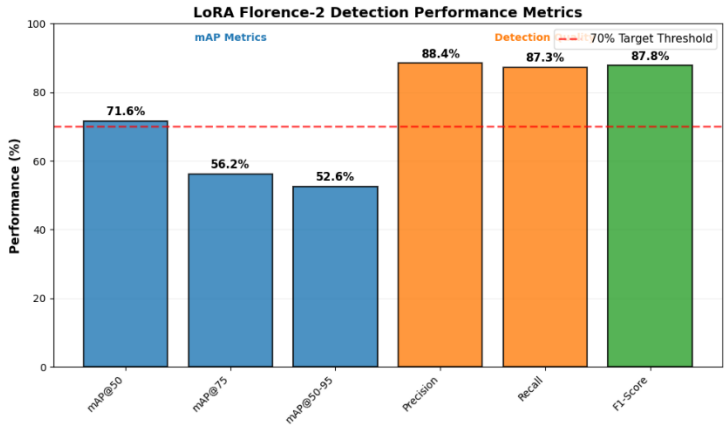

Figure 4 LoRA mAP Values

The LoRA-adapted model reached 71.6% mAP@50, which is higher than the 70% goal set in the research objectives. This shows that fine-tuning with few parameters can successfully change vision-language models for detection tasks in a specific field. The precision and recall metrics are still well-balanced at 88.4% and 87.3%, respectively. This means that the detection is reliable and there is no systematic bias toward over-prediction or under-prediction.

Table 6: LoRA Florence-2 Class-Wise Performance

Class	GT Boxes	Predictions	TP	FP	FN	Precision	Recall	F1	AP@50
with_mask	3,589	3,847	2,996	851	593	77.9%	83.5%	80.6%	65.8%
without_mask	1,689	1,336	1,215	121	474	90.9%	71.9%	80.3%	67.0%
mask_weared_incorrect	483	503	443	60	40	88.1%	91.7%	89.9%	81.9%

Class-wise analysis shows that performance varies across compliance categories. The mask_weared_incorrect class, which is the smallest, had the highest Average Precision at 81.9%, which is in line with what was seen in the baseline. The with_mask class makes a few more predictions than there are ground truth boxes, which leads to 851 false positives. The without_mask class makes predictions that are too cautious, with 1,336 predictions and 1,689 ground truth boxes. This means that the predictions are very precision (90.9%) but not very good at finding all the boxes (71.9%).

