engineering proceedings

MDPI

Proceeding Paper

Applying Multimodal Large Language Models in Factory Monitoring Platforms †

Chien-Yi Huang *, Chia-Yu Tsai, Yu-Chian Wei and Wei-Yang Lyu

Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei 10608, Taiwan; t111370201@ntut.org.tw (C.-Y.T.); t111370215@ntut.org.tw (Y.-C.W.); lwy1878@gmail.com (W.-Y.L.)

* Correspondence: jayhuang@mail.ntut.edu.tw

† Presented at the 9th Eurasian Conference on Educational Innovation 2026 (ECEI 2026), Da Nang City, Vietnam, 30 January–2 February 2026.

Abstract

To mitigate the difficulty of real-time PPE compliance monitoring in electronic factories, we developed an intelligent platform combining You Only Look Once version 10 (YOLOv10) real-time detection with Large Language Model Meta AI (Llama) 3.2 Vision multimodal large language model. Using real and simulated factory footage annotated via Roboflow, YOLOv10 detects violations, and Llama 3.2 vision generates natural-language descriptions. Results showed a 94.8% accuracy, a 94.81% recall, and a 91.82% F1-score, with robust performance under complex lighting/occlusion and clear semantic reports. The platform significantly improves factory safety and management efficiency, offering strong practical value.

Keywords: MLLMs; Llama 3.2 vision; YOLOv10; monitoring platform

1. Introduction

In the Industry 4.0 era, strict cleanroom personal protective equipment (PPE) compliance is critical for product yield and worker safety in electronics manufacturing. Traditional manual inspection is inefficient and non-real-time. Although computer vision has been widely adopted, most existing systems rely on single-modal object detection, lacking semantic interpretation of violations. We developed an intelligent monitoring platform integrating You Only Look Once version 10 (YOLOv10) real-time detection with Large Language Model Meta AI (Llama) 3.2 Vision multimodal large language model (MML), enabling instant violation identification and automatic natural-language report generation to significantly reduce managerial workload and enhance decision-making efficiency.

2. Materials and Methods

2.1. Research Procedure

This study was conducted in five sequential phases. First, the problem was defined, focusing on PPE non-compliance in cleanrooms (e.g., missing cleanroom suits, anti-static garments, masks, or gloves). Violation categories were established through on-site observation and analysis of factory regulations.

Data were collected from two sources: (1) real factory surveillance footage obtained through collaboration with a domestic electronics manufacturer, covering diverse lighting and operational scenarios; and (2) simulated scenarios recorded in a mock cleanroom environment. All videos were frame-extracted using Open Source Computer Vision Library (OpenCV 4.9.0.80), annotated on the Roboflow platform (classes: without_gloves,

Check for updates

Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 3 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

without_mask, without_suit), and exported in YOLOv10-compatible format. Image quality was enhanced using Contrast Limited Adaptive Histogram Equalization (CLAHE) for improved contrast.

After evaluating multiple multimodal LLMs (MLLMs), Llama 3.2 vision was selected for semantic report generation due to its strong vision-language alignment and fine-tuning capability, paired with YOLOv10 as the real-time detector. System integration combined both models into a monitoring platform supporting video upload and live camera streaming, with a frontend displaying bounding boxes and natural-language violation reports. Final stages involved iterative fine-tuning using real factory data, adjusting prompts and parameters to reduce missed detections and improve description accuracy followed by on-site deployment and continuous optimization of the dataset and processing pipeline to enhance robustness under complex lighting and occlusion.

2.2. Data Collection and Processing

Python was selected as the primary development language for this study due to its powerful ecosystem, clean and readable syntax, and extensive support for machine learning and AI development. Python also demonstrates strong performance in MLLM applications and asynchronous processing, making it the most suitable choice for implementing the proposed system.

2.2.1. OpenCV

We employed OpenCV as the primary tool for image and video processing. OpenCV is an open-source computer vision library widely used for object detection, image enhancement, and motion tracking. Its efficiency and Python compatibility make it well-suited for the frame-by-frame analysis required in this study. Using OpenCV, the system loads raw video data and extracts frames sequentially, converting each into a processable format and storing them for further analysis. The frame extraction interval can be adjusted as needed to maintain flexible and precise sampling.

