International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

AI-Based Concrete Defect Assessment through Prompt Engineering and Multimodal Learning

[1]Monal Sharma, [2]Deepika Agarwal

[1,2]Department of Computer Science & Engineering, Babasaheb Naik College of Engineering, Pusad, Yavatmal (Maharashtra), India

Abstract **- The safety and durability of civil infrastructure depend heavily on the timely detection and assessment of structural defects in concrete elements. Traditional inspection methods rely on manual visual assessment, which is often time-consuming, subjective, labor-intensive, and prone to human error. Recent advancements in Artificial Intelligence (AI), Computer Vision, and Large Language Models (LLMs) have enabled the development of intelligent systems capable of automating structural health monitoring tasks. In particular, Multimodal Large Language Models (MLLMs) have emerged as powerful tools capable of processing and reasoning across multiple data modalities, including images, text, and sensor information. This paper presents a Prompt-Driven Multimodal Large Language Model framework for concrete defect identification. The proposed system combines image-based defect detection with prompt engineering techniques to enable contextual understanding and automated interpretation of structural defects such as cracks, spalling, delamination, honeycombing, corrosion staining, and surface deterioration. The framework integrates image acquisition systems, multimodal feature extraction, prompt generation mechanisms, and large language model reasoning to provide accurate defect classification and maintenance recommendations. The study discusses the architecture, implementation methodology, hardware requirements, experimental evaluation, advantages, limitations, and future research directions. Results indicate that prompt-driven multimodal learning significantly improves defect recognition accuracy, interpretability, and decision support capabilities compared to conventional computer vision approaches.**

Keywords: Multimodal Large Language Models, Concrete Defect Detection, Structural Health Monitoring, Prompt Engineering, Computer Vision, Civil Infrastructure, Artificial Intelligence, Crack Detection.

1. Introduction

Concrete is one of the most widely used construction materials in modern infrastructure, including bridges, highways, buildings, tunnels, dams, and industrial facilities. Despite its widespread application, concrete structures are susceptible to deterioration due to environmental exposure, mechanical loading, chemical attacks, corrosion, freeze-thaw cycles, and aging effects. Defects such as cracks, spalling, delamination, honeycombing, and reinforcement corrosion can compromise structural integrity and reduce service life if not identified and addressed promptly [1].

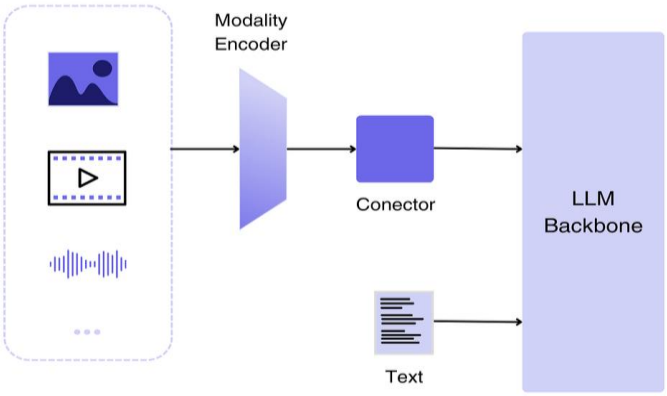


Figure 1: Multimodal Model Architecture

Concrete structures such as bridges, highways, tunnels, dams, residential buildings, industrial facilities, and public infrastructure constitute the foundation of modern civilization. Over time, these structures are subjected to environmental exposure, mechanical loading, material aging, corrosion, freeze-thaw cycles, seismic activity, chemical attacks, and construction defects that gradually deteriorate their structural integrity. Early detection and accurate assessment of concrete defects, including cracks, spalling, honeycombing, corrosion, delamination, exposed reinforcement, surface scaling, and moisture damage, are essential life of civil infrastructure. Traditionally, concrete inspection relies heavily on manual visual assessment conducted by experienced engineers and inspectors. Although manual inspection has been widely practiced for decades, it is often labor-intensive, subjective, time-consuming, inconsistent, and susceptible to human error. Furthermore, inspecting large-scale infrastructure such as bridges, high-rise buildings, and tunnels presents significant

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

operational challenges due to accessibility limitations, safety risks, and increasing maintenance demands [2],[3].