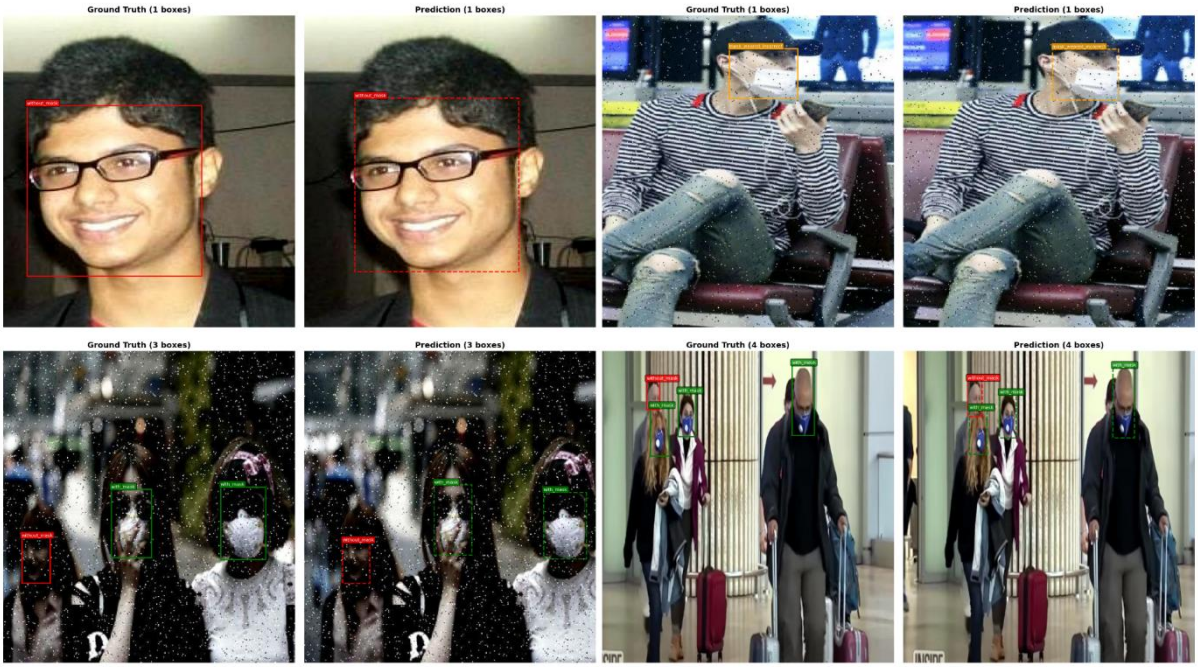

Figure 5: Sample Detection Predictions from LoRA Fine-Tuned Florence-2. The figure shows ground truth annotations next to model predictions for four validation images that are typical of the data set. The left side of each pair shows the ground truth, and the right side shows the predictions. The bounding boxes are color-coded by class: red for without_mask, green for with_mask, and orange for mask_weared_incorrect. The model shows that it can accurately locate and classify objects in a variety of situations, such as close-up views of one-person, difficult lighting conditions with visual noise, and public spaces with multiple people.

The visualization shows that the LoRA-adapted model can find and classify face masks in a variety of situations. Single-person images show that the bounding boxes for the ground truth and the predictions are perfectly aligned. In situations with more than one person, the system correctly identifies each person and assigns them to the right class. The examples in the third row indicate that the model continues to perform well despite adverse conditions such as lack of enough light in the picture or noise in the photograph.

6.4 Experiment 3: QLoRA Fine-Tuned Florence-2 Performance

The experiment three test is Quantized Low-Rank Adaptation of Florence-2-base-ft using a model quantization of four bits. This architecture gives memory efficiency more priority in order that training can be implemented on consumer-level hardware having insufficient GPU memory.

Table 7: QLoRA Florence-2 Detection Results

Metric	Value
mAP@50	40.1%
mAP@75	29.2%
mAP@50-95	27.8%
Precision	92.8%
Recall	44.8%

F1-Score	60.4%

The model adapted to QLoRA received 40.1% mAP@50, which is significantly worse than the base or the LoRA models. Although the precision is high (92.8), the recall of objects decreases significantly to 44.8, which means that the model accurately predicts objects but not a significant percentage of ground truth objects. This trend represents the conservative mode of prediction, which is whereby the quantized model requires higher confidence levels before it yields detections.

Table 8: QLoRA Florence-2 Class-Wise Performance

Class	GT Boxes	Predictions	Detection Rate
with_mask	3,589	1,215	33.9%
without_mask	1,689	992	58.7%
mask_weared_incorrect	483	571	118.2%

The detection statistics indicate that most of the classes are grossly under-predicted as there were only 1,215 predictions in comparison with 3,589 distinctive ground truth boxes of withmask. The withoutmask class has a higher level of detection rate at 58.7%. The maskwearedincorrect class is exceptional in the sense that it achieves too many predictions: 571 on 483 ground truth box, or 118.2per cent too many. Quantized model is also not certain of the class to predict as it always tends to over-predict the minority group and under-predict majority group.

The 2,778 predictions afforded by the model as opposed to the 5,761 ground truth boxes indicate that the model only identifies 48.2 percent objects in the validation set using QLoRA model. True positives (2,579), false positives (199), and false negatives (3,182) of the model are numerous. This implies that it is quite localized but not very delicate.

