



Comparative Evaluation of Zero-Shot Vision-Language Models and YOLO for Image-Based Weld Defect Severity Assessment

¹Muhammad A'Zom Ar-Rabaqi, ¹Mohamad Yamin*

¹Gunadarma University, Jakarta, Indonesia

Corresponding Author: *mohay@staff.gunadarma.ac.id

Article submitted:

3 July 2026

Revised:

12 August 2026

Accepted:

17 August 2026

Abstract

Automated visual inspection of welded joints increasingly uses supervised deep-learning models, particularly YOLO object detectors. Although effective for defect localization, these approaches require domain-specific annotated data and provide limited engineering-oriented textual explanations. This study compared supervised YOLO-based pipelines with zero-shot Vision-Language Models (VLMs) for image-based weld-defect severity assessment. A public Welding Defect–Object Detection dataset containing 1,983 images was used. A four-level ordinal severity rubric was developed as an AWS D1.1-informed engineering heuristic rather than a direct implementation of AWS acceptance criteria because the source images lacked consistent physical scale calibration. Two supervised detectors, YOLOv8n and YOLO26n, and two zero-shot VLMs, NVIDIA Nemotron and Meta Llama 3.2 Vision-90B, were evaluated. All pipelines processed 401 validation-and-test images, while primary comparison against independently rated human references used an adjudicated subset of 100 images. Human inter-rater agreement was high (Cohen's $\kappa = 0.839$; weighted $\kappa = 0.929$). Nemotron achieved the highest accuracy (55.0%), followed by YOLOv8n (50.0%), YOLO26n (48.0%), and Llama Vision-90B (38.0%), and the highest Level-4 recall (0.871). Both YOLO pipelines showed zero recall for Level 2. Although unadjusted McNemar testing found $p = 0.030$ for Nemotron versus Llama Vision-90B, no pairwise comparison remained significant after Holm correction. VLM explanation quality was moderately associated with classification correctness ($r = 0.533$; $p = 0.0001$). None of the methods was sufficiently reliable for standalone engineering acceptance decisions, but their complementary failure patterns support VLMs as an auxiliary review layer within human-supervised weld inspection.

Keywords: Computer Vision, Explainable Artificial Intelligence, Vision-Language Model, Weld Defect Inspection, YOLO, Zero-Shot Learning.

Introduction

Welded joints are critical structural elements in buildings, ships, pressure-related systems, pipelines, transportation equipment, and heavy manufacturing. Surface or subsurface discontinuities that remain undetected may act as local stress raisers and contribute to premature damage under static or cyclic loading. Consequently, visual inspection and other non-destructive testing methods remain essential elements of quality assurance in welded structures. Engineering codes such as AWS D1.1 establish acceptance requirements for structural steel welding and provide the regulatory basis for qualified inspection procedures [1].

Manual weld inspection nevertheless presents several practical limitations. Inspection quality may depend on inspector experience, fatigue, environmental conditions, and the volume of components

requiring examination. These constraints have encouraged the development of computer-vision-based inspection systems capable of assisting human inspectors in detecting and classifying surface discontinuities. Deep-learning methods, particularly convolutional neural networks and object detectors, have therefore become increasingly relevant for industrial visual inspection [2], [3], [5].

Among object-detection methods, the YOLO family has gained substantial attention because of its favorable compromise between inference speed and detection accuracy [4]. Modern YOLO variants can identify and localize defects directly from digital images and are suitable for near-real-time industrial applications. However, conventional supervised detectors depend on annotated domain-specific datasets. Their performance may also deteriorate when imaging conditions, materials, weld appearance, or defect characteristics differ from those represented during training [5].

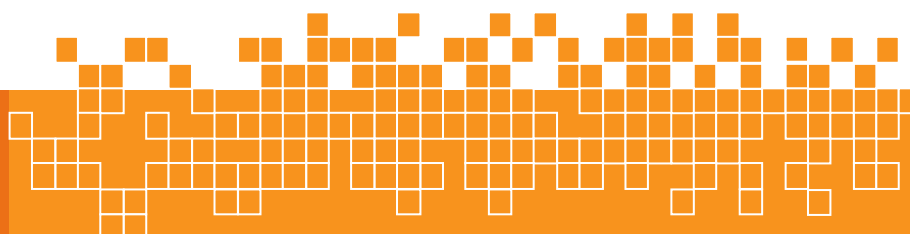
Another limitation concerns interpretability. Common explainable artificial intelligence approaches such as Grad-CAM, SHAP, and LIME primarily indicate which regions or features contributed to a model prediction [6]–[10]. Such methods are valuable for visualizing model attention or feature importance, but they do not inherently generate an explicit engineering rationale describing why a visual indication may be relevant to an inspection decision.

Recent Vision-Language Models provide a substantially different paradigm. By jointly processing images and natural language, these models can perform zero-shot visual reasoning without domain-specific parameter updating while simultaneously generating textual explanations [11]–[14]. Such properties create an opportunity for industrial inspection applications in which model predictions may be accompanied by human-readable rationales. Recent studies have begun investigating multimodal models in welding and industrial-defect applications [15]–[17], but direct evidence regarding their reliability relative to conventional supervised detectors remains limited.

A further methodological issue is that defect detection and engineering acceptance are not equivalent tasks. A bounding-box detector identifies visual regions associated with training labels, whereas engineering severity depends on characteristics including defect type, dimensions, location, loading condition, and applicable acceptance criteria. The image dataset used in the present study does not provide reliable pixel-to-millimetre calibration. Therefore, a direct implementation of AWS D1.1 acceptance criteria cannot be justified. Instead, this study employs an ordinal severity rubric informed by general AWS D1.1 principles, explicitly treating it as an experimental engineering heuristic rather than a substitute for formal code-based acceptance inspection.

Against this background, the present study compares two supervised YOLO-based pipelines with two zero-shot VLMs using a common image set. The contributions are threefold:

1. evaluation of agreement and classification performance of two AI paradigms using human reference annotations;
2. analysis of class-specific failure patterns rather than relying exclusively on aggregate accuracy; and
3. evaluation of the quality of textual explanations generated by zero-shot VLMs and their relationship with classification correctness.



Materials and Methods

A. Dataset

The study used the publicly available Welding Defect – Object Detection dataset obtained from Kaggle [21]. The original dataset consisted of 1,983 RGB images divided into 1,582 training images, 277 validation images, and 124 test images. The original object-detection annotations contained three nominal categories: Good Weld, Bad Weld, and Defect. These labels indicate visual weld condition but do not inherently represent an ordinal engineering severity scale. For training the supervised models, only the original training split of 1,582 images was used. The original validation and test subsets were combined into a pool of 401 images for the broader inter-model comparison. This combined pool is interpreted as an extended evaluation set rather than a strictly independent hold-out test set. The overall dataset composition and experimental data structure are summarized in Table 1.