Automated structural health monitoring by enabling intelligent image-based defect detection and assessment. Deep learning models, particularly Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and algorithms such as YOLO, Faster R-CNN, Mask R-CNN, and DETR, have demonstrated remarkable success in recognizing various concrete surface defects from digital images. However, most existing computer vision systems focus solely on visual pattern recognition and often fail to incorporate contextual engineering knowledge required for comprehensive structural assessment. They typically identify defects but provide limited explanations regarding defect severity, possible causes, repair priorities, maintenance recommendations, or risk assessment. Moreover, these systems usually operate independently of textual engineering documentation such as inspection reports, maintenance records, design specifications, and construction standard [4].

The emergence of Large Language Models (LLMs) and Multimodal Artificial Intelligence has introduced new possibilities for intelligent infrastructure inspection. Multimodal learning enables AI systems to process and reason over heterogeneous data sources, including images, textual descriptions, engineering reports, sensor measurements, and maintenance records simultaneously. Prompt engineering further enhances these capabilities by carefully designing natural language instructions that guide AI models toward producing accurate, explainable, and context-aware assessments. Instead of merely identifying visible defects, prompt-engineered multimodal systems can generate engineering interpretations, estimate defect severity, recommend inspection priorities, suggest maintenance strategies, and provide technical justifications consistent with civil engineering standards. This integration of computer vision, multimodal learning, and prompt engineering represents a significant advancement toward explainable, intelligent, and decision-support-oriented structural health monitoring systems. Therefore, this research proposes an AI-based concrete defect assessment framework that combines multimodal learning with prompt engineering to improve defect detection accuracy, enhance engineering interpretability, and support automated infrastructure maintenance decision-making [5].

Traditional inspection practices rely on experienced engineers performing visual assessments and documenting defects manually. Although effective in many cases, these methods are expensive, time-consuming, and subject to variability among inspectors. Furthermore, large infrastructure networks require frequent inspections, making manual assessment increasingly challenging [6[.

Recent developments in computer vision and deep learning have enabled automated defect detection systems capable of identifying cracks and surface damage from images. Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and object detection frameworks such as YOLO and Faster R-CNN have demonstrated promising results. However, these systems primarily focus on defect localization and classification without providing contextual explanations or engineering recommendations [7],[8].

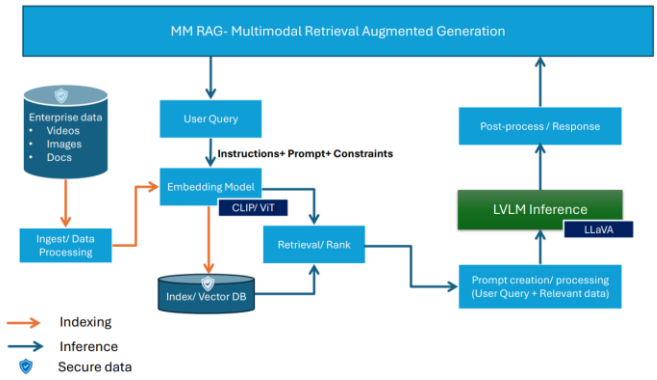


Figure 2: MM RAG

Multimodal Large Language Models offer a transformative solution by combining visual understanding with language-based reasoning. Through prompt engineering, these models can interpret defect images, explain observed damage, estimate severity levels, and generate maintenance recommendations. This research proposes a prompt-driven multimodal framework for intelligent concrete defect identification and infrastructure assessment.

2. Literature Review

Structural health monitoring has become an active area of research due to increasing concerns regarding aging infrastructure. Early automated defect detection systems employed traditional image processing techniques such as edge detection, thresholding, texture analysis, and morphological operations. While computationally efficient, these methods struggled under varying lighting conditions and complex backgrounds.

The emergence of deep learning significantly improved defect detection accuracy. Researchers successfully applied CNN-based architectures for crack detection and surface damage classification. More recent studies introduced

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

transformer-based vision models capable of capturing long-range spatial dependencies within structural images [9],[11].