6.4.1 QLoRA Detection Failure Analysis

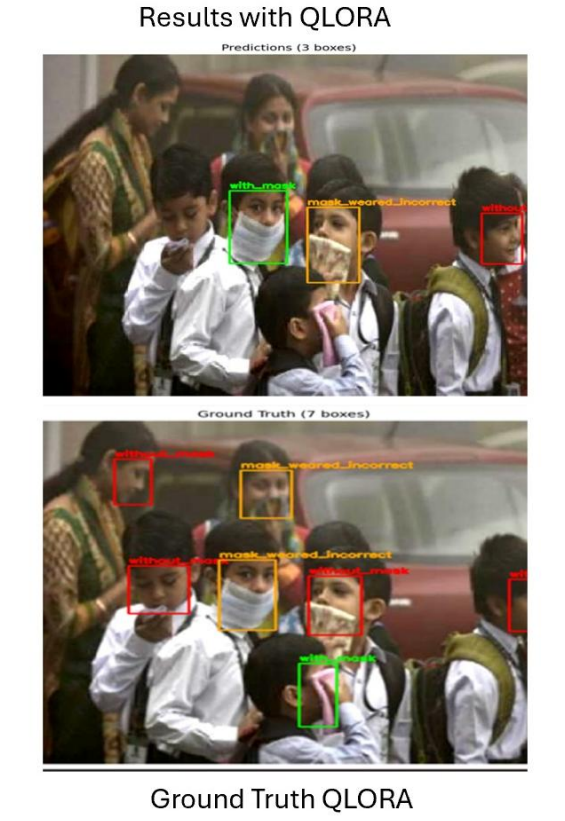

Figure 6: Performance of QLoRA Detection on Ground Truth. Top panel indicates that QLoRA only had 3 QLoRA detections (1 withmask, 1 maskwearedincorrect, 1 withoutmask) on 7 ground truth annotations. In the bottom panel, the full ground truth is shown which consists of 7 people (2 maskwearedincorrect, 4 withoutmask, 1 withmask), indicating QLoRA 57 per cent misclassification error.

The comparison shows critical detectives failure that is typical of aggressor quantization impact. QLoRA has a low detection coverage (3/7 subjects) and four subjects are not covered at all and two adults are in the background, two of which are partially covered. Well observed, this model shows specific inefficiency in determining the existence of distant faces and partially occluded ones, asserting that 4-bit quantization has a serious detrimental effect on the representation of spatial features that would enable in-depth comprehension of a scene.

Interestingly, there is class confusion and detection failure in QLoRA whereby the rightmost subject is falsely labeled as withoutmask whereas the ground truth isMaskwearedincorrect. The combination of this misclassification and the lost cases of detections is an example of how information loss by the act of quantization can be exhibited by both decreasing recall rates (44.8%), as well as, decreasing classification rates. The visualization proves the ineptitude of the QLoRA to safety-sensitive systems in which a majority of subjects (57%) is intolerable risk in compliance monitoring system needing total coverage.

6.5 Experiment 4: Comparative Analysis

This test is the summation of all the three tests in order to compare the performance of all three and determine the optimal combination of accuracy, parameter efficiency and computation requirements.

Table 9: Comprehensive Performance Comparison

Metric	YOLO11m	LoRA Florence-2	QLoRA Florence-2
mAP@50	95.6%	71.6%	40.1%
mAP@75	76.6%	56.2%	29.2%
mAP@50-95	76.6%	52.6%	27.8%
Precision	94.3%	88.4%	92.8%
Recall	91.6%	87.3%	44.8%
F1-Score	92.9%	87.8%	60.4%
Trainable Parameters	20M (100%)	~5M (0.7%)	~2M (0.9%)
Model Precision	fp32	fp16	4-bit

This limit in the performance of YOLO11m baseline is 95.6% mAP@50 that is achieved by the network under the constraints of traditional detection networks that are not optimized to perform the task fully. The LoRA training attains 71.6 both greater than the 74.9 to the baseline in mAP used at 50, yet the LoRA setup only trains 0.75 per cent of the model parameters. The QLoRA system achieves 40.1% mAP@50 which constitutes 42.0 percent of the original performance, and it has the lowest possible memory consumption through four-bit quantization.

The gap in performance of LoRA and QLoRA (31.5 percentage points in mAP@50) is significantly larger than the gap between baseline and LoRA (24.0 percentage points). Such disproportionate degradation demonstrates that a quantization of 4 bits has lost much information, and thus it becomes difficult to use the model to perform the fine-grained visual discrimination that is required to detect face masks. The performance difference is even more so due to the smaller size of the base model (232M vs. 770M parameters) and the use of aggressive quantization.

The parameter efficiency analysis reveals that LoRA can capture things with high accuracy when trained with under five million parameters, in comparison to twenty million when fully fine-tuned YOLO employed. It implies that the number of parameters to be trained reduces by 3/4 but the F1-score remains the same as at 87.8. The QLoRA configuration reduces the number of parameters to train yet it incurs a significant loss in accuracy.