Table 1. Characteristics of the Welding Defect Dataset

Attribute	Description
Dataset	Welding Defect – Object Detection
Total images	1,983
Training images	1,582
Validation images	277
Original test images	124
Extended evaluation pool	401
Original annotation	YOLO-style bounding boxes
Original categories	Good Weld, Bad Weld, Defect
Experimental severity output	Ordinal Level 1–4

Representative examples of the original Good Weld, Bad Weld, and Defect classes are shown in Fig. 1.

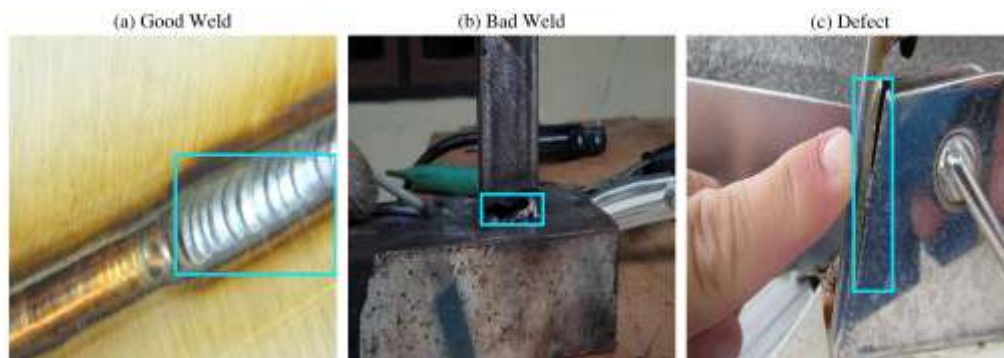


Fig. 1. Examples of the original dataset classes (Good Weld, Bad Weld, and Defect)

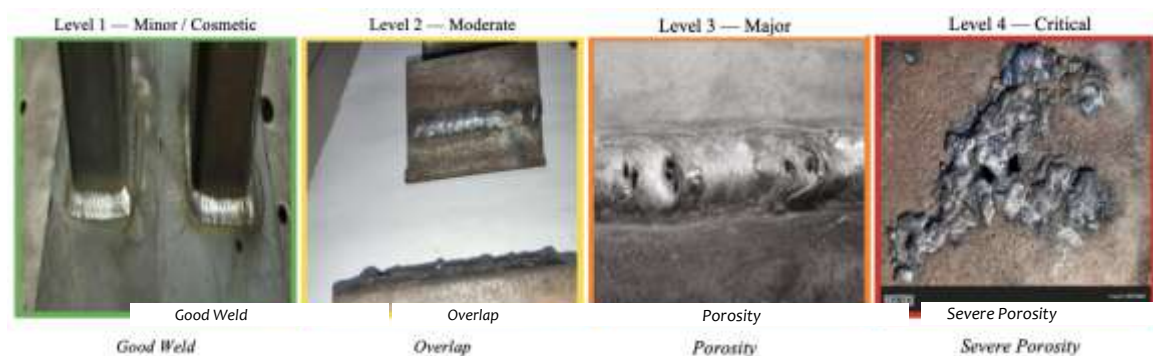
B. AWS D1.1-Informed Ordinal Severity Rubric

Because the original dataset did not contain severity labels, an ordinal four-level rubric was constructed to allow a common comparison among human raters, YOLO-derived classifications, and VLM predictions. The rubric was informed by general engineering principles reflected in AWS D1.1 but does not reproduce formal AWS D1.1 acceptance criteria. The dataset lacks consistent physical reference dimensions; therefore, defect size cannot be converted reliably from pixels to millimetres. The four experimental severity levels and their corresponding visual interpretation are summarized in Table 2.

Table 2. AWS D1.1-Informed Experimental Ordinal Severity Rubric

Level	Interpretation	Visual criterion	Experimental action interpretation
1	Low/no visible severity	Clean and continuous weld appearance without meaningful visible discontinuity	Acceptable visual appearance
2	Minor indication	One small isolated visual discontinuity while overall weld continuity remains	Review/monitor
3	Major indication	Clearly visible discontinuity disrupting weld continuity	Repair-oriented review
4	Critical indication	Crack-like indication, extensive discontinuity, or multiple visible defects	Immediate expert review

Representative weld images corresponding to the four experimental severity levels are presented in Fig. 2.

**Fig. 2.** Representative images for the four experimental severity levels

C. Human Reference Annotation

Human reference annotation was performed independently by two raters on an overlapping subset of 100 images. The second rater independently annotated the 100-image overlap without access to the first rating. Disagreements were subsequently resolved through adjudication to produce the final reference severity for the primary performance comparison. Before adjudication, the two raters agreed on 88 of 100 cases. Inter-rater reliability was evaluated using both unweighted and quadratic-weighted Cohen's Kappa. Because inter-rater agreement measures consistency rather than absolute code compliance, the resulting labels are referred to as human reference annotations rather than definitive AWS acceptance ground truth.

D. Supervised YOLO Models

Two supervised detectors were evaluated: YOLOv8n and YOLO26n. Both nano-scale variants were independently trained using the same original training dataset of 1,582 images. Input resolution was 640×640 pixels, batch size was 16, and training was conducted for up to 100 epochs using an NVIDIA T4 GPU in Google Colab. Image augmentation was restricted to the training split and included horizontal flipping, HSV variation, translation, scaling, and mosaic augmentation. The principal training configurations of YOLOv8n and YOLO26n are listed in Table 3.

Table 3. Training Configuration of YOLOv8n and YOLO26n

Parameter	YOLOv8n	YOLO26n
Input size	640×640	640×640
Maximum epochs	100	100

Parameter	YOLOv8n	YOLO26n
Batch size	16	16
Model scale	Nano	Nano
Framework	Ultralytics	Ultralytics
Hardware	NVIDIA T4	NVIDIA T4

A schematic comparison of the two supervised YOLO pipelines is illustrated in Fig. 3. Note that The main difference lies in the prediction head; both models were trained for 100 epochs using 640×640 -pixel images, a batch size of 16, and the Adam optimizer.

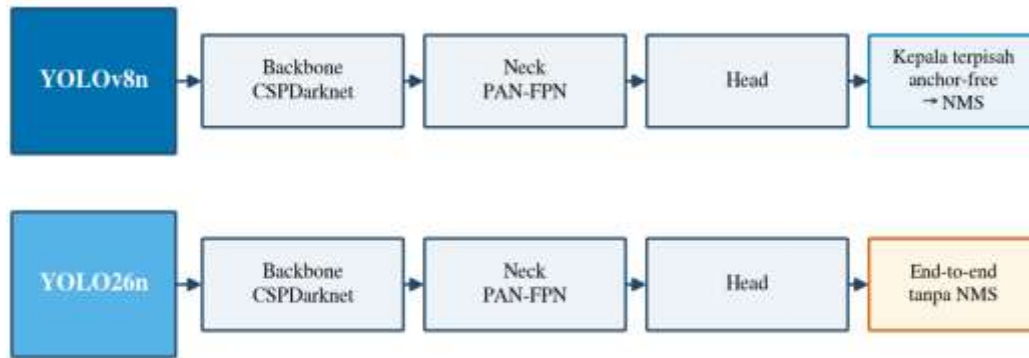


Fig. 3. Schematic comparison of the YOLOv8n and YOLO26n pipelines

E. Bounding-Box-to-Severity Mapping

YOLO produces object-level detections rather than an image-level severity grade. Therefore, a deterministic post-processing rule was required to convert the detected bounding boxes into the experimental four-level severity scale. Bounding boxes classified as *Good Weld* were excluded from defect-area calculations. The area of the largest detected defect relative to the total image area was then used as a heuristic severity indicator:

$$R_A = \left(\frac{A_{\text{defect,max}}}{A_{\text{image}}} \right) \times 100\% \quad (1)$$