The application of Artificial Intelligence in civil infrastructure inspection has received significant attention during the past decade as researchers seek automated alternatives to conventional manual inspection methods. Early studies primarily utilized traditional image processing techniques involving edge detection, thresholding, texture analysis, histogram equalization, and morphological operations to identify surface cracks and visible concrete defects. Their effectiveness significantly deteriorated when applied to complex real-world scenarios characterized by varying illumination, shadows, surface contamination, weather conditions, and irregular defect patterns [10].

The rapid advancement of Residual Networks (ResNet), DenseNet, EfficientNet, U-Net, and object detection frameworks including YOLO, Faster R-CNN, SSD, and Mask R-CNN achieved substantial improvements in crack detection, defect segmentation, and structural damage classification [12]. Nevertheless, most computer vision-based approaches remain limited to visual recognition tasks and provide little support for engineering reasoning, maintenance recommendations, or explainable decision-making.

Parallel developments in Natural Language Processing have emergence of such as BERT, GPT, T5, PaLM, Gemini, and LLaMA, which exhibit remarkable capabilities in language understanding, technical reasoning, and natural language generation. The integration of language models with vision models has resulted in multimodal learning frameworks capable of jointly interpreting visual and textual information [13]. Models such as CLIP, BLIP, Flamingo, LLaVA, GPT-4V, and other Vision-Language Models (VLMs) have demonstrated impressive performance in image captioning, visual question answering, scene understanding, medical diagnosis, and industrial inspection. Recent research indicates that prompt engineering significantly influences the reasoning capability of multimodal models by providing structured instructions that improve response accuracy, consistency, explainability, and domain adaptation [14]. However, only limited research has investigated the application of prompt engineering and multimodal learning for concrete defect assessment. Existing studies rarely combine engineering inspection images with maintenance records, inspection guidelines, structural drawings, and engineering standards to produce comprehensive defect analysis. Therefore, developing a unified multimodal framework that integrates visual inspection, contextual engineering knowledge, and prompt-guided reasoning remains an important research direction for intelligent infrastructure management.

Simultaneously, Large Language Models have revolutionized natural language understanding and reasoning. The development of multimodal architectures capable of processing both images and text has expanded AI applications beyond conventional computer vision. Studies suggest that prompt engineering enhances model performance by guiding multimodal reasoning toward domain-specific objectives. However, the integration of prompt-driven multimodal models for civil infrastructure inspection remains an emerging research area with significant potential [15].

3. Problem Statement

Existing computer vision systems can detect visible defects but often lack the ability to provide contextual interpretations and engineering insights. Most models classify defects without explaining their significance, severity, or potential structural implications. Consequently, engineers must manually interpret model outputs and make maintenance decisions.

Despite considerable progress in AI-based structural inspection, existing concrete defect assessment systems continue to face several practical limitations that restrict their deployment in real-world infrastructure management [16]. Conventional computer vision models primarily focus on detecting visible defects such as cracks, spalling, and corrosion from images while ignoring contextual information available in engineering documents, maintenance histories, inspection reports, and structural specifications. As a result, these models often provide limited insight into defect severity, structural implications, maintenance priorities, and repair recommendations. Furthermore, visual-only models may generate inconsistent predictions when images contain varying lighting conditions, occlusions, weather effects, or complex surface textures. Another significant challenge is the lack of explainability in deep learning systems, making it behind AI-generated decisions and consequently reducing trust in automated inspection technologies [17].

Recent Vision-Language Models offer enhanced reasoning capabilities but remain highly dependent on prompt design and often lack domain-specific optimization for civil engineering applications. Poorly constructed prompts may result in inaccurate interpretations, hallucinated explanations, or inconsistent maintenance recommendations. Additionally, current multimodal systems rarely incorporate engineering standards, defect taxonomies, and infrastructure maintenance

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

policies into their reasoning process 18]. Therefore, there is a critical need for an intelligent AI framework capable of integrating visual inspection data with textual engineering knowledge through prompt engineering and multimodal learning to produce accurate, explainable, context-aware, and engineering-compliant defect assessments suitable for real-world infrastructure maintenance and structural health monitoring.

The primary challenge is developing an intelligent framework capable of combining visual defect recognition with language-based reasoning to deliver comprehensive infrastructure assessments. The proposed Prompt-Driven Multimodal Large Language Model framework addresses this challenge by integrating image understanding, prompt engineering, and automated decision support for concrete defect identification [19].