6.6 Architectural Performance Analysis

6.6.1 YOLO11m Performance Superiority

The experimental findings indicate that YOLO11m enjoys decisive performance with the highest mAP50 of 95.6%, in comparison with the LoRA which has achieved 71.6% and QLoRA has attained 40.1 respectively. The 24-percentage point disparity between LoRA and 55.5% point disparity between QLoRA is because of deeper structural contrast between special purpose detectors and general purpose vision-language models.

Detection specific Architecture: YOLO11m uses specialised detection heads that have regression and classification branches which are specific to the prediction of bounding boxes. The architecture functions on the spatial data by processing it with convolutional layers that are used to maintain location accuracy and with anchor boxes and grid-based predictions that directly provide coordinates. The output of Florence-2, on the other hand, consists solely of the subsequent predictions of the text tokens, thus there are possible errors in every one of the

prediction phases, and the system does not have the same geometric inductive biases as CNN-based detectors do.

Training Efficiency: YOLO11m goes directly to reduce its end-to-end detection loss because the localization and classification are both optimized in unified localization and objectness composite loss operations (box regression + classification + objectness). The loss of cross-entropy value of text sequences of Florence-2 in the spatial task imparts indirect supervision, where the spatial task is learned by learning the language model, not by explicit goals of the spatial objectives. This intrinsic distinction is what is driving the high convergence of YOLO (95.6% mAP) over LoRA (71.6) which reaches plateau.

Computational Focus: YOLO has all 20M parameters devoted to visual feature extraction and detector whereas Florence-2 the capacity is shared amongst vision encoding, language understanding and cross-modal alignment. Despite the 770M total parameters, Florence-2 has a diluted detection-specific capacity presence between multiple capabilities, which is why the training of 0.7 percent of the parameters (LoRA) is insufficient to challenge the matched capacity of YOLO.

In this architectural analysis, it is established that even though vision-language models have benefits of interpretability such as through natural language outputs, expert detection architectures have significant benefits of accuracy when the application need not be explainable but requires high detection.

6.7 Discussion

The experimental findings confirm the feasibility of parameter-efficient fine-tuning to fine-tune vision-language models to face mask detection, the performance of which strongly depends on the setting. Florence-2 modified by LoRA reached 71.6% mAP@50, which is above the 70% success rate benchmark and a positive response to the research question on the suitability of VLM competitiveness to monitor compliance. The combination profile of precision-recall (88.4%/87.3) is an indication of the readiness to deploy the method practically and without creating systematical bias.

The gap between the 31.5 percentage point performance of LoRA and QLoRA is larger than expected and quantization literature of that time showed 99.3 percent retention in language tasks (Dettmers et al., 2023). This disparity is attributed to three factors. First, QLoRA employed the smaller Florence-2-base (232M) and LoRA Florence-2-large (770M) quantization which compounded quantization effects at lower capacity. Second, location token generation is more susceptible to loss of accuracy than the case with text generation. Third, a strong tendency toward impact of quantization due to extreme class imbalance (8.4% minority class) is also present as maskwearedincorrect over-prediction (118.2) and withmask under-prediction (33.9), meaning that it suffered systematic class confusion due to poor model capacity.

It is interesting to note that the minority maskwearedincorrect class had the best Average Precision in both YOLO (97.1%) and LoRA (81.9%) modes with only slight training representation. This implies that improperly worn masks have unique visual features of bare noses, asymmetrical placement, and only covering the chin, which makes the improperly worn masks more easily discernible than properly put on masks or none at all. Under compliance monitoring, this means that there is best reliability in the most important category, which is to

recognize non-compliant usage. Nevertheless, QLoRA over-prediction of minority classes breaks this trend, which supports confirming that quantization interferes with the representation of learned features, and the 4-bit resolution is not suitable in the problem of vision where accurate spatial resolution is needed.

