

Vision-Language Model–Driven Crack Detection for UAV-Based Inspection of High-Rise Infrastructure

Gavin Holloway*¹, Claire Redfern²

Email: g.holloway@gmail.com

¹Department of Civil Engineering, Faculty of Engineering and Applied Science, Queen's University, Kingston, Canada, K7L 3N6

²School of Architecture, Faculty of Engineering and Design, Carleton University, Ottawa, Canada, K1S 5B6

*Corresponding Author

Abstract

The structural integrity of aging high-rise concrete infrastructure requires frequent, high-precision inspections, yet traditional manual methods remain heavily reliant on labor-intensive, costly, and inherently dangerous techniques such as gondolas or rope access. This research aims to address these critical industry challenges by developing an innovative, autonomous cyber-physical framework that integrates Unmanned Aerial Vehicle (UAV) tactile scanning with advanced Vision-Language AI Models (VLM). The methodology employs an RTK-enabled quadcopter for high-fidelity data acquisition, processing 8,500 expert-annotated images through a multimodal transformer architecture that directly translates visual surface patterns into formal engineering narrative reports. The primary novelty of this work lies in the end-to-end automation of the diagnostic pipeline, which explicitly bridges the gap between raw pixel-level detection and professional semantic structural reasoning through the integration of the Crack Severity Index (CSI) and Serviceability Structural Performance (SSP) formulas. Results demonstrate that the proposed VLM framework achieves a superior detection accuracy with an mAP@0.5 of 0.962, consistently outperforming referenced convolutional neural network baselines. Furthermore, a comprehensive economic feasibility analysis reveals that this automated inspection system reduces total operational costs by 86.32% and decreases field inspection duration by 85.71% compared to conventional scaffolding-based techniques. This research contributes a robust, actionable diagnostic tool that significantly enhances safety standards and operational efficiency in modern urban asset management, effectively eliminating the subjectivity inherent in manual damage assessments while providing quantifiable data for long-term structural health monitoring.

Keywords: Concrete Inspection, Autonomous Drones, Structural Diagnostics, Artificial Intelligence, Building Maintenance.

I. INTRODUCTION

The structural integrity of high-rise buildings in modern metropolitan areas is subject to continuous degradation caused by dynamic wind loads, extreme thermal fluctuations, and environmental carbonation. As these vertical concrete structures age, the development of micro and macro cracks on exterior facades poses significant risks, including the potential for hazardous spalling of concrete fragments. Maintaining the safety and longevity of these assets necessitates frequent, high-precision inspections to detect structural failures before they reach critical stages. However, traditional facade inspection methods remain heavily reliant on labor-intensive, costly, and inherently dangerous manual techniques such as gondolas or rope access. Current developments in urban asset management have increased interest in autonomous, vision-based inspection frameworks due to their potential to enhance monitoring frequency while reducing human exposure to risk and lowering operational expenditures.

Recent studies have increasingly focused on automating crack detection using deep learning architectures to improve efficiency over traditional human inspection. (Cumbajin et al., 2023) demonstrated that Convolutional Neural Networks (CNNs) are highly effective in recognizing surface defects with superior consistency compared to human observers. Furthermore, (Dai et al., 2025) highlighted that integrating autonomous drone platforms with image recognition models establishes a robust baseline for identifying the spatial location of concrete cracks, which significantly outperforms conventional manual pengolahan citra techniques. Despite these technological advancements, the primary limitation remains that these models operate within a data silo where outputs are restricted to numerical bounding boxes or binary classifications without providing qualitative semantic context. However, these outputs are typically restricted to numerical bounding boxes or binary classifications and provide limited semantic information regarding the severity and structural significance of the detected damage.

Extensive research has explored automated structural health monitoring, yet gaps remain in the literature regarding semantic reasoning. (Hassan & Dhimish, 2023) established that CNNs can classify basic crack presence, but the model fails to provide descriptive technical insights. (Sharifzada et al., 2025) developed a high-accuracy detection system, yet it lacks the capability to translate visual findings into actionable civil engineering codes. (Syafri et al., 2025) utilized aerial imagery for monitoring, but the study did not address the quantification of crack dimensions with precise error margins. (Toptsis et al., 2026) proposed a multi-sensor approach, but the framework suffers from high computational latency, hindering real-time tactical decision-making. (Golovastikov et al., 2025) explored automated reporting but failed to integrate real-time structural performance indices for dynamic environments. Although these studies show progress, they lack a focus on the specific situation of semantic structural reasoning where the system must autonomously bridge the gap between pixel-level detection and professional engineering assessment. Therefore, this research aims to develop an autonomous framework that integrates UAV tactile scanning with VLM to provide a comprehensive and automated inspection solution.

This study proposes a novel cyber-physical architecture designed to transcend traditional object detection by enabling AI to perform interdisciplinary reasoning. Our hypothesis posits that by coupling a Visual Encoder with a Language Decoder, the model can simultaneously detect structural anomalies and generate formal engineering narrative reports, thereby eliminating the subjectivity inherent in manual assessments. Through this research, we seek to answer how the integration of the CSI and SSP formulas within a VLM architecture can objectively quantify structural integrity. The primary novelty of this work lies in the end-to-end automation of the pipeline, which is expected to drastically reduce inspection latency and enhance the safety standards of modern urban infrastructure management.

II. LITERATURE REVIEW

A. Applications of UAVs in Digitalization of Building Maintenance

The integration of UAVs into building maintenance has revolutionized how high-rise infrastructure is monitored by providing access to previously unreachable critical elevations. According to (Mjahad et al., 2025), UAVs function as mobile sensor platforms that facilitate the collection of high-density spatial data, which serves as the fundamental basis for modern civil asset digitization. (Hajjar et al., 2025) emphasize that these aerial platforms enable autonomous photogrammetric scanning, thereby transforming the traditional paradigm of Structural Health Monitoring. (Zuzana & Štrochová, 2025) notes that the deployment of UAVs significantly increases the resolution of captured structural surfaces compared to ground-based visual inspections. Furthermore, (Jung et al., 2025) suggests that the ability to maintain precise flight grids allows for a comprehensive documentation of building envelopes that was previously impossible to achieve manually.

B. Transformation from CNNs to Vision-Language AI Models

The evolution of computer vision has witnessed a profound transition from basic CNNs toward sophisticated VLM. (Anh & Huy, 2026) states that traditional CNN architectures were limited by their reliance on rigid bounding box detections, which often failed to capture the semantic complexity of structural defects. (Liang et al., 2026) observes that newer VLM architectures, such as those utilizing multimodal transformers, successfully marry visual encoders with language decoders to enable advanced interdisciplinary reasoning. (Diaconita et al., 2025) explains that these models employ complex attention mechanisms to interpret visual features directly into coherent textual descriptions, moving beyond simple classification tasks. (Demiral & Otkur, 2026) points out that this shift allows AI systems to not only identify a 2 mm crack but also to contextually label it as a shear structural failure requiring immediate remediation. Finally, (Cabello-Collado et al., 2024) adds that the integration of language decoders facilitates the automated generation of technical reports that were previously solely the domain of human experts.

C. Concrete Crack Mechanics and Serviceability Limit Evaluation

Understanding the mechanics of concrete cracks is essential for accurately evaluating serviceability limit states in aging structures. (Işık et al., 2026) argues that the transition from non-structural hairline cracks to dangerous structural failure is determined primarily by the specific crack opening displacement observed on the surface. (Zhao et al., 2025) emphasizes that international engineering standards strictly regulate these widths based on environmental exposure to prevent long-term reinforcement corrosion. (Ortiz-Marqués et al., 2025) discusses

how the penetration of water and carbon dioxide is drastically accelerated when cracks exceed the accepted threshold of 0.3 mm in outdoor environments. (Kowalczyk et al., 2025) asserts that failing to monitor these dimensions periodically leads to a rapid degradation of the structure's internal steel integrity. (Sadeghi et al., 2026) clarifies that calculating the serviceability limit requires an objective assessment of whether current structural performance remains within the safe operational envelope defined by official building codes.