2.2.2. Object Annotation with Roboflow

Roboflow is used to annotate attire-related violations, offering an intuitive interface and preprocessing tools for labeling categories such as without gloves, without a mask, and without a suit, with support for YOLOv10-compatible formats. Figure 1 shows frames annotated under various angles and lighting conditions, while Figure 2 illustrates example annotations where different bounding-box colors represent different violation types, such as without suit and without hat; gloves and mask indicate proper PPE usage.

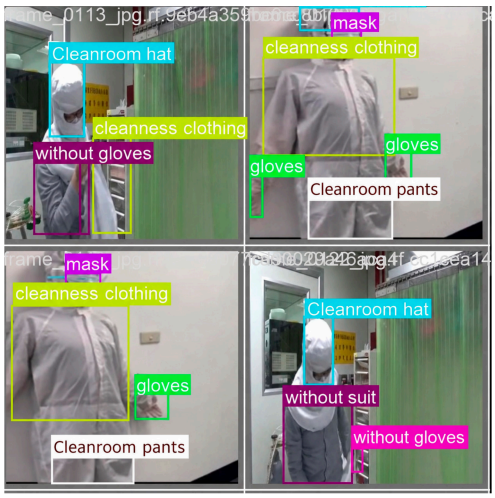


Figure 1. Extracting video frames from various viewpoints.

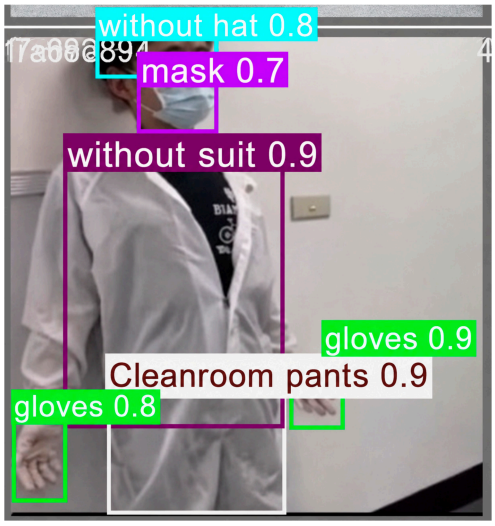


Figure 2. Roboflow labels the corresponding objects using color-coded bounding boxes.

2.3. Model Selection

In this study, Llama 3.2 vision is selected as the primary model for the following reasons.

1. Strong multimodal capabilities: It processes both images and text, performing classification, detection, and semantic description generation, making it suitable for factory monitoring tasks.
2. Robustness in complex environments: By integrating visual features with textual interpretation, the model maintains stable performance under varying lighting and complex backgrounds.
3. High scalability and multi-task support: It identifies attire violations, generates descriptions, and provides explanatory analyses, aligning well with industrial monitoring requirements.

Additionally, YOLOv10 is adopted as the real-time detection model due to the following reasons.

1. Excellent real-time performance: Its single-stage architecture ensures high-speed detection suitable for real-time production-line monitoring.
2. Strong multi-class detection capability: It supports custom datasets and can recognize multiple violation categories, such as missing gloves, masks, or cleanroom suits.
3. Flexible deployment: The model integrates easily with Python (3.10.11) and FastAPI (0.110.0) and can be exported in multiple formats for backend deployment.

An end-to-end PPE violation detection and semantic reporting system is implemented with the following pipeline:

Incoming surveillance footage is first processed by the custom-trained YOLOv10 model for real-time violation detection. The model identifies non-compliance categories (without_gloves, without_mask, without_suit) and draws red bounding boxes around violation regions (examples shown in Figures 3 and 4). Upon detection, the current frame containing the red boxes is automatically cropped and forwarded to the semantic generation module. The cropped image is then fed to Llama 3.2 vision, which directly generates a clear natural-language description without requiring additional prompt engineering (e.g., "The worker is not wearing gloves and a protective suit during operation."). The final output integrates (1) the violation image with YOLOv10 red bounding boxes, and (2) the corresponding textual report produced by Llama 3.2 vision. This unified report significantly improves readability for managers and accelerates response times.