In this study, $R_A < 5\%$ was mapped to Level 2, $5\% \leq R_A \leq 15$ to Level 3, and $R_A > 15\%$ to Level 4. If two or more separate defect bounding boxes were detected within the same image, the image was directly assigned to Level 4. These thresholds were introduced as experimental design parameters to enable a consistent image-level severity comparison across the evaluated methods. They were informed by the general engineering logic of AWS D1.1 but were not derived from, nor validated against, the numerical acceptance criteria specified in the standard. Accordingly, the proposed mapping should be interpreted as an experimental severity heuristic rather than as an AWS D1.1 acceptance rule. The complete bounding-box-to-severity decision process is illustrated in Fig. 4.

F. Zero-Shot Vision-Language Models

Two VLMs from different model families were evaluated: NVIDIA nemotron-3-nano-omni-30b-a3b-reasoning and Meta llama-3.2-90b-vision-instruct. The models were accessed through NVIDIA NIM. No fine-tuning or parameter updating was performed using the weld dataset; therefore, the evaluation represented a zero-shot inference setting. Both VLMs received the same system-level severity instructions and returned a severity grade, predicted defect type, and short textual explanation. The

zero-shot VLM inference workflow and the structured output format are illustrated in Fig. 5 and Fig. 6. No model weight updates are performed; all predictions are generated in a zero-shot inference setting.

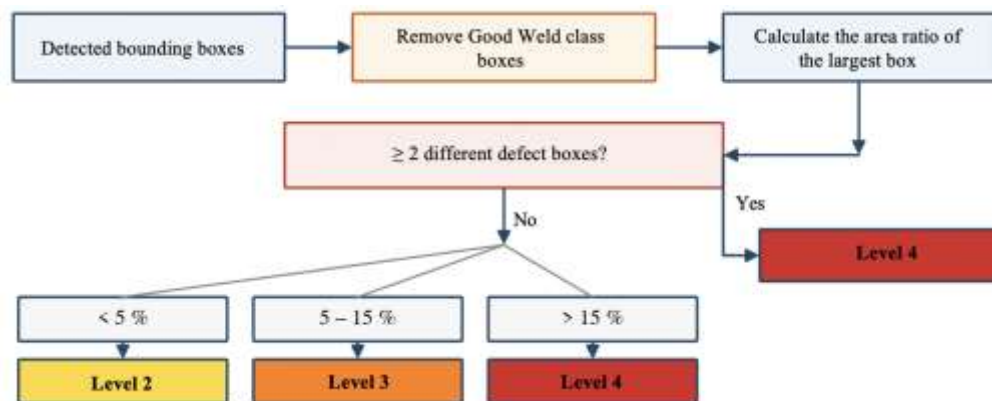


Fig. 4. Bounding-box-to-severity mapping rule

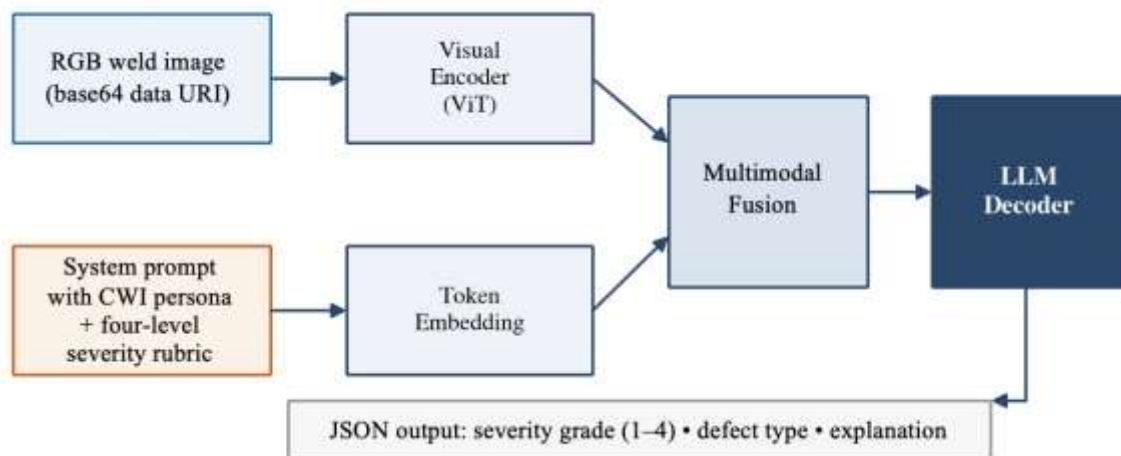


Fig. 5. Zero-shot VLM inference workflow and structured JSON output

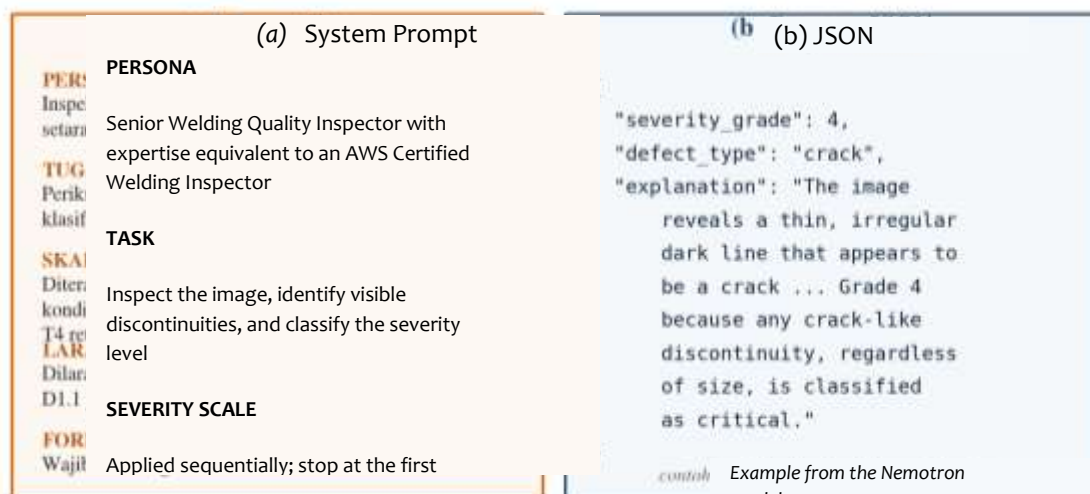


Fig. 6. Structure of the system prompt and an example of the structured JSON response generated by the VLM.