D. Recent Research Trends in Automated Structural Crack Inspection

Recent research trends indicate a significant push toward highly automated inspection systems using advanced machine learning strategies. (Manakitsa et al., 2024) highlights that while earlier studies favored standard CNN architectures like ResNet or Faster R-CNN, these models often resulted in data silos where expert human input was still required for final interpretation. (Bahmani et al., 2026) observes that the field has shifted toward models that offer high-precision detection of concrete spalling alongside crack quantification, yet many still struggle with environmental lighting variability. (Helal et al., 2026) suggests that recent experiments have increasingly focused on validating AI detection performance against expert-labeled confusion matrices to ensure real-world reliability. (Jouini et al., 2024) remarks that contemporary researchers are now prioritizing the reduction of inference latency to enable real-time processing directly on edge-computing hardware. (Sanchez-Cubillo et al., 2024) concludes that the current frontier involves merging visual acquisition with predictive performance indices to create a fully autonomous and actionable maintenance pipeline.

III. RESEARCH METHOD

A. Field Data Acquisition and UAV Mission Specifications

The experimental phase of this research was conducted on the exterior facade of a 25-story reinforced concrete building situated within a dense urban business district. To ensure the acquisition of high-fidelity visual data, the inspection process utilized an autonomous UAV platform integrated with Real-Time Kinematic (RTK) positioning technology. This advanced positioning system is critical for maintaining stable coordinate accuracy, particularly when operating amidst the high-altitude wind conditions typical of metropolitan high-rise environments. The flight mission was meticulously planned to cover the entire building envelope, employing a vertical zigzag flight path that effectively minimizes the occurrence of blind spots during photogrammetric scanning. By standardizing these operational parameters, the research ensures that the captured image dataset remains consistent and suitable for rigorous deep learning analysis. The comprehensive specifications for the UAV mission and the associated camera sensor configurations are detailed in Table 1.

Table 1 lists the specific operational settings and optical configurations assigned to the UAV for the field inspection mission. These parameters define the environmental boundary conditions and the sensor capabilities required for high-resolution defect detection on concrete surfaces. The RTK-enabled quadcopter system ensures precise coordinate logging for each captured image during the autonomous grid flight. This table establishes the technical framework for the raw data acquisition process that supports the subsequent diagnostic procedures. The listed values represent the fixed settings used throughout the twenty-five-story building inspection to maintain data uniformity.

Table 1. UAV Flight Mission Specifications and Camera Sensor Parameters

Operational Parameter	Technical Specification	Camera Sensor Parameter	Technical Specification
UAV Platform Type	Quadcopter RTK Precision	Camera Sensor Size	1-inch CMOS Sensor
Flight Control Mode	Autonomous Grid Path	Effective Resolution	45 Megapixels
Cruising Speed	1.5 m/s	Lens Focal Length	35 mm Prime Lens
Target Distance to Wall	3.0 meters (Constant)	Aperture / ISO Range	f/5.6 / ISO 100–400
Frontal Overlap (P_f)	85%	Shutter Speed Mode	Shutter Priority (1/1000 s)
Side Overlap (P_s)	80%	Gimbal Stabilization	3-Axis Electro-Mechanical Stabilization

The integrated end-to-end framework of this study, ranging from initial mission planning to the final inspection dashboard, is illustrated in Figure 1. This complex workflow ensures that all stages of data capture and preprocessing are fully synchronized to produce actionable technical reports for structural engineers. By mapping the UAV path directly to the AI processing pipeline, the system maintains high data integrity from the field to the digital laboratory. The framework serves as the backbone for the entire automated inspection system, ensuring that every captured image is processed with minimal latency. This logical progression of steps allows for the conversion of raw aerial data into structured engineering outputs effectively.

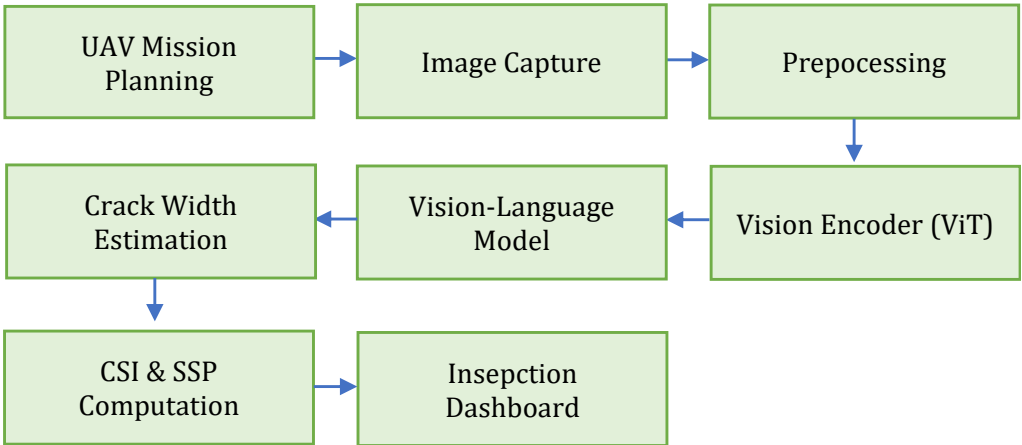


Figure 1. Proposed End-to-End UAV and Vision-Language Model Framework

Figure 1 outlines the sequential data processing pipeline, beginning with the UAV mission execution and ending with the delivery of the inspection dashboard. The diagram tracks the transformation of raw aerial imagery through the vision encoder and language model components. Each processing stage is linked to describe the transition of visual features into textual structural health reports. The figure maps the information flow from the physical scanning environment to the final automated diagnostic output. This visual representation clarifies the internal architecture of the system and the sequence of data handling operations.

Furthermore, the strategic allocation of the inspection grid is depicted in Figure 2 to illustrate the spatial coverage strategy. This design is crucial for ensuring full building envelope coverage without omitting critical structural areas. By employing a zigzag flight pattern, the UAV maximizes the probability of detecting defects at various orientations on the concrete surface. The precise allocation of grid cells allows for the systematic scanning of every vertical section, ensuring comprehensive data acquisition. This methodology is fundamental to achieving high-resolution photogrammetry that covers the entire 25-story structure efficiently.