https://doi.org/10.3390/engproc2026141001

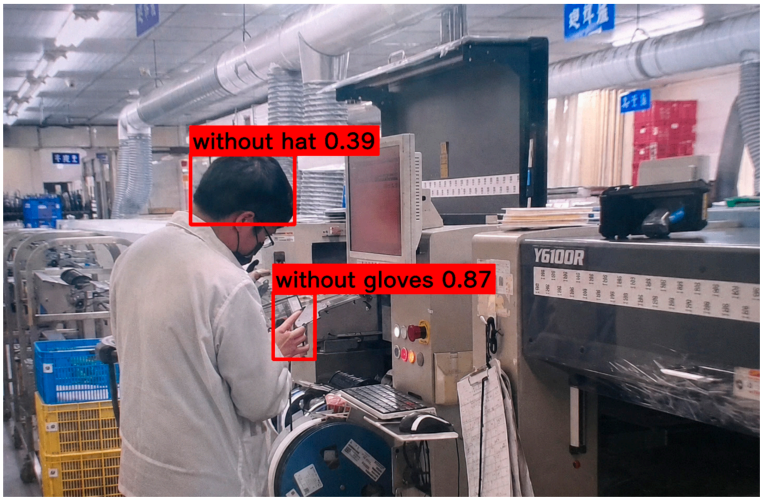


Figure 3. Actual annotation results from the factory.

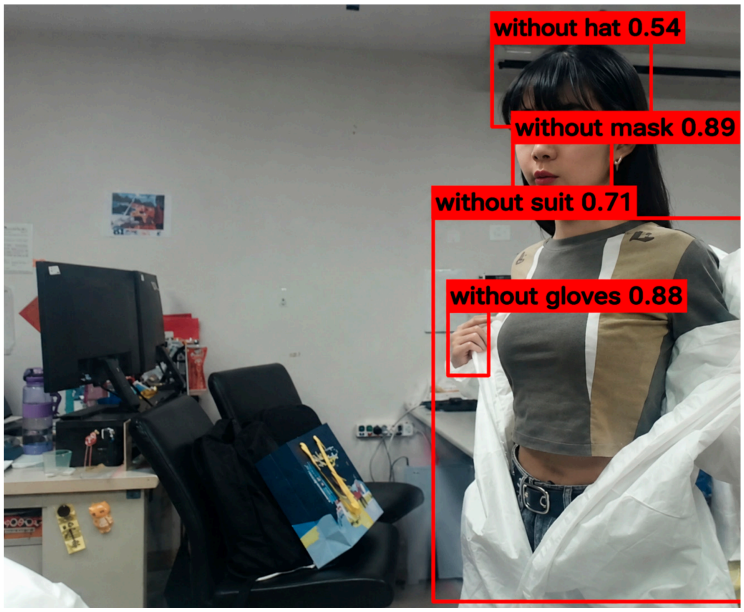


Figure 4. Annotation results from laboratory-simulated scenarios.

2.4. Model Testing and Fine-Tuning

To ensure that the monitoring platform achieves high recognition accuracy and practical usability, this study adopts separate fine-tuning strategies for YOLOv10 and LLaMA 3.2 vision. Multi-layer evaluation methods—including confusion matrices and human-expert assessment—are applied to validate the overall performance and stability of the system.

2.5. Model Training and Configuration

Llama 3.2 vision was used without parameter fine-tuning; performance was optimized solely via structured English prompt engineering and manual validation of outputs. YOLOv10 was trained from scratch on the custom dataset using Roboflow for annotation and augmentation (rotation, scaling, and CLAHE). Hyperparameters were tuned to maximize recall while maintaining >60 frames per second inference on the NVIDIA GeForce RTX 4060.

https://doi.org/10.3390/engproc2026141001

3. Results

3.1. System Interface

To enhance user experience, we designed an intuitive and user-friendly monitoring platform interface (Figure 5). The platform provides key functions including real-time video display, violation detection, visualization of annotated results, automatic semantic-description generation, and historical record tracking. In addition to presenting detection results instantly, the system also allows users to retrieve records, enabling managers to efficiently monitor factory conditions and maintain oversight of on-site activities.