G. Evaluation Metrics

YOLO detection performance was evaluated using precision, recall, mAP50, and mAP50–95. Image-level severity performance was evaluated using accuracy, unweighted Cohen's Kappa, quadratic-weighted Cohen's Kappa, class-specific precision, recall and F1-score, and pairwise exact McNemar tests. Because six pairwise McNemar comparisons were performed, the raw p-values were additionally interpreted using Holm family-wise error correction.

H. Human Evaluation of VLM Explanations

Textual explanations produced by the VLMs were evaluated by two human raters using five Likert-scale dimensions: accuracy of visual description, appropriateness of severity reasoning, engineering reasoning depth, absence of apparent hallucination, and practical usefulness. A total of 68 image-model explanation samples were rated. The relationship between aggregated explanation score and classification correctness was examined using Pearson correlation for samples with available reference labels.

Results

On the 100-image independently rated subset, raw inter-rater agreement was 88.0%. The unweighted Cohen's Kappa was 0.839, while quadratic-weighted Kappa reached 0.929. These results indicate high consistency between the two human raters under the proposed experimental rubric.

A. YOLO Object-Detection Performance

The bounding-box detection performance of YOLOv8n and YOLO26n is summarized in Table 4 using mAP50, mAP50–95, precision, and recall.

Table 4. Object Detection Performance of YOLOv8n and YOLO26n

Model	mAP50	mAP50–95	Precision	Recall
YOLOv8n	0.6564	0.4004	0.6524	0.6203
YOLO26n	0.6007	0.3850	0.6085	0.5933

YOLOv8n achieved higher bounding-box detection metrics than YOLO26n on the available evaluation data. Detection accuracy and severity-classification accuracy should not be interpreted as equivalent because the latter additionally depends on the bounding-box-to-severity transformation.

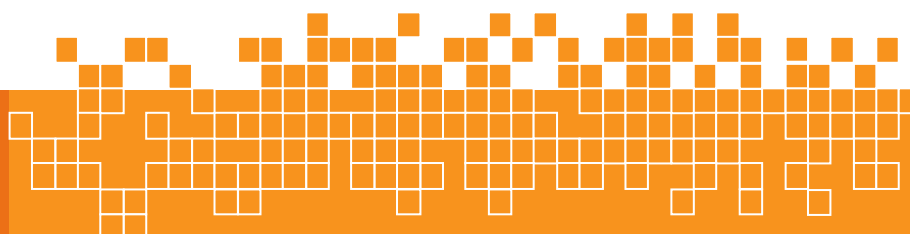
B. Severity Classification Performance

The image-level severity classification performance of the four automated methods against the adjudicated human reference annotations is presented in Table 5.

Table 5. Image-Level Severity Classification Performance Against Human Reference Annotations

Method	Accuracy	Unweighted Kappa	Weighted Kappa
YOLOv8n-based pipeline	50.0%	0.300	0.650
YOLO26n-based pipeline	48.0%	0.272	0.621
Nemotron	55.0%	0.376	0.626
Llama Vision-90B	38.0%	0.199	0.436

The class-specific error distributions underlying these aggregate performance measures are further illustrated by the confusion matrices in Fig. 7.



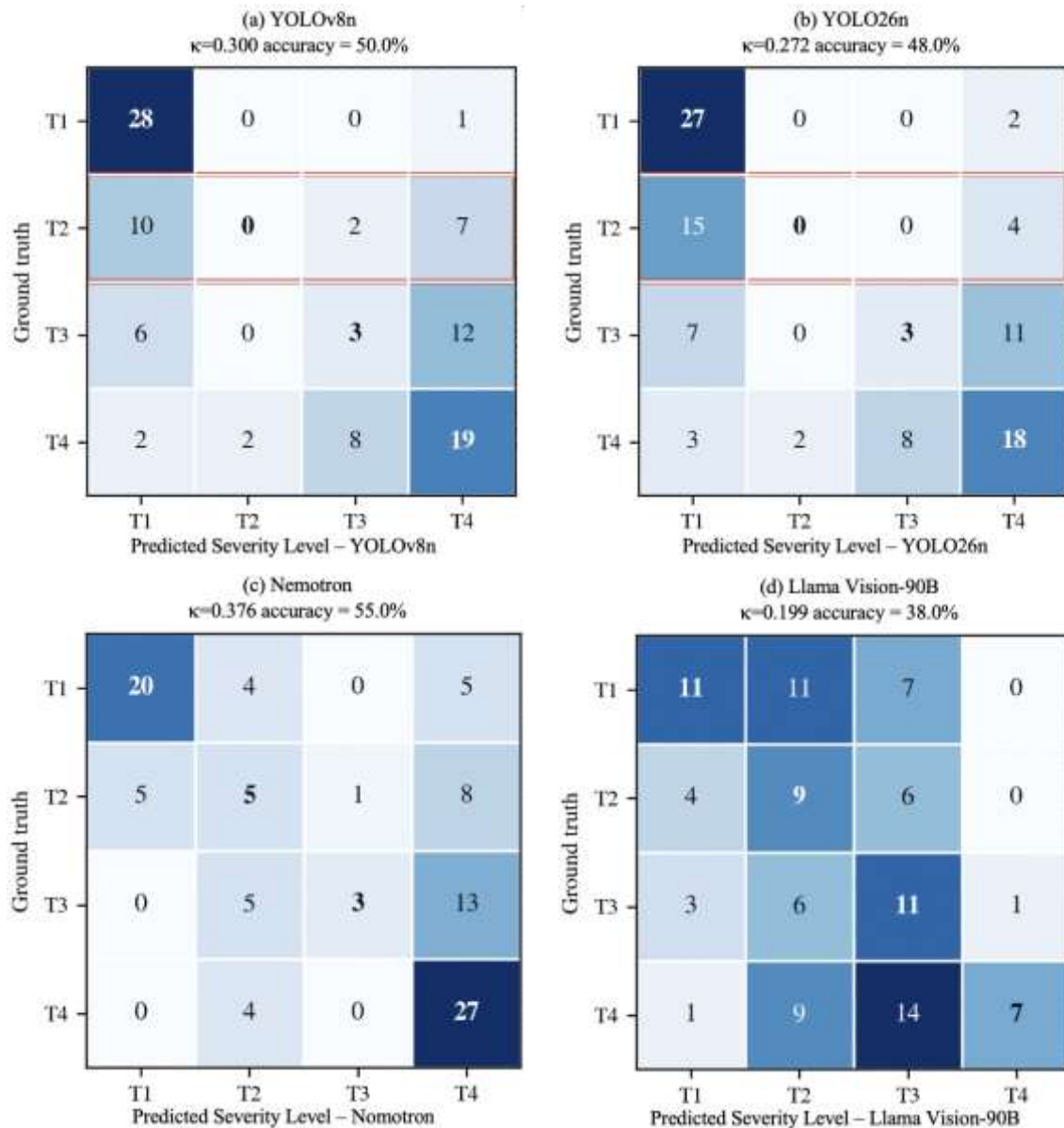


Fig. 7. Confusion matrices of the four automated methods against human reference annotations

C. Class-Specific Failure Patterns

To examine the most safety-relevant class-specific differences, precision and recall for Severity Levels 2 and 4 are compared in Table 6.