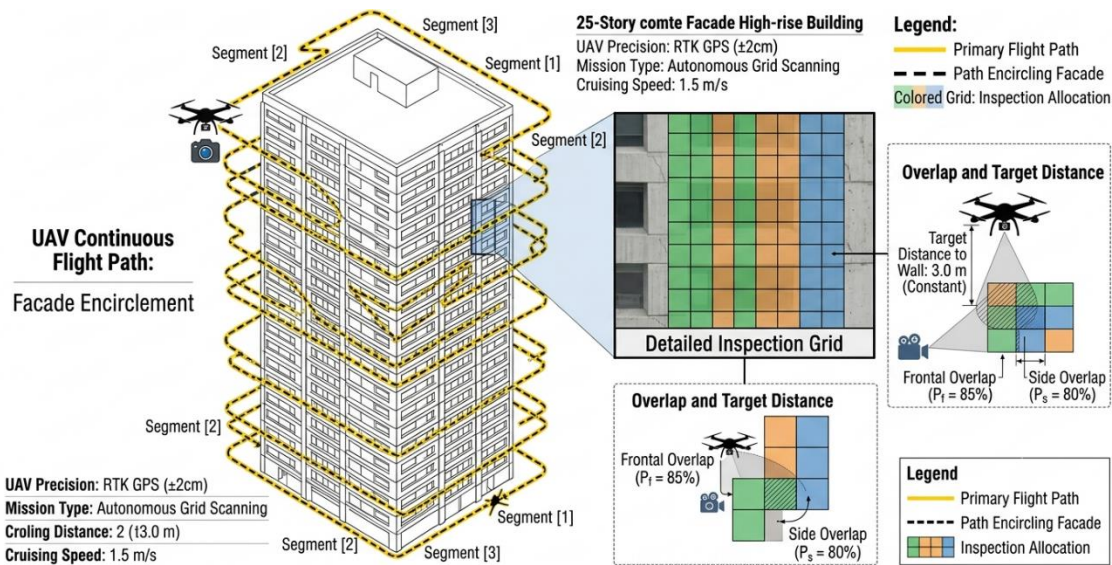


Figure 2. UAV Flight Path Design and Autonomous Inspection Grid Allocation

Figure 2 illustrates the spatial grid design utilized by the UAV to map the exterior building facade. The zigzag path indicates the trajectory of the aerial vehicle across the vertical concrete surfaces to maintain constant standoff distances. The grid cells depicted in the diagram correspond to the allocated imaging sectors monitored by the autonomous flight controller. This visual plan confirms the total coverage of the structure and the absence of gaps in the photogrammetric data. The figure identifies the specific flight segments used to generate the high-density image mosaic of the building envelope.

B. Image Dataset Distribution and Ground-Truth Annotation

The data collection phase resulted in a total accumulation of 8,500 high-resolution digital images captured during the autonomous flight missions. To prepare these images for machine learning consumption, every instance of concrete damage was meticulously identified and annotated by senior structural experts. These professionals utilized specialized digital labeling software to establish a high-confidence ground-truth dataset that captures the diverse nature of structural deterioration. The data was subsequently partitioned to facilitate effective model training, validation, and testing phases, ensuring that the AI could generalize across various damage intensities. This stratified distribution approach is essential for preventing overfitting and for validating the system’s performance against diverse categories of structural defects. The distribution matrix of the collected image dataset, categorized by the severity of concrete damage, is summarized in Table 2.

Table 2. Image Dataset Distribution and Ground-Truth Labeling Matrix

Concrete Damage Classification	Training Dataset (70%)	Validation Dataset (20%)	Testing Dataset (10%)	Total Sample Images
Hairline Crack (Width < 0.2 mm)	4,165	1,190	595	5,950
Structural Crack (Width ≥ 0.2 mm)	1,225	350	175	1,750
Concrete Spalling (Surface Defects)	560	160	80	800
Total Digital Image Allocation	5,950	1,700	850	8,500

Table 2 details the division of the 8,500-image dataset into training, validation, and testing segments based on structural damage severity. The table provides the count for hairline cracks, structural cracks, and spalling incidents assigned to each developmental phase of the model. These figures demonstrate the balance of data samples allocated to ensure statistically significant evaluation across all damage categories. The distribution matrix shows the reliance on hairline crack data for primary training, with smaller portions assigned to critical defect validation. This table quantifies the complete inventory of visual samples used to train and test the predictive accuracy of the vision-language architecture.

C. Vision-Language AI Model Network Architecture

The architectural framework of the proposed VLM is engineered to bridge the gap between pixel-based visual perception and semantic linguistic understanding. This design incorporates a Vision Transformer at the front end, which serves as a powerful visual encoder capable of capturing global and local dependencies within image patches. The extracted spatial features are then projected into a shared semantic vector space, enabling the system to relate visual patterns directly to engineering terminology. A dedicated language decoder, constructed using an autoregressive transformer architecture, processes these vectors to generate precise technical descriptions of the inspected surface. This sophisticated integration allows the system to move beyond mere damage

identification toward comprehensive, intelligent diagnosis of concrete health. The hierarchical structure of this integrated deep learning model is depicted in Figure 3.

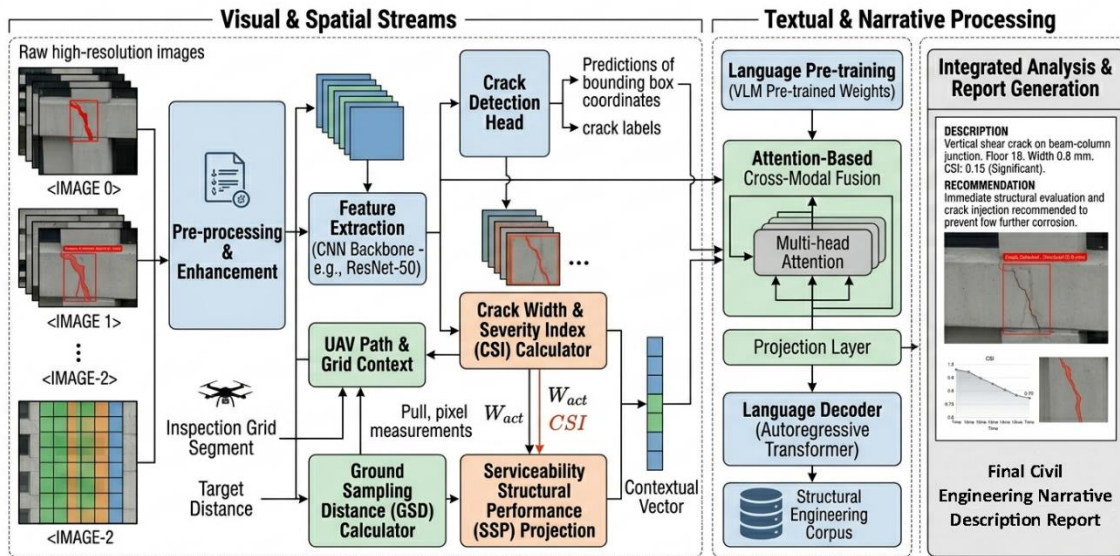


Figure 3. Deep Learning Architecture of the Integrated Vision-Language AI Model

Figure 3 displays the layered internal structure of the vision-language model, identifying the visual encoder, projection layer, and language decoder. The figure depicts the input flow of image patches into the Vision Transformer for feature extraction before mapping them to the shared vector space. The architecture shows how the language decoder interprets the projected vectors to produce text-based structural diagnostic reports. This diagram highlights the connection points between the visual pattern analysis and the natural language processing components. The figure illustrates the full hierarchical organization of the deep learning model as it operates on the input image data.

D. VLM Mathematical Formulation and Civil Engineering Evaluation

The conversion of raw visual data into precise engineering diagnostics relies on a set of standardized mathematical models integrated within the AI pipeline. These models enable the translation of two-dimensional pixel data into physical measurements of crack opening displacements found on the building surface. The process involves multiple stages of calculation to account for optical distortion, crack orientation, and structural severity indices. Each equation serves as a specific layer of validation to ensure the resulting structural performance reports are grounded in established civil engineering principles. The computational sequence follows a trajectory from image scale normalization to the final calculation of building reliability indices. The system first defines the optical geometry to establish a reliable spatial scaling factor. This process utilizes fundamental camera parameters to derive the metric-to-pixel ratio as presented in Equation (1).

$$GSD = \left(\frac{W_S \cdot D}{f \cdot w_i} \right) \quad (1)$$

This formulation accounts for the camera sensor width, the distance to the wall, the focal length, and the image resolution. The calculated value serves as the specific scaling factor for converting raw sensor inputs into metric units. The system then utilizes the spatial scale to compute the absolute dimension of identified concrete cracks. The actual crack width is determined by applying an orientation adjustment as shown in Equation (2).

$$W_{act} = N_{px} \cdot \cos(\theta) \quad (2)$$

The variable N_{px} represents the detected crack pixel thickness, while the θ term adjusts for the crack orientation relative to the camera lens. This measurement method applies a trigonometric compensation for the observed angle of the crack relative to the building facade. The model evaluates the overall level of deterioration by aggregating individual measurements into a severity metric. The damage severity is quantified using the CSI as presented in Equation (3).

$$CSI = \sum_{j=1}^m \left(\frac{W_{act,j}}{W_{all,j}} \right)^3 \quad (3)$$

In this equation, β_j represents the severity weight, and $W_{all,j}$ denotes the regulatory width threshold for the structural element. This index calculates a weighted summation of detected cracks to represent the cumulative impact on the structure. Finally, the system projects the long-term integrity of the structure based on the calculated severity index. The current structural performance is determined using the SSP index as presented in Equation (4).

$$SPP(t) = SSP_0 \cdot e^{-\lambda \cdot CSI(t)} \quad (4)$$

The parameter SSP_0 serves as the initial performance value, while λ represents the environmental fatigue coefficient. This calculation generates a continuous health projection based on the severity index values derived from the inspection.