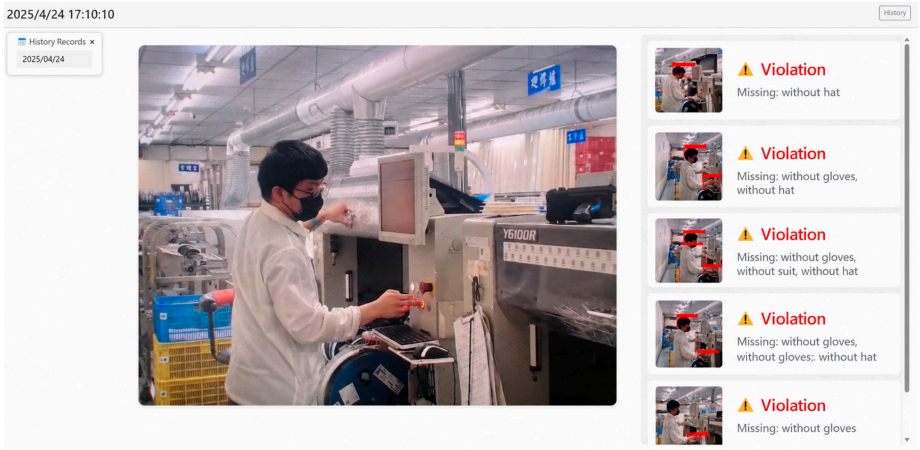


Figure 5. Interface design of the recognition platform.

3.2. Examples of Detection Performance

The following examples illustrate the effectiveness of the monitoring platform. The detection pipeline consists of two stages: YOLOv10 performs real-time attire-violation detection and marks non-compliant areas with red bounding boxes, after which the frames are processed by LLaMA 3.2 vision to generate semantic descriptions. Figure 6 shows a worker not wearing a mask or gloves. Despite complex lighting conditions, the system accurately identifies and highlights the violation. Figure 7 presents the generated descriptions, indicating violations such as without gloves and without a hat, along with relevant suggestions. Overall, the system maintains high accuracy under uneven lighting and complex backgrounds and provides clear and interpretable descriptions of attire violations.



Figure 6. YOLOv10 detection results marked with red bounding boxes.

https://doi.org/10.3390/engproc2026141001

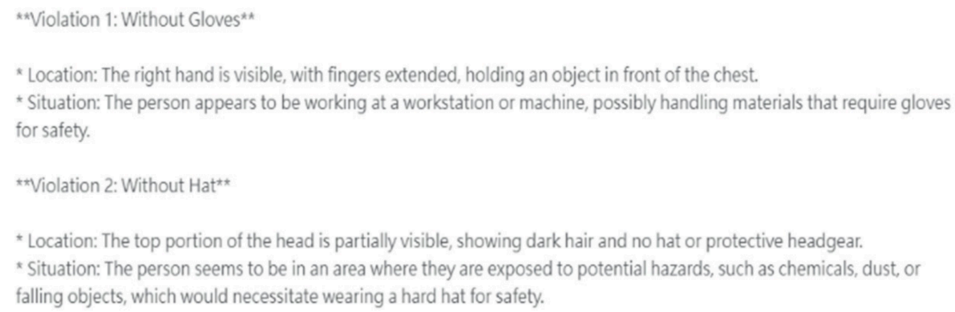


Figure 7. Llama 3.2 vision generating semantic descriptions.

3.3. Model Performance Evaluation

To evaluate the system's recognition performance, we used a confusion matrix and classification metrics. The test dataset includes varied lighting, backgrounds, and worker movements to simulate real factory conditions. YOLOv10's detection results are compared with ground-truth labels to assess classification accuracy.

The system was evaluated using a diverse test set of 500 samples covering various lighting and motion conditions. The classification results are as follows.

- True positive (TP): 146 (correctly identifying compliant workers).
- True negative (TN): 328 (correctly identifying workers with violations).
- False positive (FP): 18 (misclassifying violating workers as compliant).
- False negative (FN): 8 (misclassifying compliant workers as violating).

The results show that TP and TN cases constitute the majority of predictions, indicating strong overall classification performance. Additionally, the false-negative rate for violation detection is extremely low, demonstrating the system's high reliability in identifying non-compliant attire.

To evaluate the classification performance of the proposed attire-compliance recognition system, accuracy, precision, recall, and F1-score were calculated based on the previously obtained TP, TN, FP, and FN values.