Table 6. Class-Specific Precision and Recall for Severity Levels 2 and 4

Model	Precision L2	Recall L2	Precision L4	Recall L4
YOLOv8n-based pipeline	0.000	0.000	0.487	0.613
YOLO26n-based pipeline	0.000	0.000	0.514	0.581
Nemotron	0.278	0.263	0.509	0.871
Llama Vision-90B	0.257	0.474	0.875	0.226

Both YOLO-based severity pipelines produced zero recall at Level 2. This result should not be interpreted as proof that the underlying YOLO detectors are intrinsically incapable of recognizing minor defects because severity assignment also depended on the shared post-processing rule. Nemotron achieved high Level-4 recall but showed an over-escalation tendency, while Llama Vision-90B

demonstrated high Level-4 precision with very low Level-4 recall. The complete precision and recall profiles across all severity levels are shown in Fig. 8.

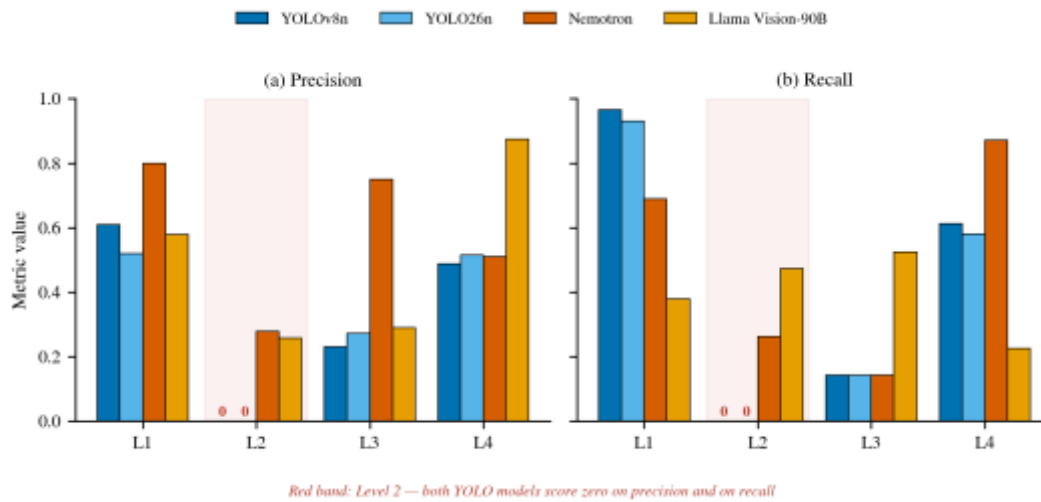
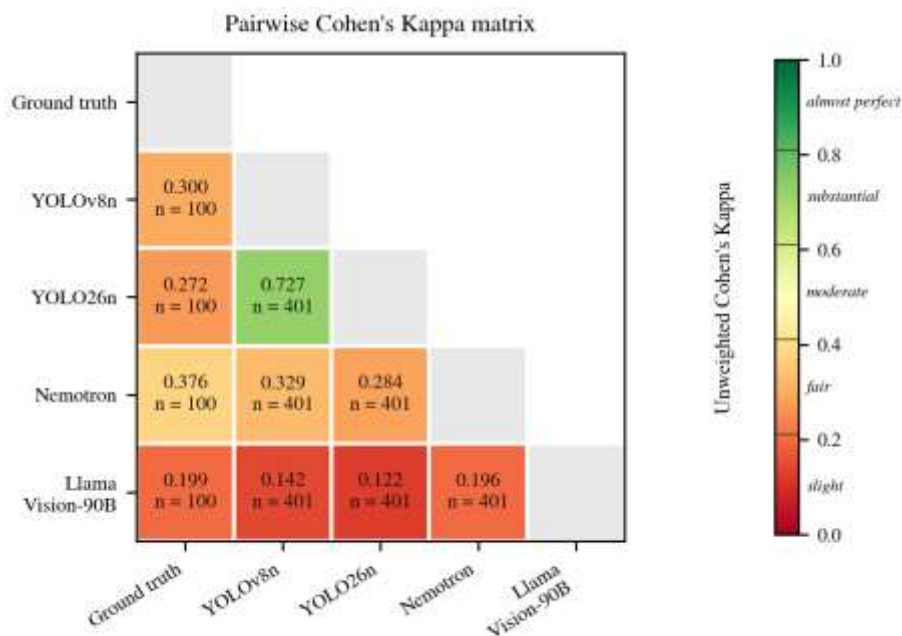


Fig. 8. Precision and recall by severity level for all four methods

D. Agreement Between Automated Models

Agreement between YOLOv8n and YOLO26n on the complete 401-image extended evaluation pool was relatively high ($\kappa = 0.727$). By contrast, $\kappa_{\text{YOLOv8n}, \text{Nemotron}} = 0.329$, $\kappa_{\text{YOLO26n}, \text{Nemotron}} = 0.284$, and $\kappa_{\text{Nemotron}, \text{Llama}} = 0.196$. The high agreement between the two YOLO pipelines should not automatically be interpreted as higher correctness because both pipelines were trained on the same source dataset and used the same post-processing severity mapping. The overall pattern of pairwise agreement among the automated methods is visualized in the Cohen's Kappa matrix shown in Fig. 9.



Human inter-rater reliability reference (Rater 1 vs Rater 2): $\kappa = 0.839$ — almost perfect

Fig. 9. Pairwise Cohen's Kappa agreement matrix

E. Pairwise Statistical Comparison

The raw and Holm-adjusted results of the six pairwise McNemar comparisons are reported in Table 7.

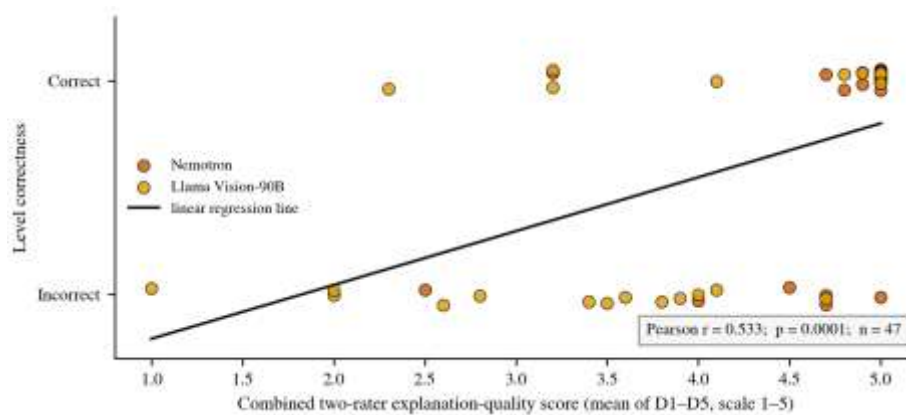
Table 7. Pairwise McNemar Test Results Before and After Holm Correction