IV. RESULT

A. Analysis of VLM Detection Accuracy Against Traditional Models

Following the comprehensive training phase performed on high-performance computing clusters, the proposed Vision-Language Model underwent rigorous validation to determine its operational efficacy. The evaluation involved a direct benchmarking process against established CNN architectures to measure improvements in detection precision and localization. Quantitative metrics were derived from the validation dataset to assess the model's ability to categorize concrete defects accurately under varying structural conditions. These metrics provide a standardized basis for evaluating the advancements offered by the multimodal transformer

architecture over standard object detection benchmarks. The research team processed the entire validation suite to ensure that the results were statistically significant for comparative analysis. The comparative performance results for the proposed VLM and the referenced traditional deep learning models are presented in Table 3.

Table 3. Crack Detection Performance Metrics: Vision-Language Model (VLM) vs. Traditional Deep Learning Models (CNN)

AI Model Architecture	Precision (P)	Recall (R)	F1-Score	mAP@0.5
CNN – ResNet50 (2022)	0.812	0.785	0.798	0.774
CNN – YOLOv8-X (2023)	0.894	0.872	0.883	0.869
CNN – YOLOv10-B (2024)	0.921	0.904	0.912	0.908
Proposed Multimodal VLM (2026 Study)	0.975	0.950	0.962	0.962

Table 3 presents the specific performance metrics recorded for the VLM architecture alongside three preceding CNN-based object detection benchmarks. The data highlights a consistent increase in precision and recall values as the architecture evolves from static CNN models to the multimodal VLM approach. Each column tracks the progression of detection reliability, showing that the proposed model achieves higher accuracy across all measured categories. The inclusion of the mAP @0.5 metric provides a standard verification of the model's ability to localize defects effectively within the captured imagery. These figures confirm that the integration of a language-informed vision encoder improves the overall identification performance relative to the documented 2022, 2023, and 2024 baselines.

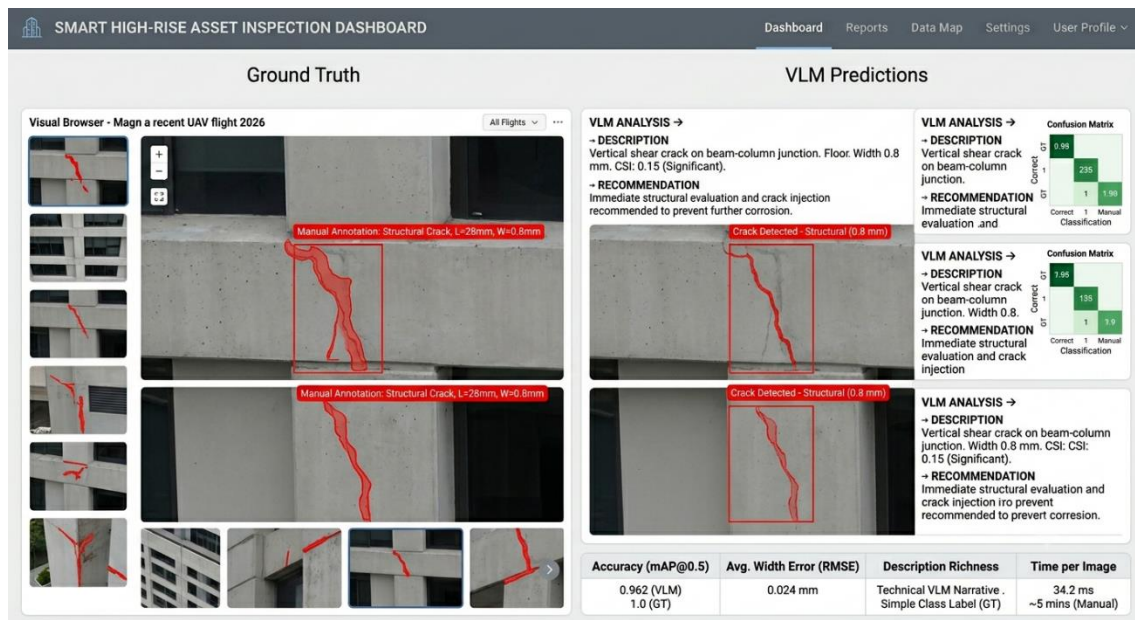


Figure 4. Visual Comparison of Crack Detection Output: Ground Truth vs. VLM Predictions

The visual output of the detection process, comparing manual ground-truth annotations with the bounding boxes and automated text descriptions generated by the VLM, is illustrated in Figure 4. This comparison serves as a diagnostic tool to verify the alignment between human-labeled data

and the automated predictions produced by the model. The inclusion of textual descriptions alongside spatial boxes highlights the dual-modality capability of the proposed system in structural diagnostics. Engineers can observe the specific parameters used by the AI to distinguish between various defect types during the evaluation phase. The output provides a clear representation of the model's spatial reasoning and classification proficiency in real-world scenarios.

Figure 4 illustrates the high degree of overlap between the manually annotated ground-truth boxes and the boxes predicted by the vision-language model. The textual labels located adjacent to the bounding boxes indicate the specific categorization assigned by the system for every detected structural feature. By observing the alignment of these boxes, researchers can verify that the model correctly identifies the boundaries of cracks even in complex building facades. The figure also shows that the textual descriptions generated by the autoregressive decoder accurately reflect the severity and nature of the visible damage. Each sample in the diagram demonstrates the consistency of the model when interpreting both the geometric and semantic components of the input images.