Accuracy, defined as the proportion of correctly predicted samples [Equation (1)], reached 94.8%, demonstrating excellent overall discrimination between compliant and non-compliant PPE attire.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

Precision, the ratio of true positives to all predicted positives [Equation (2)], reached 89.02%, indicating reliable violation identification with an acceptable false alarm rate.

$$Precision = TP/(TP + FP) \quad (2)$$

Recall, the ratio of true positives to all actual positives [Equation (3)], reached 94.81%, demonstrating excellent sensitivity and minimal missed violations.

$$Recall = TP/(TP + FN) \quad (3)$$

The F1-score, the harmonic mean of precision and recall [Equation (4)], reached 91.82%, confirming excellent balance and robust overall performance.

$$F1\text{-}score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (4)$$

https://doi.org/10.3390/engproc2026141001

Figure 8 shows the precision–recall curves for all classes, with recall on the *x*-axis and precision on the *y*-axis. The overall mean average precision at 0.50 intersection over union threshold reaches 0.730, indicating solid detection performance. Classes such as without pants (0.994) and cleanroom clothing (0.946) remain close to the upper-right corner, maintaining high precision even at high recall. In contrast, anti-static shoes (0.489) and without a mask (0.776) exhibit lower stability with noticeable precision drops at higher recall levels. Future work will focus on targeted data augmentation and threshold optimization for these challenging classes.

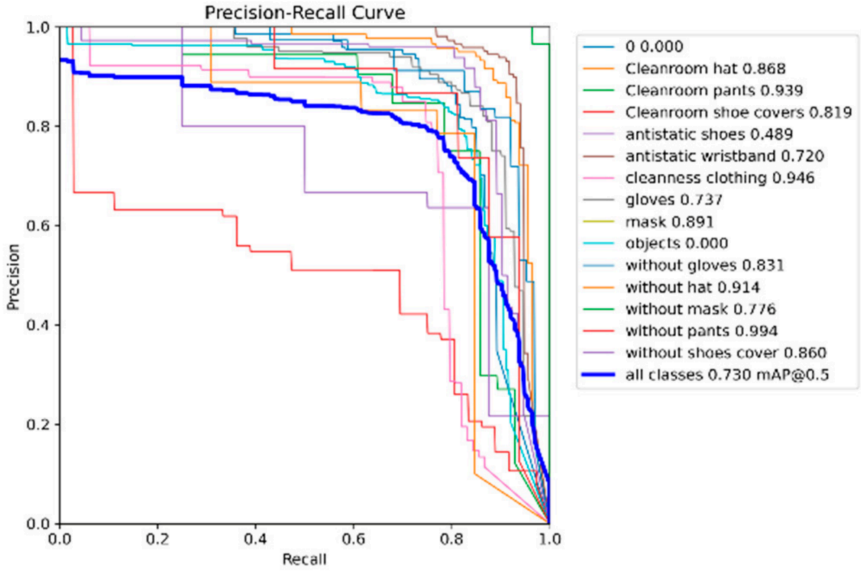


Figure 8. Precision–recall curve.

Figure 9 shows the F1-Confidence curves for all classes. The average peak F1-score reaches 0.69 at moderate confidence levels (0.5–0.8). Cleanroom clothing and cleanroom pants exhibit stable curves with slow F1 decline, while without shoes cover and without a suit display larger fluctuations. Future work will apply targeted data augmentation to improve the stability of the latter categories.