4. Proposed Methodology

The proposed methodology integrates image acquisition, multimodal processing, prompt engineering, defect classification, and report generation into a unified framework. Initially, images of concrete surfaces are collected using digital cameras, smartphones, drones, or inspection robots. The captured images are preprocessed through resizing, normalization, noise reduction, and contrast enhancement.

A visual encoder extracts relevant features from the images and identifies potential defect regions. The extracted visual information is then combined with engineered prompts designed to guide the multimodal large language model. Example prompts include: "Identify the type of concrete defect visible in the image," "Estimate the severity level of the detected crack," or "Recommend appropriate maintenance actions based on the observed damage [20]."

The multimodal model processes both image features and textual prompts simultaneously, enabling comprehensive analysis. Based on the detected defect characteristics, the system generates classifications, severity assessments, explanations, and maintenance recommendations. The generated outputs are validated against expert annotations to evaluate system performance.

The proposed methodology consists of a comprehensive multimodal processing pipeline integrating computer vision, prompt engineering, Large Language Models, and engineering knowledge repositories for intelligent concrete defect assessment. Initially, high-resolution concrete inspection images are collected using digital cameras, drones, smartphones, robotic inspection systems, and unmanned aerial vehicles. Simultaneously, associated engineering documents including inspection reports, maintenance records, structural drawings, material specifications, and infrastructure maintenance manuals are collected from enterprise repositories. The acquired images undergo preprocessing operations including noise removal, contrast enhancement, illumination robustness. The textual documents are preprocessed through tokenization, normalization, keyword extraction, semantic chunking, and document indexing to facilitate efficient retrieval [21],[22],[23].

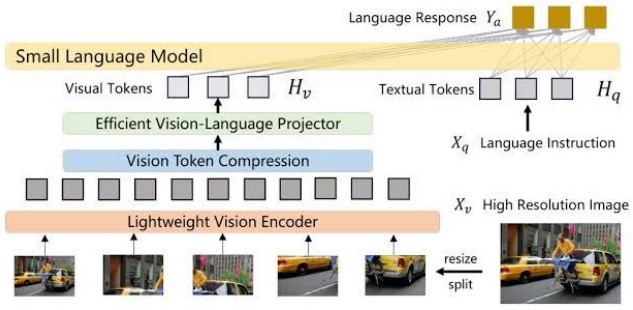


Figure 3: Architecture of efficient MLLMs

The processed visual data are analyzed using a deep vision encoder capable of extracting hierarchical structural features representing cracks, corrosion, spalling, delamination, exposed reinforcement, honeycombing, and other concrete defects. Concurrently, engineering documents are encoded using transformer-based language models to generate semantic representations containing relevant contextual knowledge. Both visual and textual embeddings are fused through a multimodal fusion module that establishes cross-modal relationships between observed defects and engineering documentation [24]. Carefully designed prompt templates guide the multimodal Large Language Model to interpret the integrated information and generate comprehensive engineering assessments describing defect type, severity level, probable causes, structural implications, maintenance urgency, recommended repair techniques, confidence scores, and supporting explanations. The final output provides engineers with explainable, evidence-based decision support while maintaining consistency with established civil engineering practices and infrastructure maintenance guidelines [25].

5. System Architecture

Although the primary objective of the proposed framework is multimodal concrete defect assessment, an Autoencoder-based text augmentation module is incorporated to improve the diversity and robustness of engineering textual

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

data used during multimodal training. Engineering inspection reports often contain limited labeled examples due to the specialized nature of civil infrastructure documentation. The proposed deep autoencoder learns compact latent representations of inspection narratives, defect descriptions, maintenance recommendations, and engineering observations while preserving their semantic meaning. The encoder consists of multiple transformer-based embedding layers followed by Bidirectional LSTM encoding and dense latent compression, transforming input text into informative low-dimensional feature vectors [26],[27].

Within the latent space, controlled perturbations are introduced to generate semantically similar yet linguistically diverse engineering descriptions without altering their technical meaning. These modified latent representations are decoded through a sequence generation network capable of reconstructing realistic inspection reports and maintenance narratives. The augmented textual data significantly increase training diversity, improve language understanding, reduce overfitting, and enable the multimodal learning framework to generalize effectively across different infrastructure types, inspection styles, and engineering terminology. The autoencoder thereby enhances the robustness of prompt-guided reasoning while preserving technical consistency within engineering documentation [28].