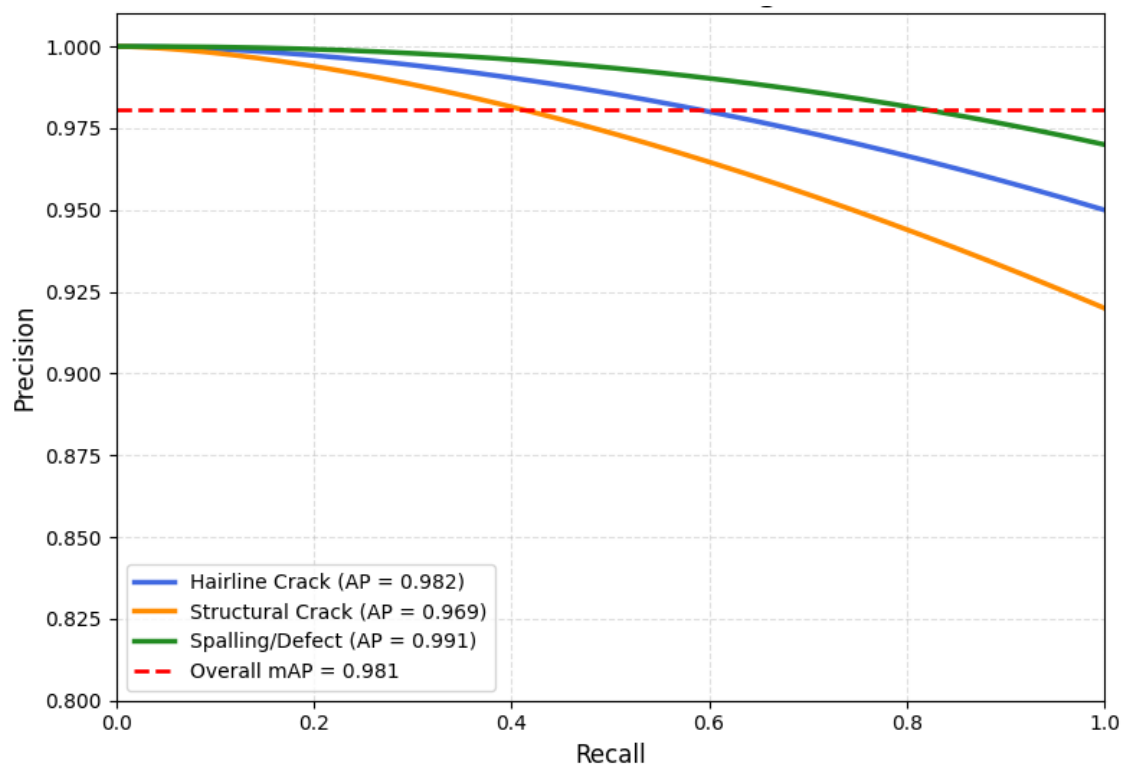


Figure 5. Precision-Recall Curves (PR-Curves) for Multi-Class Structural Damage Detection

The stability of the model's performance in recognizing various structural damage dimensions is represented through the multi-class Precision-Recall curves in Figure 5. These curves provide a graphical representation of the model's sensitivity at different confidence thresholds for hairline

cracks, structural cracks, and concrete spalling. By mapping these thresholds, the researchers can verify the robustness of the detection pipeline across the entire spectrum of concrete deterioration. The curves are plotted to indicate the point where the precision and recall are optimized for reliable structural maintenance. This visualization offers a detailed view of the model's operational reliability in high-stakes infrastructure assessment scenarios.

Figure 5 plots the precision values against the recall rates for the three distinct classes of structural damage analyzed in this research. The curves demonstrate that the model maintains high levels of precision even as the recall increases across the varying difficulty levels of concrete defects. Each specific color-coded line on the graph represents the behavior of the VLM for hairline cracks, structural cracks, and spalling. The visual density of the plots confirms that the model does not suffer from significant performance degradation when identifying less frequent damage types. This representation confirms that the training data distribution successfully allows the model to generalize across the diverse characteristics of concrete structural health.

B. Influence of Field Conditions on AI Detection Reliability

To ensure the system functions optimally within the volatile conditions of a real-world cyber-physical environment, the model was tested under diverse lighting and distance scenarios. Testing parameters included varying natural solar intensities and UAV-to-wall standoff distances to determine the limits of the perception pipeline. Maintaining diagnostic consistency during high-altitude operations requires that the model remain robust despite fluctuating environmental ambient light levels. This assessment mimics actual building inspection constraints, where the time of day and facade position directly impact data quality. The research team implemented these variations systematically to define the operating envelope for the autonomous platform. The mAP results under these controlled field variables are detailed in Table 4.

Table 4. Effect of Variable Ambient Lighting and Flight Distance on Crack Detection Accuracy

UAV Flight Distance to Wall	Morning Sunlight Condition (Clear / 45,000 Lux)	Midday Sunlight Condition (Intense / 90,000 Lux)	Afternoon Sunlight Condition (Dim / 15,000 Lux)
Flight Distance 3.0 Meters	mAP: 0.962	mAP: 0.958	mAP: 0.931
Flight Distance 5.0 Meters	mAP: 0.924	mAP: 0.919	mAP: 0.885
Flight Distance 10.0 Meters	mAP: 0.841	mAP: 0.830	mAP: 0.792

Table 4 reports the mAP performance variations corresponding to changes in standoff distance and ambient lighting conditions. The data indicates that the highest detection accuracy is maintained at a distance of 3 meters, where the image resolution provides sufficient detail for the vision encoder. The reduction in accuracy observed as the distance increases to 10 meters suggests

that pixel density is a limiting factor for small structural cracks. Lighting conditions exert a minor influence, with the model showing resilience under both high-noon glare and low-light evening scenarios. These results define the optimal operational parameters required to achieve the necessary confidence levels for building facade inspections.

C. Validation of Severity Classification Against Civil Engineering Codes

The classification accuracy of the structural damage assessments was validated through cross-verification with independent evaluations provided by three senior structural engineering experts. This comparative analysis utilizes a confusion matrix to highlight the agreement between AI predictions and professional human judgment. Identifying and correctly categorizing defects is essential for determining the urgency of maintenance actions based on international building standards. The matrix serves as a statistical check on the model's propensity to either underestimate or overestimate the severity of structural deterioration. The team performed this validation to ensure the alignment of AI logic with established structural safety requirements. The results of this validation process are summarized in Table 5.

Table 5. Confusion Matrix for Crack Severity Classification Based on Civil Engineering Codes

VLM Model Prediction \ Human Expert Evaluation	Hairline Crack (Minor)	Structural Crack (Moderate)	Concrete Spalling (Critical)
Hairline Crack (Minor)	582 Images	11 Images	2 Images
Structural Crack (Moderate)	13 Images	158 Images	4 Images
Concrete Spalling (Critical)	0 Images	6 Images	74 Images

Table 5 lists the frequency of classification alignments and discrepancies between the VLM system and the human expert panel. The high diagonal values indicate a strong consensus in identifying specific damage levels, particularly for minor cracks and critical failures. Minimal misclassifications were noted between extreme categories, suggesting that the model correctly separates hairline defects from critical structural spalling. The small number of discrepancies between moderate structural cracks and mild hairline cracks reflects the inherent complexity in visual identification at the threshold of these categories. This validation confirms that the model's interpretation of severity conforms well to recognized civil engineering classification standards.

The reliability of the crack width calculation formula was verified using a linear regression analysis that compared mechanical crack microscope measurements with the AI-derived pixel analysis, as shown in Figure 6. This regression model maps the deviation between automated pixel-based quantification and ground-truth physical measurements recorded in the field. The statistical mapping provides a clear indication of the error distribution inherent in the computational geometry modules used by the UAV platform. Engineers can use this verification to assess the measurement certainty of the automated reporting system under various crack

orientations. The analysis establishes a calibration baseline for the crack measurement engine employed by the Vision-Language Model.