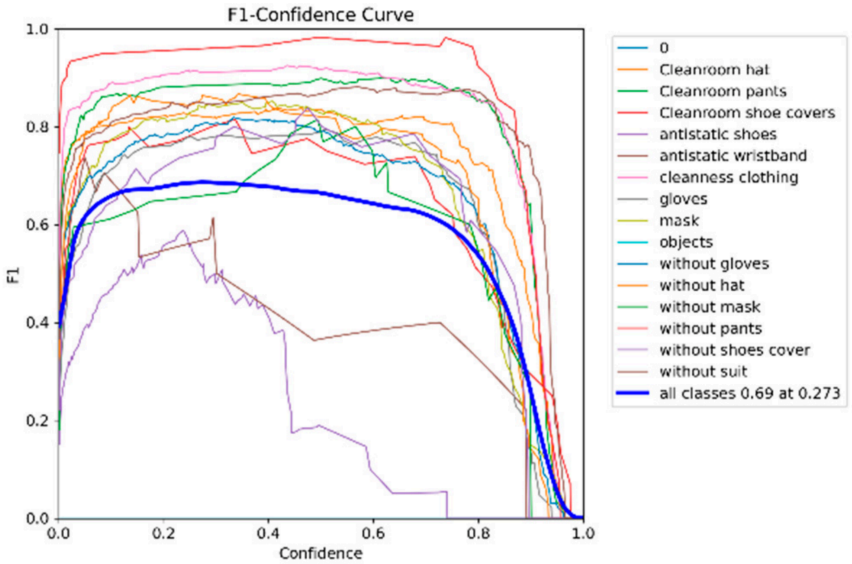


Figure 9. F1-score curve.

https://doi.org/10.3390/engproc2026141001

3.4. Semantic Description

In this study, the LLaMA 3.2 vision model was incorporated to generate semantic descriptions, converting the attire-compliance detection results into natural-language narratives. This enhances the intuitiveness and immediacy of information presented to managers. The semantic outputs not only describe the detected violations—for example, They do not have on a hat or They are not wearing gloves—but also provide explanations of potential safety risks associated with these violations, as illustrated in Figure 10.

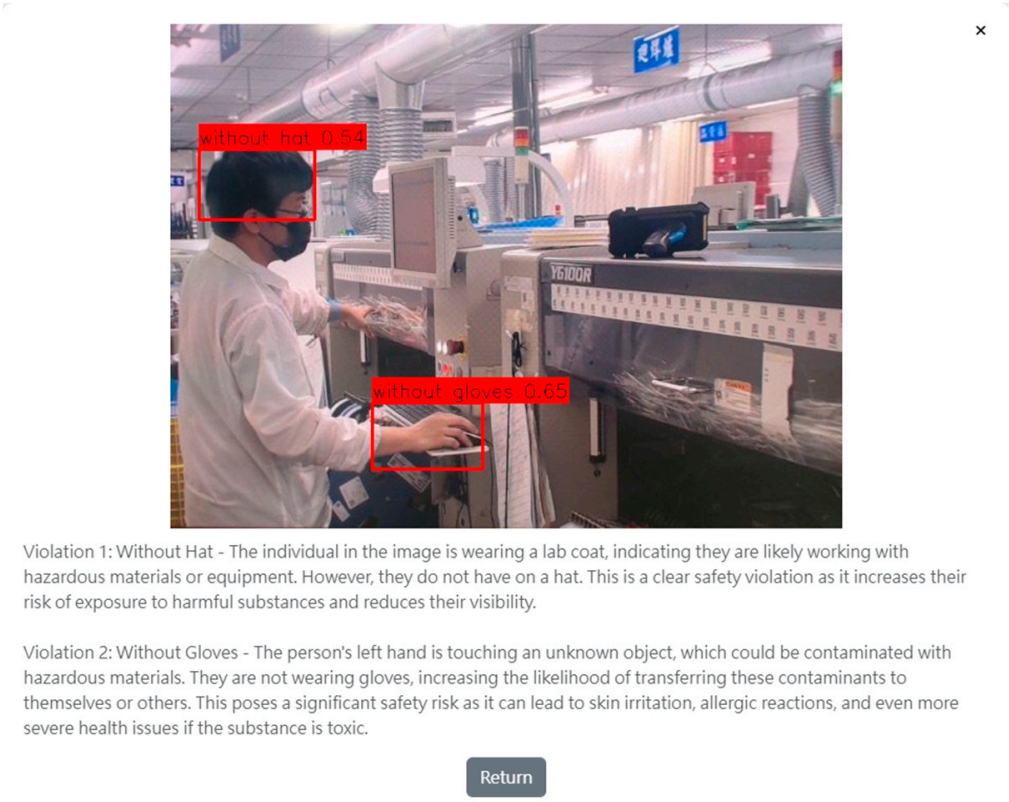


Figure 10. Examples of semantic descriptions.