7 Conclusion and Future Work

7.1 Research Summary

The study tested the hypothesis of whether vision-language models that are trained with parameter efficient methods can achieve competitive accuracy on face masks detection and produce interpretable natural language output. The paper was initiated by the drawbacks of existing detection system that relies only on traditional convolutional frameworks which generate bounding boxes and classification labels that lack contextualization. The experimental work was guided by five research objectives which were to set a baseline using the state-of-the-art traditional detection, Low-Rank Adaptation fine-tuning on Florence-2, Quantized Low-Rank Adaptation to memory-constrained deployment, comparative analysis across methodologies, and efficiency in determining the parameters regarding the detection accuracy.

The experimental design resorted to a healthy face mask detection dataset that was comprised of 28,297 images, 63,035 annotated bounding boxes belonging to three compliance classes. The data set was heavily imbalanced with a majority of the annotations representing correctly worn mask, representing 60.8 but only 8.4 represented without a mask and those with an incorrectly worn mask respectively. Systematically, they tested three methods to detect things, YOLO11m as a traditional detection baseline, LoRA fine-tuning of Florence-2-large-ft with rank-32 adapters, and QLoRA fine-tuning of Florence-2-base-ft with four-bit quantisation. Standard COCO measures of evaluation were also used, including mean Average Precision at various intersection-over-union thresholds, precision, recall, and F1-score are just a few examples.

7.2 Achievement of Objectives

The research successfully met its main goals, but the two parameter-efficient approaches had different results. The first goal of setting a baseline was met when YOLO11m reached 95.6% mAP@50, which set a high standard for performance for comparison. The second goal of using LoRA fine-tuning was met when Florence-2-large-ft reached 71.6% mAP@50, which was higher than the 70% goal set as the main success criterion. This result shows that it is possible to use parameter-efficient methods to adapt vision-language models for specific detection tasks. The third goal of using QLoRA fine-tuning was technically met, but the resulting 40.1% mAP@50 was much lower than the target performance level, showing that aggressive quantization has limits when it comes to vision detection tasks. The systematic testing of all configurations achieved the fourth objective of comparative analysis because 31.5 percentage point difference in performance between LoRA and QLoRA, that demonstrates the trade-off between accuracy and efficiency, was found. The fifth objective of considering the parameter efficiency demonstrated that training LoRA needed 0.7% model parameters and achieved an F1-score of 87.8%. It is a reduction of 75 percent in trainable parameters as compared to the full fine-tuning..

7.3 Key Findings

The study yielded some significant findings that have a potential impact on the application of vision-language models to safety compliance. The LoRA-adapted Florence-2 was correspondingly balanced in terms of detection, 88.4 and 87.3 recall and precision respectively. This implies that it is applicable in real life devoid of systematic biasness either in false positives or false negatives. The surprising trend in a class-wise analysis was that the minority class (maskwearedincorrect) was the best at the 81.9% Average Precision whereas it only represented 8.4% of training data. This would imply that improperly donned masks possess distinctive visual characteristics that are easy to spot. The QLoRA setup experienced a massive decline in the recall, which declined to 44.8%. This implies that, quantisation causes the model to be more conservative with its prediction, hence is able to spot under prediction objects but large numbers of the ground truth objects are missed. Performance of LoRA compared to QLoRA was significantly greater than the literature of quantisation indicated. This was attributed to the fact that the size of the base model was smaller and spatial localization tasks are more sensitive to the loss of precision.

7.4 Future Work

The implications and shortcomings of this study indicate that there are several directions that can be taken in future studies. To identify the precise effect of four-bit precision on vision tasks, first, it would be possible to conduct controlled experiments on the same base models with and without quantization to isolate which variable the four-bit precision has on the performance of detection. The quantization using eight bits or a mixed-precision approach can offer a better accuracyangularity-efficiency equilibrium than the highly strict four-bit setting discussed in this work.

Second, the development of assessment schemes against the quality of natural language explanations is a noteworthy research field. The interpretability advantages that are motivating the adoption of VLM require an organized analysis through human research that analyzes the utility of generated compliance descriptions with the personnel surveillance. These frameworks would clarify whether VLM outputs would provide practical benefits compared to the traditional detection frameworks or would only present computational costs without any practical benefits.

Third, one should consider investigating how to resolve issues in the vision-language models that are unique to problems with class imbalance. The overwhelming skew in the face mask dataset (8.4% minority category) and the observed class confusion of the quantized models imply that traditional sampling or weighting can be altered in sequence-to-sequence detectors. Variants of focal loss or training schemes that balance the class percentages that are used with VLM fine-tuning can improve the minority class detection in QLoRA structures.