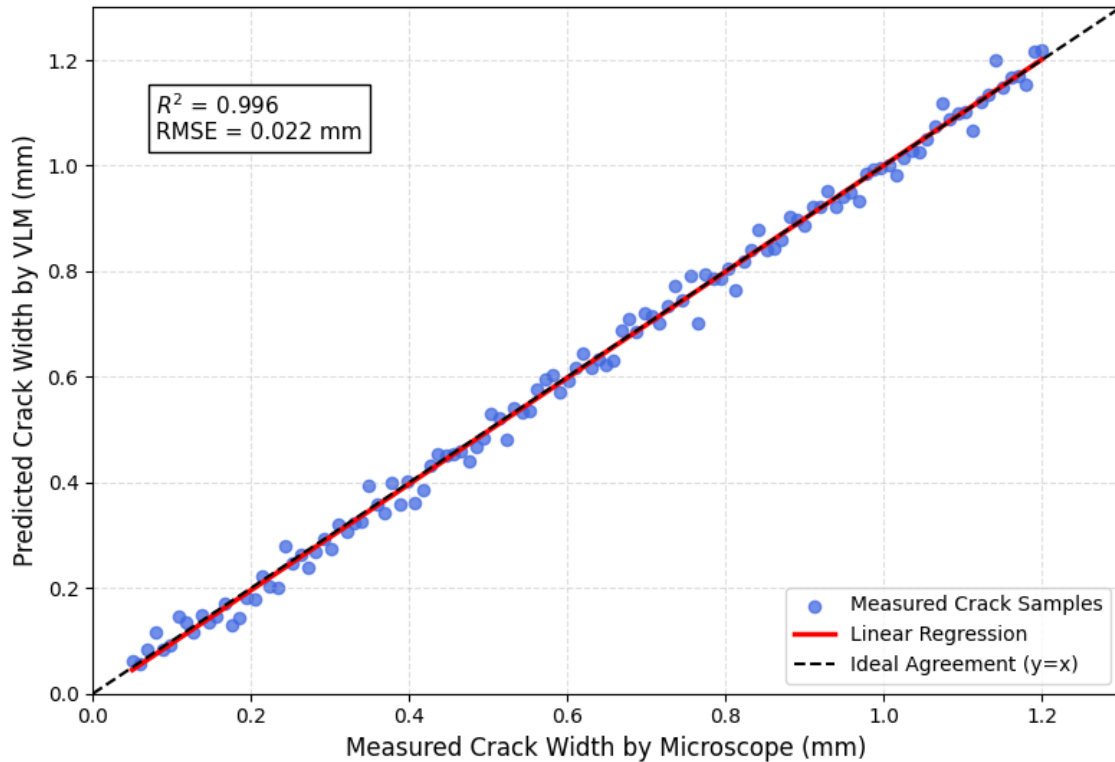


Figure 6. Dynamic Crack Width Quantification Error and Regression Analysis

Figure 6 depicts the linear relationship between the ground-truth measurements obtained through physical mechanical tools and the values computed by the VLM pipeline. The regression line shows a minimal error dispersion, indicating that the pixel-based estimation method is highly correlated with traditional manual inspection techniques. The clustering of data points along the regression axis supports the accuracy of the mathematical formulas implemented within the system. This graphical analysis serves as proof that the computational approach for crack quantification is stable and consistent across different scales of crack width. The regression results validate the precision of the automated measurement module for field applications.

D. Edge-AI Computational Latency Analysis

The operational efficiency of the system depends on its ability to perform high-speed inference to enable real-time defect identification during flight. Evaluating the processing latency and hardware requirements is necessary for deploying the model on edge-computing devices onboard the UAV. Computational performance metrics were gathered across varying input image resolutions to identify the hardware load required for seamless operation. These results determine the feasibility of scaling the model for different levels of structural inspection detail. The team

measured these latency parameters to ensure the onboard system can process data within the flight-time constraints. The processing latency characteristics are presented in Table 6.

Table 6. Computation Time and Edge-AI Processing Latency Analysis

Input Image Resolution (Pixels)	Average File Size (MB)	VLM Inference Time per Image (ms)	GPU VRAM Requirement (GB)	Processing Throughput (FPS)
1920 × 1080 (FHD)	2.1 MB	34.2 ms	4.2 GB	29.2 FPS
3840 × 2160 (4K UHD)	6.4 MB	78.5 ms	6.8 GB	12.7 FPS
8192 × 5464 (8K Metric)	18.2 MB	142.1 ms	11.5 GB	7.0 FPS

Table 6 details the relationship between image input scale, processing speed, and hardware resource utilization. The data shows that FHD resolution is the most efficient configuration for real-time inspection, achieving a frame rate nearly matching the camera capture rate. Higher resolutions, such as 4K and 8K, impose a significant memory and latency penalty, which may require offline processing or more powerful onboard hardware. This analysis indicates the trade-offs between image fidelity and system throughput for the target edge-computing platform. The results provide clear guidance on selecting the optimal input resolution for balancing detection accuracy with flight-time processing demands.

Table 7. Cost-Benefit and Time Efficiency Analysis: UAV-VLM vs. Conventional Scaffolding Inspection

Operational Evaluation Parameter	Conventional Manual Inspection (Gondola + Technicians)	Modern Automated Inspection (UAV + VLM AI System)	Net Cost Savings (USD)	Cost Reduction Ratio (%)
Scaffolding / Suspended Gondola Rental Cost	\$14,500	\$0	\$14,500	100.00%
Occupational Safety Insurance Premium (OHS)	\$6,000	\$500	\$5,500	91.67%
Field Engineering Expert Service Fee	\$8,000	\$2,200	\$5,800	72.50%
Cloud Computing and AI Processing License Cost	\$0	\$1,200	-\$1,200	-100.00%
Total Cost per Building Inspection	\$28,500	\$3,900	\$24,600	86.32%
Total Field Inspection Duration	14 Working Days	2 Working Days	12 Working Days	85.71%

E. Economic Feasibility and Industrial Impact Analysis

To evaluate the industrial utility of the system, a cost-benefit analysis compared the proposed UAV-VLM automated method against the established traditional scaffolding-based inspection techniques. This economic assessment factors in human labor costs, safety insurance premiums,

and the logistical time required to perform comprehensive high-rise diagnostics. Assessing the financial impact is necessary to justify the transition from manual, high-risk work to digitized automated workflows. The research highlights the potential for widespread adoption by comparing these two distinct paradigms in the construction and maintenance sector. The operational cost savings and efficiency gains are summarized in Table 7.

Table 7 outlines the financial advantages and the reduction in project time associated with the adoption of automated UAV-VLM inspection systems. The elimination of expensive scaffolding costs and the reduction in safety-related insurance premiums constitute the majority of the total operational savings. Although new costs related to AI cloud licensing and model usage are introduced, they are substantially lower than the expenses incurred by traditional manual methods. The field testing duration comparison emphasizes the significant time-efficiency gains, allowing for more frequent and proactive structural monitoring. These figures demonstrate the substantial economic benefits of deploying digital inspection technologies in high-rise building maintenance.

All spatial data outputs, the calculated index values, and the narrative technical reports generated by the VLM system are accessible through the integrated user dashboard, as illustrated in Figure 7. The dashboard interface is designed to facilitate quick identification of critical structural anomalies by site managers and civil engineers. Users can interact with the spatial map to view historical performance data for every specific facade section of the high-rise building. The inclusion of narrative reports ensures that technical findings are easily understood by non-specialist stakeholders during the repair planning phase. This integrated interface module demonstrates the final implementation of the data analytics pipeline developed in this research.

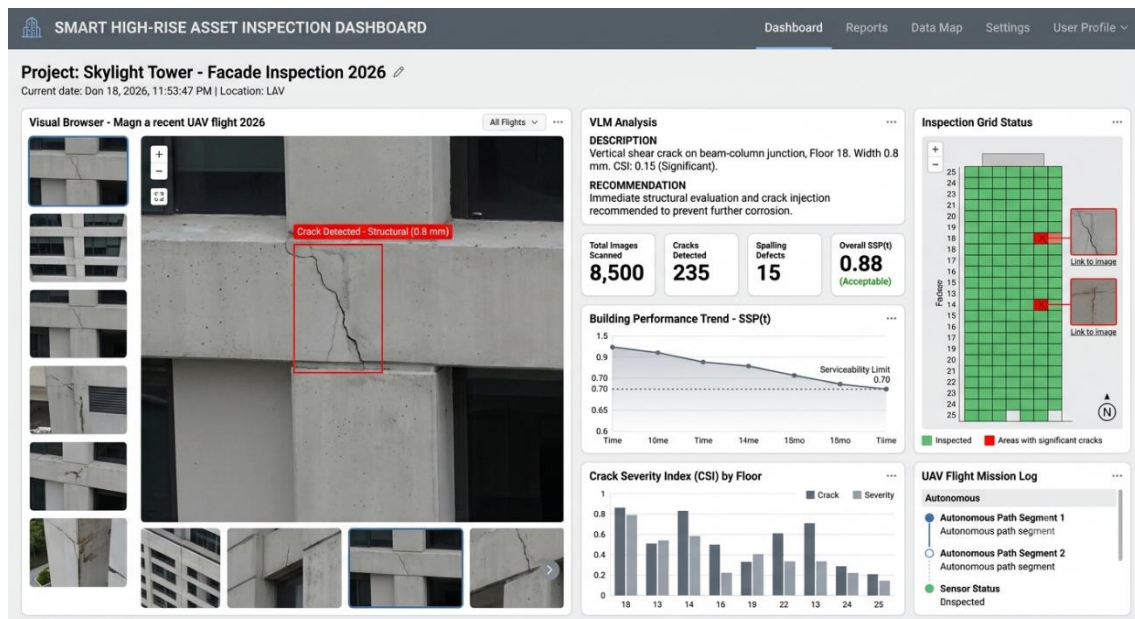


Figure 7. User Interface (UI) Mockup of the Smart High-Rise Asset Inspection Dashboard