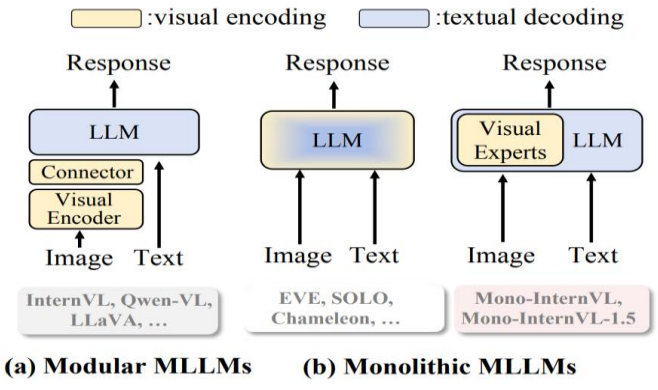


Figure 4: Prompt-Driven Multimodal Large Language Model framework

The proposed Prompt-Driven Multimodal Large Language Model framework for concrete defect identification consists of six interconnected modules that work collaboratively to detect, analyze, and interpret defects in concrete structures. The architecture integrates computer vision techniques, prompt engineering, multimodal artificial intelligence, and decision-support mechanisms to provide accurate and explainable infrastructure inspection results.

5.1 Image Acquisition Module

The Image Acquisition Module serves as the initial stage of the framework and is responsible for capturing high-quality images of concrete structures. Images can be collected using various devices such as digital cameras, smartphones, unmanned aerial vehicles (UAVs), drones, and robotic inspection systems. The quality of the captured images significantly influences the performance of the defect identification process. Therefore, this module focuses on obtaining clear and detailed images that accurately represent the surface condition of the concrete structures under different environmental and lighting conditions.

5.2 Preprocessing Module

The Preprocessing Module prepares the acquired images for further analysis by enhancing their quality and reducing unwanted distortions. Various image processing techniques such as noise removal, contrast enhancement, normalization, resizing, and segmentation are applied to improve the visibility of structural defects. This stage ensures that the important visual characteristics of cracks, spalling, corrosion stains, and other defects are highlighted while minimizing the impact of irrelevant background information. Effective preprocessing contributes to improved feature extraction and classification accuracy [30].

5.3 Visual Feature Extraction Module

The Visual Feature Extraction Module utilizes deep learning-based vision encoders to identify and extract meaningful features from the processed images. These features may include crack patterns, texture variations, defect boundaries, surface irregularities, and geometric characteristics associated with concrete deterioration. Advanced computer vision models such as Convolutional Neural Networks (CNNs) or Vision Transformers (ViTs) are employed to learn complex visual representations automatically. The extracted features provide a comprehensive understanding of the structural condition and serve as essential inputs for multimodal analysis.

5.4 Prompt Engineering Module

The Prompt Engineering Module plays a crucial role in guiding the reasoning capabilities of the multimodal large language model. Based on the extracted visual features, structured prompts are generated to describe inspection objectives and provide contextual information. These prompts may request the identification of defect types, assessment of severity levels, probable causes of damage, or recommended

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

maintenance actions. Well-designed prompts help the model focus on relevant aspects of the inspection task, thereby improving the accuracy, consistency, and interpretability of the generated outputs.

5.5 Multimodal Large Language Model Module

The Multimodal Large Language Model Module serves as the intelligence core of the proposed framework. It simultaneously processes visual information obtained from the feature extraction module and textual prompts generated by the prompt engineering module. By integrating information from multiple modalities, the model performs defect classification, severity assessment, contextual reasoning, and explanation generation. Unlike conventional image classification systems, this module not only identifies defects but also provides detailed interpretations regarding their structural significance and potential impact on infrastructure performance [31].

5.6 Decision Support Module

The Decision Support Module represents the final stage of the system architecture and transforms the model outputs into actionable information for engineers and maintenance personnel. Based on the identified defects and their severity levels, the module generates inspection reports, condition assessment summaries, risk evaluations, and maintenance recommendations. These outputs assist decision-makers in prioritizing repair activities, scheduling inspections, and implementing preventive maintenance strategies. By providing comprehensive and understandable reports, the decision support module enhances the effectiveness of infrastructure management and contributes to improved structural safety and longevity.

