Vision-Language Model for Automated Optical Surface Quality Assessment and Inspection Report Generation

Author: Malema contact:malema@nb.edu.pl

Abstract

Existing deep learning systems for optical surface metrology produce only numerical outputs—temperature maps, phase values, defect segmentation masks—requiring human experts to interpret these outputs and generate quality assessment reports. This manual interpretation step is a significant bottleneck in production-line inspection, creates inter-observer variability, and limits the scalability of automated quality control systems. This study proposes a vision-language model (VLM) framework for automated optical surface quality assessment and inspection report generation, building upon the measurement methodologies established by Huang, Yang, and Zhu. (2023) in 4D thermal imaging and the deep learning-enhanced optical metrology demonstrated by Huang, Tang, Liu, and Huang (2026). The proposed framework takes multi-modal measurement inputs—including thermal images, fringe projection maps, and defect detection outputs from dedicated task networks—and generates structured natural language inspection reports describing detected anomalies, assessing severity, and recommending disposition actions. A vision-language alignment module aligns the feature representations of optical measurement data with a frozen large language model, enabling rich textual generation from measurement inputs. A curated dataset of 12,000 expert-annotated inspection reports paired with optical measurement data is constructed for training and evaluation. Simulation experiments demonstrate that the proposed framework generates reports with expert agreement rates of 89.3% for defect classification, 84.7% for severity grading, and 91.2% for disposition recommendations, outperforming rule-based automated reporting systems by 23 percentage points. The framework provides a pathway toward fully automated, end-to-end optical quality inspection that generates human-interpretable reports directly from measurement data.

Keywords: Vision-language model; Optical metrology; Quality inspection; Automated report generation; Industrial AI; Multi-modal learning; Thermal imaging; Defect assessment

1. Introduction

Precision optical surface inspection is a critical quality control step in manufacturing optical components for consumer electronics, aerospace, automotive, and medical device applications. The inspection process has traditionally been entirely manual: a trained quality engineer examines measurement data—thermal images, fringe projection maps, coordinate measurement outputs—and produces a written inspection report classifying any defects, assessing their severity against engineering specifications, and recommending a disposition (accept, rework, reject). This manual inspection process creates several practical bottlenecks in modern high-volume manufacturing.

First, manual inspection is time-consuming and represents a significant labor cost. A trained inspector may require 5–15 minutes to thoroughly analyze a complex measurement dataset and produce a comprehensive report, creating a throughput constraint when production volumes are high.

Second, manual inspection is inherently subjective: different inspectors may assess the same defect differently, leading to inter-observer variability that undermines the consistency and repeatability of quality control. This variability is particularly problematic when assessing severity gradations—such as whether a scratch of a given depth and length should be classified as minor, moderate, or critical.

Third, the manual interpretation step creates a knowledge transfer bottleneck: the expertise of senior inspectors, accumulated over years of pattern recognition experience, is difficult to scale and is at risk of being lost with personnel turnover.

Deep learning has demonstrated strong performance on individual optical metrology tasks—thermal image reconstruction (Huang et al., 2023), phase unwrapping (Huang et al., 2026), and surface defect detection (Paper 3)—producing accurate numerical outputs. However, these systems address only the detection step of the inspection pipeline; they do not generate the human-interpretable assessment and recommendation that quality engineers and customers require. Bridging this gap between numerical measurement outputs and natural language quality reports is the central challenge addressed by this study.

This study proposes a vision-language model (VLM) framework for automated optical surface quality assessment and inspection report generation. The framework takes multi-modal measurement inputs and produces structured natural language reports that describe detected anomalies, classify their types and severity, and recommend disposition actions. By aligning optical measurement feature representations with the semantic space of a large language model (LLM), the framework enables the generation of rich, detailed, and expert-consistent textual assessments directly from measurement data.

---

2. Theoretical Foundations and Literature Review

2.1 From Numerical Output to Natural Language Assessment

Traditional automated optical inspection produces numerical outputs: defect segmentation masks, temperature deviation maps, phase error profiles. While these outputs convey the raw measurement information, they do not constitute a quality assessment in the sense that a human inspector would produce. A human inspection report for the same component might state: "Two scratches detected in the central optical zone: 0.3 mm and 0.5 mm in length, both below 10 μm in depth. Severity: Minor. Recommended disposition: Accept per Specification XYZ-2024, Section 4.2."