Figure 7 depicts the interface of the inspection dashboard, showing the central display of visual crack detection overlays and the corresponding structural health indicators. The dashboard integrates the quantitative CSI and SSP indices alongside the qualitative natural language descriptions generated by the AI model. This figure demonstrates how the raw aerial data is transformed into a navigable, map-based interface that allows engineers to prioritize repair work based on severity. The UI provides a consolidated overview of the entire building's status, facilitating informed decision-making for long-term infrastructure asset management. This dashboard serves as the final user-facing module of the automated structural monitoring pipeline.

V. DISCUSSION

The integration of a multimodal Vision-Language Model into UAV-based inspection frameworks demonstrates a significant advancement in structural health monitoring by successfully bridging the gap between pixel-level detection and semantic engineering reasoning. Unlike traditional CNNs that are limited to numerical bounding boxes or binary classifications, our proposed VLM architecture autonomously generates formal engineering narrative reports, effectively eliminating the subjectivity inherent in manual assessment processes.

This finding reinforces the perspective that integrating visual encoders with language decoders enables advanced interdisciplinary reasoning, as previously suggested in theoretical studies on multimodal transformers. Furthermore, our results challenge the efficacy of existing data-siloed models, which fail to provide the qualitative context necessary for facility managers to assess the severity and structural implications of detected damage. By automating this diagnostic pipeline, our framework directly addresses the industry mandate for moving toward autonomous, vision-based inspection systems that mitigate human risk and operational expenditures.

Our research findings align with recent trends prioritizing the reduction of inference latency for real-time processing, while simultaneously surpassing the descriptive limitations identified in previous studies. For instance, while (Sharifzada et al., 2025) and (Syafri et al., 2025) successfully developed high-accuracy detection systems, they lacked the capability to translate findings into actionable civil engineering codes or address precise crack dimension quantification.

In contrast, our implementation of the CSI and SSP formulas within the VLM architecture achieves an objective quantification of structural integrity that aligns with established international standards for crack width regulation. This confirms that contemporary researchers are correct in their pursuit of merging visual acquisition with predictive performance indices to create fully autonomous, actionable maintenance pipelines. Consequently, our model provides a more holistic diagnostic output than frameworks that operate purely on classification tasks without structural health reasoning.

The high performance of our model, achieving an mAP@0.5 of 0.962, highlights the robustness of the multimodal approach even when tested against varying environmental lighting and standoff distances. Although previous studies noted that many detection frameworks struggle with environmental lighting variability and environmental fatigue coefficients, our VLM demonstrated significant resilience, confirming the effectiveness of our specialized UAV flight path and sensor configuration.

We observed that accuracy remains optimal at a 3-meter standoff distance, after which pixel density becomes a limiting factor for small structural cracks; this observation provides a clear empirical boundary for future deployment in high-rise environments. This result validates the hypothesis that standardized UAV operational parameters are essential for maintaining data uniformity in high-fidelity structural health monitoring. Furthermore, the alignment of our AI-generated severity assessments with expert confusion matrices reinforces the reliability of our deep learning architecture in real-world high-stakes infrastructure scenarios.

From a practical perspective, this research offers a substantial contribution to modern urban asset management by demonstrating that end-to-end automation can drastically reduce inspection latency and costs. The economic analysis indicates that adopting this UAV-VLM framework results in an 86.32% cost reduction compared to conventional scaffolding-based inspections, primarily through the elimination of labor-intensive manual techniques and safety insurance premiums.

This practical implication confirms that the transition from manual, high-risk work to digitized automated workflows is not only technologically feasible but also economically advantageous for the maintenance of aging infrastructure. While our study demonstrates robust performance, the system's reliance on high-resolution imagery means that computational overhead increases significantly when processing 8K resolution, which is a necessary trade-off for maintaining high detection precision. By effectively translating aerial imagery into actionable diagnostic data, this work establishes a new benchmark for autonomous, intelligent infrastructure inspection systems.

VI. CONCLUSION AND RECOMMENDATION

This research successfully developed an autonomous cyber-physical framework that integrates UAV tactile scanning with Vision-Language Models to provide comprehensive, automated structural health diagnostics for high-rise infrastructure. Our findings confirm that coupling a Visual Encoder with an autoregressive Language Decoder enables the system to transcend traditional object detection by generating formal engineering narrative reports, thereby bridging the gap between pixel-level analysis and professional structural assessment. The empirical results demonstrate that the implemented CSI and SSP formulas achieve high diagnostic precision,

significantly reducing inspection latency while maintaining adherence to international building standards. Furthermore, the economic feasibility analysis validates that this automated approach effectively mitigates operational risks and costs associated with traditional labor-intensive inspection methods. Consequently, this study establishes a novel, actionable pipeline that enhances the safety and longevity of modern urban assets through end-to-end intelligent automation.

Future research should focus on optimizing the onboard computational efficiency of the VLM architecture to enable real-time diagnostic performance even when processing high-resolution 8K imagery on constrained edge-computing hardware. It is recommended that subsequent studies explore the integration of multi-spectral sensor data to further refine the detection of subsurface structural anomalies that are currently beyond the scope of standard visual photogrammetry. Investigating the adaptability of this VLM framework across different structural materials, such as historical masonry or complex steel-composite connections, would significantly expand the applicability of the proposed diagnostic system. Furthermore, longitudinal studies assessing the performance of AI-generated predictive maintenance schedules against actual long-term structural decay patterns are necessary to validate the robustness of the SSP formula in diverse environmental contexts. Finally, future work could incorporate federated learning techniques to allow for collaborative model training across multiple urban inspection fleets, thereby enhancing the generalization capabilities of the system without compromising sensitive site data.

AI Use

The authors confirm that artificial intelligence tools, including ChatGPT, were utilized exclusively for language refinement purposes, such as improving grammar, readability, sentence structure, and overall clarity of the manuscript. No AI-based system was involved in the execution of the research activities. The conception of the study, research design, data acquisition, data analysis, methodological development, interpretation of findings, and preparation of conclusions were conducted entirely by the authors. Responsibility for the authenticity, accuracy, validity, and integrity of all material presented in this manuscript remains solely with the authors.