Pair	Raw p	Holm-adjusted p	Significant after Holm correction?
YOLOv8n vs YOLO26n	0.7266	0.9738	No
YOLOv8n vs Nemotron	0.4869	0.9738	No
YOLOv8n vs Llama Vision-90B	0.1480	0.7400	No
YOLO26n vs Nemotron	0.2810	0.9480	No
YOLO26n vs Llama Vision-90B	0.2370	0.9480	No
Nemotron vs Llama Vision-90B	0.0300	0.1800	No

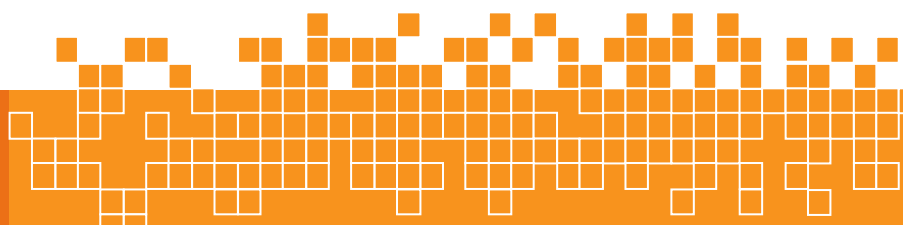
After controlling the family-wise error rate using the Holm procedure, none of the pairwise accuracy differences remained statistically significant at $\alpha = 0.05$. Therefore, Nemotron may be described as having the highest observed accuracy, but not as statistically superior to the other methods.

F. Human Evaluation of VLM-Generated Explanations

The highest weighted inter-rater Kappa for explanation assessment was obtained for appropriateness of severity criteria ($\kappa_w = 0.802$), followed by practical usefulness ($\kappa_w = 0.784$). The lowest agreement was observed for the hallucination-related dimension ($\kappa_w = 0.317$). For 47 explanation samples with corresponding human reference severity labels, the combined explanation score showed a statistically significant positive correlation with classification correctness ($r = 0.533$; $p = 0.0001$). The relationship between explanation-quality score and classification correctness is visualized in Fig. 10.

**Fig. 10.** Relationship between explanation-quality score and classification correctness**G. Representative Failure Cases**

Three qualitative cases illustrate the quantitative trends: a crack-like indication correctly escalated by both VLMs but missed by both YOLO-based pipelines; a visually clean weld in which a linear tool-related artifact was misinterpreted by a VLM as a crack-like feature; and a case showing sensitivity to VLM model capacity. These examples demonstrate that inter-model agreement does not guarantee agreement with the human reference and that VLMs can misread non-defect linear structures as critical indications. The three representative qualitative cases and the corresponding predictions from the evaluated methods are presented in Fig. 11.



(a) Case 1 — crack caught by the VLMs, missed by YOLO



Source	Level	Defect type
Ground truth	Level 4	Cold Crack
YOLOv8n	Level 1	Good Weld
YOLO26n	Level 1	none
Nemotron	Level 4	crack
Llama Vision-90B	Level 4	crack

Both supervised detectors produce a false negative on a critical defect, whereas both zero-shot VLMs identify it correctly.

(b) Case 2 — over-escalation on a clean weld



Source	Level	Defect type
Ground truth	Level 1	No significant defect
YOLOv8n	Level 1	Good Weld
YOLO26n	Level 1	Good Weld
Nemotron	Level 4	crack
Llama Vision-90B	Level 1	none

The zero-tolerance crack rule forces Nemotron prone to reading this linear non-defect artefacts as cracks, raising the level to critical.

(c) Case 3 — explanation stability across calls



Source	Level	Defect type
Nemotron — pilot run	Level 4	incomplete fusion
Nemotron — full run	Level 4	crack
YOLOv8n	Level 4	Bad Weld
YOLO26n	Level 1	none
Llama Vision-90B	Level 4	crack

The same image submitted twice to the same model returns an identical severity level but a different defect type. Level stability does not guarantee explanation stability — the two must be reported separately.

Fig. 11. Three representative qualitative failure cases with predictions from all methods

Discussion

A. Overall Performance and Suitability for Engineering Inspection

The present results indicate that none of the evaluated automated methods demonstrated sufficient standalone reliability for engineering acceptance decisions. As summarized in Table 5, the highest observed accuracy was achieved by Nemotron at 55.0%, followed by the YOLOv8n-based pipeline at 50.0%, YOLO26n-based pipeline at 48.0%, and Llama Vision-90B at 38.0%. The corresponding unweighted Cohen's Kappa values remained relatively modest, with the highest value of 0.376 obtained by Nemotron. The confusion matrices in Fig. 7 further demonstrate that prediction errors were distributed across multiple severity levels rather than being confined to a small number of isolated cases.

These results suggest that the principal contribution of the study should not be interpreted as identifying a universally superior model. Instead, the findings demonstrate that supervised object-detection pipelines and zero-shot VLMs produce different error structures when applied to image-based weld severity assessment. Considering the safety-sensitive nature of welded structural components, the observed performance levels are not sufficient to support autonomous acceptance or rejection decisions.

The evaluated models are therefore more appropriately positioned as decision-support tools within a human-supervised inspection workflow.

B. Performance of the Supervised YOLO-Based Pipelines

At the object-detection level, YOLOv8n demonstrated consistently higher performance than YOLO26n, as shown in Table 4. YOLOv8n achieved an mAP50 of 0.6564 and recall of 0.6203, compared with 0.6007 and 0.5933, respectively, for YOLO26n. However, these differences at the detection level did not translate directly into equivalent differences in image-level severity classification. As shown in Table 5, the final severity accuracies of the two pipelines were relatively similar, at 50.0% and 48.0%.

This distinction emphasizes that object localization and severity classification represent different stages of the overall inspection pipeline. The YOLO models were originally trained using the nominal categories *Good Weld*, *Bad Weld*, and *Defect*, whereas the four-level severity output was subsequently obtained using the bounding-box-to-severity mapping illustrated in Fig. 4. Consequently, classification performance reflects not only the capability of the detector, but also the assumptions embedded in the post-processing rule.

The high pairwise agreement between the two YOLO-based pipelines provides additional evidence for this interpretation. As illustrated in Fig. 9, YOLOv8n and YOLO26n achieved a Cohen's Kappa of 0.727 on the 401-image extended evaluation pool, which was substantially higher than their agreement with the human reference annotations. This high internal agreement should therefore not be interpreted as independent confirmation of prediction correctness. Because both models were trained on the same source dataset and subsequently processed through the same severity-mapping scheme, part of their agreement may arise from shared methodological assumptions.

C. Interpretation of the Level-2 Blind Spot

One of the most important class-specific findings is the complete failure of both YOLO-based severity pipelines to correctly classify Level-2 cases. As shown in Table 6, both pipelines produced Level-2 precision and recall values of zero. The broader class-specific precision and recall distributions presented in Fig. 8 confirm that this failure was not merely a consequence of low aggregate accuracy, but represented a systematic class-level weakness.