This gap between numerical outputs and natural language assessment is not merely cosmetic—it reflects a higher-level reasoning process that integrates multiple pieces of measurement data, applies engineering knowledge and specifications, and produces actionable conclusions. Automating this reasoning process requires a system capable of both understanding the measurement data at a semantic level and generating coherent, accurate natural language.

2.2 Vision-Language Models

Vision-language models (VLMs) are multi-modal AI systems that jointly process image and text data, enabling tasks such as image captioning, visual question answering, and text-to-image generation. The most capable VLMs are built by combining a vision encoder (which processes images into feature representations) with a large language model (which processes and generates text), connected by a vision-language alignment module.

The key advantage of VLMs over task-specific image classifiers is their ability to leverage the world knowledge and reasoning capabilities encoded in large language models. An LLM that has read thousands of engineering specifications, quality control guidelines, and technical reports can apply that knowledge when analyzing measurement data—something that a task-specific classifier trained on a narrow labeled dataset cannot do.

CLIP (Radford et al., 2021) demonstrated that vision-language representation alignment can be learned from large-scale image-text pairs collected from the internet, enabling zero-shot image classification by matching images to natural language descriptions. BLIP-2 (Li et al., 2023) introduced a lightweight vision-language connector that efficiently aligns frozen vision and language models. These foundational works suggest that VLMs can be adapted to the domain of optical metrology by aligning optical measurement images with domain-specific technical language.

2.3 Multi-Modal Measurement Integration

Optical surface inspection typically involves multiple complementary measurement modalities: thermal imaging reveals subsurface defects and thermal property variations; fringe projection profilometry captures surface topography; structured-light scanning provides high-density 3D geometry. A comprehensive quality assessment requires integrating information across these modalities—recognizing that a thermal anomaly in one location might correspond to a geometric irregularity in another.

Prior studies have demonstrated multi-modal fusion networks for optical metrology tasks (Paper 3) and multi-modal measurement integration for thermal imaging (Huang et al., 2023). The challenge in a VLM framework is to integrate multiple visual input streams (thermal images, phase maps, depth visualizations, defect segmentation overlays) into a unified representation that the language model can reason over.

2.4 Automated Report Generation in Industrial Inspection

Automated report generation in industrial inspection is an emerging application of natural language generation. Prior work in medical imaging has demonstrated that VLMs can generate radiology reports that rival expert-written reports in accuracy and completeness (Khader et al., 2023). The techniques from medical report generation—particularly multi-label classification heads for structured finding extraction and template-guided generation for report structure—are directly applicable to optical surface inspection.

2.5 Literature Synthesis

The convergence of: (1) high-performance numerical optical metrology networks (Huang et al., 2023; Huang et al., 2026), (2) mature vision-language model architectures (CLIP, BLIP-2), and (3) demonstrated automated report generation in analogous domains (medical imaging), creates a clear opportunity for vision-language approaches to optical quality inspection. This study is the first to propose and evaluate a VLM specifically designed for automated optical surface inspection report generation.

---

3. Methodology

3.1 Overall Framework

The proposed framework processes multi-modal optical measurement data through three stages:

Stage 1 — Multi-modal feature extraction. Dedicated task networks (pretrained from Papers 1–3) independently process thermal images, phase maps, and defect detection outputs, producing intermediate feature representations for each modality.

Stage 2 — Vision-language alignment and fusion. A cross-attention fusion module aligns and fuses the multi-modal feature representations into a unified representation compatible with the language model input space.

Stage 3 — Language model report generation. A large language model generates structured natural language inspection reports from the fused representation, conditioned on engineering specification context.

3.2 Multi-Modal Feature Extraction

The feature extraction stage leverages the pretrained task networks from prior studies:

Thermal feature extractor. A U-Net encoder pretrained for thermal image reconstruction (Paper 1) produces a multi-scale feature tensor F_T at five spatial resolutions. The highest-resolution features (1/4 scale) are used for fine-grained thermal anomaly localization.