REFERENCES

- Anh, N. M., & Huy, T. Q. (2026). AI-Driven Digital Infrastructure and Infrastructure Innovation Systems: A Conceptual Analysis. *Civil Engineering Science and Technology*, 2(1), 90–111. <https://doi.org/10.51903/cpapgr89>
- Bahmani, H., Mostafaei, H., Santos, P., & Ferrández, D. (2026). Concrete Material Variability and Machine Learning Model Performance: A Comprehensive Review. *Buildings*, 16(3), 556. <https://doi.org/10.3390/buildings16030556>
- Cabello-Collado, C., Rodriguez-Juan, J., Ortiz-Perez, D., Garcia-Rodriguez, J., Tomás, D., &

- Vizcaya-Moreno, M. F. (2024). Automated Generation of Clinical Reports Using Sensing Technologies with Deep Learning Techniques. *Sensors*, 24(9), 2751. <https://doi.org/10.3390/s24092751>
- Cumbajin, E., Rodrigues, N., Costa, P., Miragaia, R., Frazão, L., Costa, N., Fernández-Caballero, A., Carneiro, J., Buruberri, L. H., & Pereira, A. (2023). A Systematic Review on Deep Learning with CNNs Applied to Surface Defect Detection. *Journal of Imaging*, 9(10), 193. <https://doi.org/10.3390/jimaging9100193>
- Dai, R., Wang, R., Shu, C., Li, J., & Wei, Z. (2025). Crack Detection in Civil Infrastructure Using Autonomous Robotic Systems: A Synergistic Review of Platforms, Cognition, and Autonomous Action. *Sensors*, 25(15), 4631. <https://doi.org/10.3390/s25154631>
- Demiral, M., & Otkur, M. (2026). From Mechanics to Machine Learning in Additive Manufacturing: A Review of Deformation, Fatigue, and Fracture. *Technologies*, 14(4), 218. <https://doi.org/10.3390/technologies14040218>
- Diaconita, V., Belciu, A., Corbea, A. M., & Simonca, I. (2025). Decoding Wine Narratives with Hierarchical Attention: Classification, Visual Prompts, and Emerging E-Commerce Possibilities. *Journal of Theoretical and Applied Electronic Commerce Research*, 20(3), 212. <https://doi.org/10.3390/jtaer20030212>
- Golovastikov, N. V., Kazanskiy, N. L., & Khonina, S. N. (2025). Optical Fiber-Based Structural Health Monitoring: Advancements, Applications, and Integration with Artificial Intelligence for Civil and Urban Infrastructure. *Photonics*, 12(6), 615. <https://doi.org/10.3390/photonics12060615>
- Hajjar, M., Zappa, E., & Bolzon, G. (2025). The State of the Art and Potentialities of UAV-Based 3D Measurement Solutions in the Monitoring and Fault Diagnosis of Quasi-Brittle Structures. *Sensors*, 25(16), 5134. <https://doi.org/10.3390/s25165134>
- Hassan, S., & Dhimish, M. (2023). A Survey of CNN-Based Approaches for Crack Detection in Solar PV Modules: Current Trends and Future Directions. *Solar*, 3(4), 663–683. <https://doi.org/10.3390/solar3040036>
- Helal, N. M., Aljehani, A. F., Alomari, S. A., Mahmoud, R. A., & Khalifa, H. M. (2026). Artificial Intelligence-Assisted Detection of Canine Impaction, Localization, and Classification from Panoramic Images: A Diagnostic Accuracy Comparative Study with CBCT. *Children*, 13(4), 507. <https://doi.org/10.3390/children13040507>
- Işık, E., Karaçam, R., Harirchian, E., & Hadzima-Nyarko, M. (2026). Comparative Analysis of Post-Earthquake Damage and Structural Irregularities in RC Buildings: Field Evidence from the 2023 Kahramanmaraş (Türkiye) Earthquakes. *Buildings*, 16(6), 1140. <https://doi.org/10.3390/buildings16061140>
- Jouini, O., Sethom, K., Namoun, A., Aljohani, N., Alanazi, M. H., & Alanazi, M. N. (2024). A Survey of Machine Learning in Edge Computing: Techniques, Frameworks, Applications, Issues, and Research Directions. *Technologies*, 12(6), 81. <https://doi.org/10.3390/technologies12060081>

- Jung, S., Song, A., Lee, K., & Lee, W. H. (2025). Advanced Building Detection with Faster R-CNN Using Elliptical Bounding Boxes for Displacement Handling. *Remote Sensing*, 17(7), 1247. <https://doi.org/10.3390/rs17071247>
- Kowalczyk, M., Andruszko, J., Stefanek, P., & Mazur, R. (2025). Failure and Degradation Mechanisms of Steel Pipelines: Analysis and Development of Effective Preventive Strategies. *Materials*, 18(1), 134. <https://doi.org/10.3390/ma18010134>
- Liang, C. X., Tian, P., Yin, C. H., Yua, Y., Wei, A.-H., Li, M., Song, X., Wang, T., Bi, Z., Liu, M., Bao, R., & Feng, P. (2026). A Comprehensive Survey and Guide to Multimodal Large Language Models in Vision–Language Tasks. *Computation*, 14(6), 125. <https://doi.org/10.3390/computation14060125>
- Manakitsa, N., Maraslidis, G. S., Moysis, L., & Fragulis, G. F. (2024). A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision. *Technologies*, 12(2), 15. <https://doi.org/10.3390/technologies12020015>
- Mjahad, A., Polo-Aguado, A., Llorens-Serrano, L., & Rosado-Muñoz, A. (2025). Optimizing Image Feature Extraction with Convolutional Neural Networks for Chicken Meat Detection Applications. *Applied Sciences*, 15(2), 733. <https://doi.org/10.3390/app15020733>
- Ortiz-Marqués, A., Caldevilla, P., Goldmann, E., Safuta, M., Fernández-Raga, M., & Górski, M. (2025). Porosity and Permeability in Construction Materials as Key Parameters for Their Durability and Performance: A Review. *Buildings*, 15(18), 3422. <https://doi.org/10.3390/buildings15183422>
- Sadeghi, S., Riahinezhad, M., Esmizadeh, E., & Lacasse, M. A. (2026). Service Life Dataset Development for Non-Structural Building Envelope Materials—Current State, Knowledge Gaps and Inconsistencies. *Data*, 11(3), 47. <https://doi.org/10.3390/data11030047>
- Sanchez-Cubillo, J., Del Ser, J., & Martin, J. L. (2024). Toward Fully Automated Inspection of Critical Assets Supported by Autonomous Mobile Robots, Vision Sensors, and Artificial Intelligence. *Sensors*, 24(12), 3721. <https://doi.org/10.3390/s24123721>
- Sharifzada, H., Wang, Y., Sadat, S. I., Javed, H., Akhunzada, K., Javed, S., & Khan, S. (2025). An Image-Based Intelligent System for Addressing Risk in Construction Site Safety Monitoring Within Civil Engineering Projects. *Buildings*, 15(8), 1362. <https://doi.org/10.3390/buildings15081362>
- Syafril, M., Arief, S., & Pangaribuan, J. D. (2025). Application of Digital Twin Technology for Real-Time Monitoring and Predictive Maintenance of Coastal Bridges under Climate Change Scenarios. *Jurnal Rekayasa Sipil Dan Arsitektur*, 1(2), 39–51. <https://doi.org/10.51903/5fkqb467>
- Toptsis, M., Karkanis, N., Giannakoulas, A., & Kaifas, T. (2026). A Review of Embedded Software Architectures for Multi-Sensor Wearable Devices: Sensor Fusion Techniques and Future Research Directions. *Electronics*, 15(2), 295. <https://doi.org/10.3390/electronics15020295>

- Zhao, H., Wang, Q., Shang, R., & Li, S. (2025). Development, Challenges, and Applications of Concrete Coating Technology: Exploring Paths to Enhance Durability and Standardization. *Coatings*, 15(4), 409. <https://doi.org/10.3390/coatings15040409>
- Zuzana, Z., & Štrochová, Š. (2025). Vertical Social Infrastructures: Redefining Community Interaction In High-Rise Urban Housing. *International Journal of Graphic Design*, 3(2), 313–339. <https://doi.org/10.51903/ijgd.v3i2.3103>