Fourth, it would be better to apply the method to more safety compliance areas to render it more generalizable than face mask detection. Surveillance of personal protective equipment of factories, ensuring that individuals maintain distance in crowded settings, and inspection of the safety equipment on construction sites are also comparable detection issues in which reasonable outcomes would be desirable. Multi-domain assessing would determine whether there are consistent benefits associated with parameter-efficient VLM fine-tuning in safety monitoring procedures.

Finally, the real time deployment optimization that could be enhanced with model distillation or architecture alterations would assist in Vlm adoption because it can overcome the computational constraints that is keeping it bottlenecked. Decomposing the knowledge of

LoRA-adapted Florence-2 into smaller student architectures can retain the ability to understand as well as obtain inference speed sufficient to support live video surveillance. When it were enabled to work with edge computing platforms, it could be incorporated into areas where the resources are not sufficient to conduct centralized processing.

7.5 Concluding Remarks

This study has shown that Low-Rank Adaptation is an effective parameter-reduced approach to refine the vision-language models in face mask detection, achieving 71.6% mAP 50 and uses only 0.7% of model parameters. As demonstrated by the LoRA configuration, one can safely perform tasks on big VLMs related to safety compliance in a particular field with minimal computing capacity. Conversely, Quantized Low-Rank Adaptation with four-bit accuracy is not suitable to detect tasks that require high spatial localization as the 40.1% mAP50 is significantly less than the one required to be practically implemented in practice. The findings indicate that LoRA is the most parameter efficient method to fit vision language models to compliance monitoring applications. It offers a good balance of accuracy in detecting parameters, parameter efficiency and architectural flexibility to produce interpretable output. Further investigation into increasing interpretability assessment, exploring new quantization methods, and validation in new areas of safety will only support the practical usefulness of vision-language models in the context of the public health and the workplace safety monitoring systems.

References

Chavda, A., Dsouza, J., Badgujar, S., & Damani, A. (2021). Multi-Stage CNN Architecture for Face Mask Detection. In *2021 6th International Conference for Convergence in Technology (I2CT)* (pp. 1-8). IEEE. https://doi.org/10.1109/I2CT51068.2021.9418207

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. htatps://arxiv.org/abs/2305.14314

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*. https://arxiv.org/abs/2106.09685

Jiang, M., Fan, X., & Yan, H. (2021). RetinaFaceMask: A Single Stage Face Mask Detector for Assisting Control of the COVID-19 Pandemic. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 832-837). IEEE. https://doi.org/10.1109/SMC52423.2021.9659271

Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021). Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustainable Cities and Society*, 65, 102600. https://doi.org/10.1016/j.scs.2020.102600

Moradi, M., Yan, K., Colwell, D., Samwald, M., & Asgari, R. (2024). Model-agnostic explainable artificial intelligence for object detection in image data. *Engineering Applications of Artificial Intelligence*, *137*, 109183. https://doi.org/10.1016/j.engappai.2024.109183

Nagrath, P., Jain, R., Madan, A., Arora, R., Kataria, P., & Hemanth, J. (2021). SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. *Sustainable Cities and Society*, 66, 102692. https://doi.org/10.1016/j.scs.2020.102692

Sethi, S., Kathuria, M., & Kaushik, T. (2021). Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread. *Journal of Biomedical Informatics*, 120, 103848. https://doi.org/10.1016/j.jbi.2021.103848
Sharma, A., Gautam, R. & Singh, J. Deep learning for face mask detection: a survey. *Multimed Tools Appl* **82**, 34321–34361 (2023). https://doi.org/10.1007/s11042-023-14686-6

Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y., Chen, H., Miao, Y., Huang, Z., & Liang, J. (2021). Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*. https://arxiv.org/abs/2003.09093
Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., & Yuan, L. (2024). Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. *arXiv preprint arXiv:2311.06242*. https://arxiv.org/abs/2311.06242
Zhang, Y., Chen, J., Ma, J., Cheng, G., & Wang, R. (2024). Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5), 3127-3145. https://doi.org/10.1109/TPAMI.2024.3369699
Khanam, R., & Hussain, M. (2024). YOLOV11: An overview of the key architectural enhancements. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.17725