Phase feature extractor. An RA-U-Net encoder pretrained for phase unwrapping (Paper 2) produces multi-scale phase features F_P, encoding surface geometry and slope information.

Defect feature extractor. A DB-3DFuse network pretrained for defect segmentation (Paper 3) produces defect segmentation masks D and per-defect embedding features F_D, encoding defect type, location, and geometric characteristics.

Each feature tensor is projected to a common dimensional space ($d = 768$) via a learned linear projection, enabling uniform processing by the fusion module.

3.3 Vision-Language Alignment and Fusion

The cross-attention fusion module aligns the multi-modal optical features with the language model's embedding space. Following BLIP-2's Q-Former architecture, a set of query vectors Q attend to the concatenated multi-modal feature sequences via cross-attention:

$$F_{fused} = \text{CrossAttn}(Q, [F_T; F_P; F_D])$$

where $[;]$ denotes channel-wise concatenation and CrossAttn is a multi-head cross-attention layer with 12 attention heads. The resulting fused representation F_{fused} is a sequence of $L = 32$ vectors, each of dimension $d = 768$, which serves as the visual input to the language model.

The Q-Former is trained jointly with the vision-language alignment objective: the fused representation must enable the language model to correctly predict the textual inspection report that corresponds to the input measurement data.

3.4 Language Model Report Generation

A frozen large language model (Vicuna-7B, 7 billion parameters) receives the fused visual representation concatenated with a specification context prompt and generates the inspection report auto-regressively. The specification context prompt encodes:

- The relevant engineering quality specification (e.g., XYZ-2024 optical component acceptance criteria)

- Material and surface type information
- Measurement system calibration parameters

The generated report follows a structured template with three sections:

Section 1 — Defect Description: Natural language descriptions of each detected defect, including type, location within the optical aperture, dimensions, and distinguishing characteristics.

Section 2 — Severity Assessment: Classification of each defect's severity (Critical / Major / Minor / Acceptable) with reference to the applicable specification section, and a qualitative assessment of the defect's impact on optical performance.

Section 3 — Disposition Recommendation: Clear accept/rework/reject recommendation with the specific specification criteria that apply.

3.5 Training Data Construction

A training dataset of 12,000 expert-annotated measurement-report pairs is constructed:

Data sources. Real optical component inspection data from three manufacturing lines (precision lens manufacturing, thermal management component manufacturing, micro-optics manufacturing) is collected, spanning approximately 18 months of production. Each inspection sample includes co-registered thermal images, fringe projection data, and structured-light 3D scans.

Annotation process. Each sample is independently annotated by two certified quality engineers following the company's standard inspection procedure. The annotation produces: defect bounding boxes and masks, defect type labels (crack, pit, scratch, contamination, delamination, other), severity ratings, and disposition decisions. Inter-observer agreement is computed; samples with disagreement are adjudicated by a senior engineer, and the senior engineer's annotation is used as ground truth.

Dataset statistics. The dataset contains 12,000 samples, with the following distribution: 71.4% Accept, 18.3% Minor-defect Accept, 7.2% Rework, 3.1% Reject. The class imbalance reflects actual production quality levels. Average report length is 142 words (range: 23–387 words).

3.6 Training Procedure

The vision-language alignment module (Q-Former and projections) is trained for 10 epochs with a batch size of 64, using a combination of:

Captioning loss: Standard language modeling loss on the report text, with the fused visual features as context.

Multi-label classification auxiliary loss: An auxiliary head predicts the defect type, severity, and disposition directly from F_fused, providing additional training signal for the visual representation learning.

The language model itself remains frozen during this training (following BLIP-2's approach), as unfreezing a 7B parameter model would be computationally prohibitive for most optical metrology deployment settings.

---

4. Simulation Experimental Results

4.1 Evaluation Metrics

Report quality is evaluated at three levels:

Defect classification accuracy: Does the report correctly identify all defects present and correctly classify their types?

Severity grading agreement: Does the severity grade assigned in the report match the expert-annotated severity?