Nevertheless, this result should not be interpreted as evidence that the underlying YOLO architectures are intrinsically incapable of detecting minor weld defects. The final severity assignment depended on the heuristic area thresholds defined in the mapping rule shown in Fig. 4. A Level-2 misclassification may therefore originate from several sources, including failure to localize a small defect, inaccurate bounding-box size estimation, or the inadequacy of the 5% and 15% area thresholds used to transform detector outputs into severity categories.

The appropriate interpretation is therefore that a Level-2 blind spot occurred in the evaluated YOLO-based severity pipelines under the proposed mapping scheme. Future work should include threshold-sensitivity and ablation analyses to determine whether the observed zero recall persists under alternative severity-mapping configurations. Such an analysis would allow detector-related errors to be separated more clearly from errors introduced during post-processing.

D. Contrasting Failure Profiles of the Two VLMs

The two zero-shot VLMs exhibited markedly different class-specific behaviors. As shown in Table 6 and Fig. 8, Nemotron achieved a Level-4 recall of 0.871, substantially higher than the other evaluated approaches. However, its Level-4 precision was only 0.509, indicating a tendency toward over-escalation. In contrast, Llama Vision-90B achieved a high Level-4 precision of 0.875 but a recall of only 0.226, indicating that the model frequently failed to identify cases assigned to the most severe reference category.

The confusion matrices in Fig. 7 reinforce this contrast. Nemotron produced a prediction distribution strongly shifted toward higher severity categories, whereas Llama Vision-90B more frequently underestimated high-severity cases. These results demonstrate that zero-shot VLM behavior is strongly model dependent. Consequently, the performance observed for one VLM should not be generalized to the broader class of multimodal foundation models.

From an engineering-risk perspective, the two failure profiles are not equivalent. Over-escalation can increase unnecessary manual inspection or repair-related workload, while under-detection of a critical indication may permit a potentially unsafe component to proceed without further review. For applications in which missed critical defects represent the dominant safety concern, the high Level-4 recall observed for Nemotron may be more desirable than the higher Level-4 precision but low recall of Llama Vision-90B. However, this does not establish Nemotron as an overall superior model because its aggregate accuracy remained limited and the pairwise statistical comparisons did not demonstrate significant superiority after correction.

E. Inter-Model Agreement Does Not Necessarily Indicate Correctness

The pairwise Cohen's Kappa matrix in Fig. 9 provides an important perspective that is not apparent from accuracy alone. The two YOLO-based pipelines exhibited substantial agreement with each other, whereas the two VLMs showed only slight agreement. Nemotron and Llama Vision-90B achieved a pairwise Kappa of only 0.196, despite both operating under the same zero-shot severity instructions.

This finding indicates that model agreement should not automatically be interpreted as evidence of correctness. The strong agreement between YOLOv8n and YOLO26n may partly reflect similarities in their training data and shared severity mapping, while disagreement between the VLMs may reflect differences in their underlying multimodal representations and reasoning behavior.

For future inspection-system design, disagreement itself may therefore be useful information. Rather than selecting only the model with the highest aggregate accuracy, a hybrid system could use disagreement between independently structured models as a trigger for targeted human review.

F. Statistical Significance and Interpretation of Model Ranking

Although Nemotron achieved the highest observed accuracy in Table 5, the statistical analysis does not support a conclusion of inferential superiority. The six pairwise McNemar comparisons and their Holm-adjusted values are summarized in Table 7. The original raw comparison between Nemotron and Llama Vision-90B yielded , but the Holm-adjusted value increased to . Therefore, none of the six pairwise comparisons remained statistically significant after controlling the family-wise error rate.

This result is important because it distinguishes descriptive ranking from statistical superiority. Nemotron may legitimately be described as the model with the highest observed accuracy and highest Level-4 recall in the present dataset, but the available sample does not provide sufficient evidence to conclude that its underlying classification performance is statistically superior to the other evaluated methods.

The limited size of the adjudicated reference subset is likely an important contributor to this uncertainty. A larger independently annotated test set would provide greater statistical power and narrower uncertainty intervals for future comparisons.

G. Interpretation of VLM-Generated Explanations

A distinctive feature of the zero-shot VLM approach is the ability to generate textual rationales alongside severity predictions. As illustrated in Fig. 10, the combined explanation-quality score showed a moderate positive association with classification correctness, with and . This result suggests that explanations judged to be visually specific and engineering-relevant were more likely to accompany correct severity predictions.

However, the magnitude of the correlation is not sufficiently high to justify treating explanation quality as a calibrated confidence metric. A linguistically coherent or technically plausible explanation may still accompany an incorrect classification. This limitation is also supported by the qualitative example in Fig. 11, where a VLM interpreted a non-defect linear artifact associated with an inspection tool or shadow as a crack-like indication.

The value of textual explanations should therefore be interpreted primarily as a human-readable review aid. Unlike saliency-based XAI methods that highlight influential image regions, VLM explanations can explicitly describe the visual feature being interpreted and provide an engineering-oriented rationale. Nevertheless, the present experiment evaluates the perceived quality of these explanations rather than their formal causal faithfulness to the internal model computation.

H. Qualitative Failure Cases and Their Engineering Meaning

The representative cases presented in Fig. 11 provide concrete examples of the quantitative error patterns identified in Table 6 and Fig. 8. In one case, both VLMs correctly escalated a crack-like indication to Level 4 while both YOLO-based pipelines underestimated the severity. This case illustrates that agreement between supervised detectors does not necessarily imply agreement with the human reference annotation.

A second case demonstrates the opposite failure mode. A visually clean weld was escalated by a VLM because a thin linear feature associated with an inspection probe or its shadow was interpreted as a crack-like discontinuity. This example helps explain the over-escalation behavior observed for Nemotron and illustrates a potential limitation of zero-shot visual reasoning in industrial environments where tools, shadows, cables, or surface markings may mimic defect morphology.

The third case indicates possible sensitivity to model capacity. An earlier lower-capacity VLM configuration failed to classify a visually obvious high-severity indication correctly, whereas the larger 90-billion-parameter configuration produced the intended classification. This individual observation

does not establish a general scaling relationship, but it illustrates that conclusions regarding VLM performance may depend strongly on the specific model configuration used.

Implications for a Hybrid Inspection Framework

Taken together, the class-specific performance in Table 6, the precision–recall patterns in Fig. 8, and the inter-model agreement structure shown in Fig. 9 suggest that supervised detectors and VLMs may provide complementary information rather than interchangeable predictions.

A practical future architecture could therefore employ a supervised object detector as the first inspection layer, followed by a VLM-based secondary assessment. Cases in which both methods agree on a low-risk result could proceed through the normal review workflow, whereas disagreement involving a higher VLM severity estimate could trigger mandatory human inspection. For example, such a workflow directly exploits the complementary error profiles demonstrated in Table 6 and Fig. 8 rather than attempting to replace one model family with another. It may also utilize the textual VLM explanations illustrated in Fig. 10 and Fig. 11 to provide the human inspector with additional contextual information.

However, the proposed architecture remains a design implication derived from the current results. It was not experimentally evaluated as an operational human–AI inspection system in this study. Future work should therefore test whether model disagreement and textual explanations actually improve human inspection accuracy, decision time, and reliability.