The semantic descriptions provide clear explanations of various violations, such as not wearing a hat or gloves, and highlight potential safety risks, including exposure to hazardous substances or increased contamination. Overall, the system-generated statements are clear and informative, enabling managers to quickly understand violation scenarios and supporting the broader adoption of intelligent monitoring systems in factories.

4. Discussion

4.1. Analysis of Misclassification Cases

To understand the causes of recognition errors occurring in a minority of samples, this study analyzes the misclassification cases observed in the confusion matrix and classification metrics, specifically FP and FN.

1. Analysis of FN

FNs occur when a violation is present but not detected by the system. As shown in Figure 11, the worker is not wearing gloves or a cleanroom hat, yet the system identifies only the missing mask and suit. Possible causes include occlusion or poor camera angles that obscure key features, as well as insufficient training samples for glove and hat violations, leading to weaker model learning for these categories.



Figure 11. FN cases.

2. Analysis of FP

FPs occur when the system incorrectly flags a compliant worker as being in violation. In Figure 12, although the worker is properly wearing cleanroom pants, the system misclassifies the attire as without a suit with a confidence score of 0.71. This may result from clothing wrinkles, uneven lighting, or a visually complex background, which obscure boundaries and details of the attire and reduce detection accuracy.



Figure 12. FP cases.

4.2. Performance Metrics

The proposed system was comprehensively evaluated using four standard metrics: accuracy (94.8%), precision (89.02%), recall (94.81%), and F1-score (91.82%). The accuracy of

94.8% exceeds the 90% threshold generally regarded as excellent for balanced datasets [1], demonstrating high classification stability across real and simulated scenarios. The recall of 94.81% indicates that the system effectively captures nearly all actual violations, resulting in an extremely low false-negative rate, which is critical for ensuring workplace safety and product yield. Although precision (89.02%) is slightly lower than recall, reflecting occasional false positives (compliant samples misclassified as violations), it remains well above the 85% industrial benchmark [2] and does not undermine managerial trust. The F1-score of 91.82% confirms a strong balance between precision and recall, verifying the system's robust overall performance and high practical deployment value [3].

5. Conclusions

We developed an intelligent PPE compliance monitoring system by integrating YOLOv10 with the LLaMA 3.2 vision multimodal model. YOLOv10 achieved 94.8% accuracy, 94.81% recall, and a 91.82% F1-score on real and simulated factory footage, demonstrating strong and reliable violation detection despite slightly lower precision. With LLaMA 3.2 vision, the system generates clear natural-language descriptions, improving managerial readability and situational awareness. Future enhancements may include real-time alerts, human–machine collaborative interfaces, and expanded violation categories to further improve applicability in complex factory settings.

Author Contributions: Conceptualization, C.-Y.T. and C.-Y.H.; methodology, C.-Y.T., Y.-C.W. and C.-Y.H.; software, Y.-C.W.; validation, C.-Y.T., Y.-C.W. and W.-Y.L.; formal analysis, C.-Y.T.; investigation, C.-Y.T.; resources, W.-Y.L.; data curation, Y.-C.W.; writing—original draft preparation, C.-Y.T.; writing—review and editing, W.-Y.L., Y.-C.W. and C.-Y.H.; visualization, Y.-C.W.; supervision, W.-Y.L. and C.-Y.H.; project administration, W.-Y.L. and C.-Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all individuals appearing in the figures for the publication of their images in this study.

Data Availability Statement: The dataset contains proprietary factory surveillance videos and cannot be made publicly available due to commercial confidentiality. Requests for access may be sent to the corresponding author.

Acknowledgments: The authors thank Chang Hui Electronic Industry Co., Ltd. for providing factory surveillance videos and technical assistance. Consent for acknowledgement has been obtained from the company.

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

Abbreviations

The following abbreviations are used in this manuscript:

PPE	Personal Protective Equipment
MLLMs	Multimodal Large Language Models
ESD	Electrostatic Discharge
AI	Artificial Intelligence
NMS	Non-Maximum Suppression

References

1. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
2. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *arXiv* **2020**, arXiv:2010.16061. [CrossRef]
3. Goutte, C.; Gaussier, E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In *Proceedings of the 27th European Conference on Information Retrieval (ECIR 2005), Santiago de Compostela, Spain, 21–23 March 2005*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 345–359.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.