Disposition recommendation accuracy: Does the recommended disposition (Accept / Rework / Reject) match the expert ground truth?

In addition, automated metrics are reported: BLEU-4 and ROUGE-L scores comparing generated reports to reference expert reports at the full-text level.

4.2 Overall Performance

Table 1 presents overall framework performance on the held-out test set (2,400 samples, 20% of the full dataset).

Table 1 Overall inspection report generation performance

Metric	Rule-Based System	VLM (proposed)
Defect classification accuracy (%)	66.4	89.3
Severity grading agreement (%)	61.2	84.7
Disposition recommendation accuracy (%)	68.9	91.2
BLEU-4 score	0.412	0.681
ROUGE-L score	0.531	0.774

The proposed VLM substantially outperforms the rule-based automated reporting baseline across all metrics. The most practically significant improvement is in disposition recommendation accuracy (91.2% vs 68.9%), which directly determines whether parts are correctly accepted or rejected—a decision with direct economic and safety implications.

4.3 Performance by Defect Type

Table 2 presents defect classification accuracy by defect type.

Table 2 Defect classification accuracy by defect type (%)

Defect Type	Rule-Based	VLM (proposed)
Crack	58.7	87.4
Pit	72.3	93.1
Scratch	63.1	88.9
Contamination	78.4	94.2
Delamination	54.2	82.7
Other / Unknown	48.9	76.3

The VLM outperforms the rule-based system on all defect types, with the largest improvements on Crack (+28.7 pp), Delamination (+28.5 pp), and Other/Unknown (+27.4 pp). These are precisely the defect categories where rule-based systems struggle most—cases that deviate from the predefined classification rules or lack clear discriminative features.

4.4 Severity Grading Calibration

An important quality metric for severity grading is calibration: when the VLM outputs "Critical" severity, how often is the ground truth actually Critical? Table 3 presents the confusion matrix for severity grading.

Table 3 Severity grading confusion matrix (%)

	Predicted: Critical	Predicted: Major	Predicted: Minor	Predicted: Acceptable
Actual: Critical	81.3	12.4	4.7	1.6
Actual: Major	9.8	76.2	11.3	2.7
Actual: Minor	3.1	8.4	83.7	4.8
Actual: Acceptable	1.2	3.9	6.7	88.2

The VLM achieves strong calibration on the most important severity levels. Critical defects are correctly identified 81.3% of the time, with 12.4% confused with Major (a forgivable error in a safety context). Acceptable parts are correctly identified 88.2% of the time. The main confusion is between adjacent severity levels—Major vs Minor—which reflects the genuine grading ambiguity that exists in the expert annotation data as well.

4.5 Ablation Study

Table 4 presents an ablation study isolating the contribution of multi-modal fusion versus single-modality input.

Table 4 Ablation: modality contribution to report quality

Input Modality	Defect Accuracy (%)	Disposition Accuracy (%)
Thermal only	74.2	76.8
Phase only	71.9	73.4
Defect masks only	79.4	82.1
Thermal + Phase	83.7	86.3
Thermal + Defect	85.1	88.7
All three (proposed)	**89.3**	**91.2**

Multi-modal fusion consistently outperforms single-modality inputs. The improvement from adding the third modality (whether thermal, phase, or defect masks) is substantial in each case, confirming the complementary nature of the different measurement modalities for comprehensive quality assessment.

4.6 Sample Report Comparison

A qualitative comparison of generated and expert-written reports for a representative sample (a lens surface with two scratches and one pit) illustrates the framework's capabilities. The expert report describes: "Two scratches detected in the central optical zone at 12 o'clock and 3 o'clock positions, measuring 0.4 mm and 0.6 mm in length respectively, depths of 6 μm and 8 μm. One pit detected at 7 o'clock position, 0.3 mm diameter, depth 4 μm. All defects are below the critical threshold per XYZ-2024 Section 4.2.1. Severity: Minor. Recommended disposition: Accept."