Study Limitations

The results should be interpreted in the context of several limitations. First, the four-level severity scale summarized in Table 2 represents an experimental image-based rubric informed by general AWS D1.1 principles rather than formal code-based acceptance criteria. The source dataset did not provide a consistent physical scale, preventing direct conversion of visual dimensions from pixels to millimetres.

Second, the bounding-box-to-severity mapping shown in Fig. 4 employs heuristic thresholds of 5% and 15% that have not yet been calibrated against physical defect dimensions. The systematic Level-2 failure shown in Table 6 therefore requires further sensitivity analysis before it can be attributed to detector performance.

Third, the primary performance comparison in Table 5 was based on only 100 adjudicated images. Although inter-rater consistency was high, the limited sample size reduced statistical power, as reflected in the non-significant Holm-adjusted McNemar results in Table 7.

Fourth, the broader 401-image inter-model comparison includes both validation and test images and therefore should not be interpreted as a fully independent external test set.

Fifth, the image-based analysis cannot confirm all forms of internal weld discontinuity. Consequently, the severity terminology used in the present study represents visual indications rather than definitive NDT diagnoses.

Finally, only two VLMs and two YOLO configurations were evaluated. The strong differences between Nemotron and Llama Vision-90B observed in Fig. 7, Fig. 8, and Fig. 9 demonstrate that conclusions from an individual model should not be generalized to the entire VLM paradigm.

Conclusion

This study compared two supervised YOLO-based severity pipelines with two zero-shot Vision-Language Models for image-based assessment of weld-defect severity. Nemotron achieved the highest observed accuracy of 55.0%, followed by YOLOv8n at 50.0%, YOLO26n at 48.0%, and Llama Vision-90B at 38.0%. However, after Holm correction of six pairwise McNemar comparisons, no statistically significant difference remained among the evaluated approaches. The principal contribution is therefore the identification of different and potentially complementary failure profiles rather than evidence that VLMs outperform supervised detectors. Both YOLO-based severity pipelines yielded zero Level-2 recall under the proposed mapping, whereas Nemotron showed high Level-4 sensitivity accompanied by over-escalation and Llama Vision-90B showed high Level-4 precision but low Level-4 recall. VLM-generated explanations showed a moderate positive association with classification correctness ($r = 0.533$; $p = 0.0001$), supporting their potential use as auxiliary review cues rather than calibrated confidence measures. Overall, none of the evaluated automated methods demonstrated sufficient reliability for standalone weld acceptance decisions. Future work should prioritize physical-scale calibration, expert validation against applicable welding-code criteria, threshold sensitivity analysis, annotation of the complete independent test split, repeated VLM inference for consistency assessment, and validation of a detector-VLM-human hybrid workflow.

References

- [1] American Welding Society, *AWS D1.1/D1.1M:2025, Structural Welding Code—Steel*, 25th ed. Miami, FL, USA: American Welding Society, 2025.
- [2] T. H. Ngo, H. M. Q. Tang, and Q. B. Diep, "Weld-CNN: Advancing non-destructive testing with a hybrid deep learning model for weld defect detection," *Advances in Mechanical Engineering*, 2025, doi: 10.1177/16878132251341615.
- [3] M. Torres-Torres, K. Velasquez, J. Escorcia-Gutierrez, and A. Valls, "Visual weld quality inspection system for identifying defects in GMAW joints with semantic segmentation on pre-trained models," *International Journal of Computational Intelligence Systems*, vol. 19, Art. no. 110, 2026, doi: 10.1007/s44196-026-01197-z.
- [4] R. Sapkota et al., "YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series," *Artificial Intelligence Review*, vol. 58, no. 9, Art. no. 274, 2025, doi: 10.1007/s10462-025-11253-3.
- [5] Z. Li, Y. Yan, X. Wang, Y. Ge, and L. Meng, "A survey of deep learning for industrial visual anomaly detection," *Artificial Intelligence Review*, vol. 58, Art. no. 279, 2025, doi: 10.1007/s10462-025-11287-7.
- [6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.
- [7] J. Cação, J. Santos, and M. Antunes, "Explainable AI for industrial fault diagnosis: A systematic review," *Journal of Industrial Information Integration*, vol. 47, Art. no. 100905, 2025, doi: 10.1016/j.jii.2025.100905.
- [8] M. A. Mersha, K. N. Lam, J. Wood, A. K. AlShami, and J. K. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction," *Neurocomputing*, vol. 599, Art. no. 128111, 2024, doi: 10.1016/j.neucom.2024.128111.

- [9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
- [11] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Machine Learning (ICML)*, vol. 139, 2021, pp. 8748–8763.
- [12] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, Art. no. 13, 2019, doi: 10.1145/3293318.
- [13] S. Danish, A. Sadeghi-Niaraki, S. U. Khan, L. M. Dang, L. Tightiz, and H. Moon, "A comprehensive survey of vision–language models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets," *Information Fusion*, vol. 126, Art. no. 103623, 2025, doi: 10.1016/j.inffus.2025.103623.
- [14] C. X. Liang et al., "A comprehensive survey and guide to multimodal large language models in vision–language tasks," *Computation*, vol. 14, no. 6, Art. no. 125, 2026, doi: 10.3390/computation14060125.
- [15] G. Khvatskii, Y. S. Lee, C. Angst, M. Gibbs, R. Landers, and N. V. Chawla, "Do multimodal large language models understand welding?," *Information Fusion*, vol. 120, Art. no. 103121, 2025, doi: 10.1016/j.inffus.2025.103121.
- [16] G. Yong, K. Jeon, D. Gil, and G. Lee, "Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, pp. 1536–1554, 2023, doi: 10.1111/mice.12954.
- [17] C. Sub-r-pa and R.-C. Chen, "Adapting vision–language models for few-shot industrial defect detection," *Algorithms*, vol. 19, no. 4, Art. no. 259, 2026, doi: 10.3390/a19040259.
- [18] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.
- [19] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968, doi: 10.1037/h0026256.
- [20] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947, doi: 10.1007/BF02295996.
- [21] Sukmaadhiwijaya, "Welding Defect – Object Detection," *Kaggle*, dataset. [Online]. Available: <https://www.kaggle.com/datasets/sukmaadhiwijaya/welding-defect-object-detection>. [Accessed: Aug. 9, 2026].

Authors



Muhammad A'Zom Ar-Rabaqi is a Mechanical Engineer and currently a Master's student in Mechanical Engineering at Gunadarma University, Indonesia, where he also received his Bachelor's degree in 2025. His research focuses on structural simulations, advanced material extraction, and renewable energy technologies. (email: 1muhammadazom650@gmail.com).



Mohamad Yamin  is the Center for Automotive Research Head at Gunadarma University, Jakarta, Indonesia. His undergraduate degree is in Aeronautics from Institut Teknologi Bandung, Indonesia (1993), and his doctorate (2003) is in Aerodynamics, ILR, TU-Berlin, Germany. His research interests are Computational Fluid Dynamics, Dynamics and Control, EV, Aerodynamics, Energy, and Machine Learning. (email: mohay@staff.gunadarma.ac.id).