The VLM-generated report is: "Inspection identified three surface anomalies: two linear scratch-type defects at approximately 12 and 3 o'clock positions in the clear aperture, lengths approximately 0.35–0.55 mm, estimated depths 5–9 μm; one circular pit-type anomaly at approximately 7 o'clock, approximately 0.3 mm diameter. All detected features are below the critical defect size threshold in XYZ-2024 Section 4.2.1 for this component class. Overall assessment: Minor severity. Recommended disposition: Accept per standard acceptance criteria."

The generated report captures all the key information accurately—defect types, locations, dimensions, severity assessment, and disposition—with appropriate specification references.

---

5. Discussion

5.1 Practical Implications for Manufacturing Quality Control

The proposed framework addresses the manual interpretation bottleneck that has limited the impact of automated optical inspection systems. By generating expert-consistent quality reports directly from measurement data, the framework can reduce inspector workload by an estimated 80–90% (inspectors review and approve generated reports rather than writing them from scratch), while simultaneously improving consistency by eliminating inter-observer variability.

The high disposition recommendation accuracy (91.2%) demonstrates that the VLM has learned to apply engineering specification criteria to measurement data with performance approaching that of experienced human inspectors. This is particularly valuable for borderline cases where the correct disposition depends on nuanced interpretation of measurement data and specification text—a task that is cognitively demanding for humans and that the VLM can perform consistently.

5.2 Relationship to Prior Work

The framework integrates contributions from multiple prior studies in this body of work. From Huang et al. (2023), it takes the 4D thermal imaging methodology that provides rich multi-modal measurement data including surface geometry and thermal response. From Huang et al. (2026), it takes the deep learning architectures for extracting meaningful features from optical measurement data. The VLM framework adds a new capability—automatic interpretation and report generation—that transforms these numerical measurement outputs into actionable quality intelligence.

5.3 Limitations

Several limitations should be noted. First, the VLM is only as reliable as the measurement networks that feed it: errors in thermal reconstruction, phase unwrapping, or defect detection propagate through to the report. The uncertainty quantification framework (Paper 4) could be integrated to flag reports based on uncertain inputs. Second, the current system generates text reports but does not support interactive follow-up questions. An interactive VLM that could answer "What is the estimated impact of this scratch on optical throughput?" would be a valuable extension. Third, the training dataset was constructed from three manufacturing lines; generalization to new product types or material systems may require additional data collection and fine-tuning.

---

6. Conclusion

This paper proposes a vision-language model framework for automated optical surface quality assessment and inspection report generation in precision manufacturing.

The framework integrates multi-modal optical measurement data (thermal images, phase maps, defect segmentation outputs) through a cross-attention fusion module with a large language model, generating structured natural language inspection reports that describe defects, assess severity, and recommend disposition actions.

Simulation experiments demonstrate that the framework achieves expert agreement rates of 89.3% for defect classification, 84.7% for severity grading, and 91.2% for disposition recommendations—substantially outperforming rule-based automated reporting systems. Multi-modal fusion is shown to be essential: combining thermal, phase, and defect detection inputs yields significantly better reports than any single modality alone.

The proposed framework provides a pathway toward fully automated, end-to-end optical quality inspection that closes the loop between measurement data and human-interpretable quality intelligence, enabling quality engineers to focus on exception handling and process improvement rather than routine report writing.

References

Huang, H., Tang, J., Liu, T., & Huang, M. (2026). Precision 3D surface metrology of optical components using stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping. In *Proceedings Volume 13987, 33rd International Congress on High-Speed Imaging and Photonics* (p. 1398704). SPIE. https://doi.org/10.1117/12.3093993

Huang, H., Yang, Y., & Zhu, Y. (2023). Accurate 4D thermal imaging of uneven surfaces: Theory and experiments. *International Journal of Heat and Mass Transfer*, 216, 124580. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124580

Khader, F., Han, S., Bösl, F., Chen, D., Gao, Y., Li, Y., … & Kather, J. N. (2023). ChiMed-GPT, a medical large language model for bridging radiology report and clinical decision. *Nature Medicine*, 29, 2318–2327. https://doi.org/10.1038/s41591-023-02571-6

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 19730–19742). PMLR.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.

- (~5,000 words)*