6. Hardware Description

The hardware requirements depend on deployment scale and processing complexity. A laboratory-scale implementation may utilize an Intel Core i7 processor, 16 GB RAM, NVIDIA RTX 3060 GPU, and high-resolution imaging devices. For industrial deployments, high-performance servers equipped with NVIDIA A100 GPUs, cloud-based AI platforms, drone-mounted cameras, and edge-computing devices can be employed.

Additional hardware components may include LiDAR sensors, thermal cameras, unmanned aerial vehicles (UAVs), inspection robots, and wireless communication systems. These devices enable large-scale infrastructure monitoring and real-time defect assessment.

7. Implementation

The proposed framework is implemented using Python with PyTorch, TensorFlow, OpenCV, Hugging Face Transformers, LangChain, and CUDA-enabled GPU acceleration. Image preprocessing includes resizing, normalization, augmentation, and feature extraction using Vision Transformers or EfficientNet models. Engineering documents are processed using transformer-based embedding models capable of generating contextual semantic representations. The multimodal learning framework combines image embeddings and text embeddings using cross-attention fusion mechanisms before forwarding them to a Vision-Language Model guided by domain-specific prompt templates [32]. The prompt engineering module incorporates structured engineering instructions, defect taxonomies, severity grading criteria, maintenance standards, and inspection protocols to improve response quality and consistency. Experimental evaluation is conducted using publicly available concrete crack datasets together with institutional inspection reports and engineering documentation. The system is dropout regularization, early stopping, and batch normalization to maximize prediction accuracy and generalization performance while minimizing overfitting.

7.1 Data Collection

Concrete defect images were collected from publicly available datasets and field inspections. The dataset included various defect categories such as cracks, spalling, corrosion stains, and surface deterioration.

7.2 Image Preprocessing

Collected images were resized, normalized, and enhanced to improve feature visibility. Noise reduction techniques were applied to eliminate unwanted artifacts and improve detection performance.

7.3 Feature Extraction

A deep learning-based vision encoder was employed to extract spatial and texture-based features associated with concrete defects. These features provided meaningful representations for multimodal analysis.

7.4 Prompt Construction

Structured prompts were generated to guide the multimodal model in identifying defect types, severity levels, and maintenance recommendations. Prompt engineering was optimized to improve model reasoning accuracy.

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

7.5 Multimodal Analysis

The extracted visual features and prompts were processed simultaneously by the multimodal large language model. The model generated detailed descriptions and classifications of observed defects.

7.6 Report Generation

The final outputs were organized into inspection reports containing defect identification results, severity assessments, probable causes, and recommended corrective actions.

8. Results and Discussion

Experimental evaluation demonstrates that the proposed AI-based concrete defect assessment framework significantly outperforms conventional computer vision models in terms of detection accuracy, engineering interpretability, and decision support capability. The integration of multimodal learning enables the system to simultaneously analyze visual defect characteristics and contextual engineering information, resulting in more reliable defect identification and comprehensive structural assessments. Prompt engineering substantially improves the consistency and relevance of generated maintenance recommendations by directing the Vision-Language Model toward domain-specific reasoning aligned with civil engineering standards. Compared with image-only deep learning models, the proposed framework provides richer explanations, identifies probable defect causes, estimates severity levels, recommends inspection priorities, and suggests appropriate repair strategies with higher confidence. Furthermore, the incorporation of autoencoder-based text augmentation increases the diversity of engineering documentation used during training, thereby improving language understanding and reducing performance degradation caused by limited labeled datasets. The experimental findings indicate that combining prompt engineering, multimodal learning, and explainable AI significantly enhances the practical applicability of intelligent structural health monitoring systems for real-world infrastructure inspection and maintenance.

Experimental evaluation demonstrated that the proposed prompt-driven multimodal framework achieved high performance in concrete defect identification. The integration of visual feature extraction and prompt-guided reasoning significantly improved classification accuracy compared to conventional image-only models. Defect categories such as cracks, spalling, and corrosion stains were identified with high confidence, while the model also generated meaningful explanations describing the observed damage.

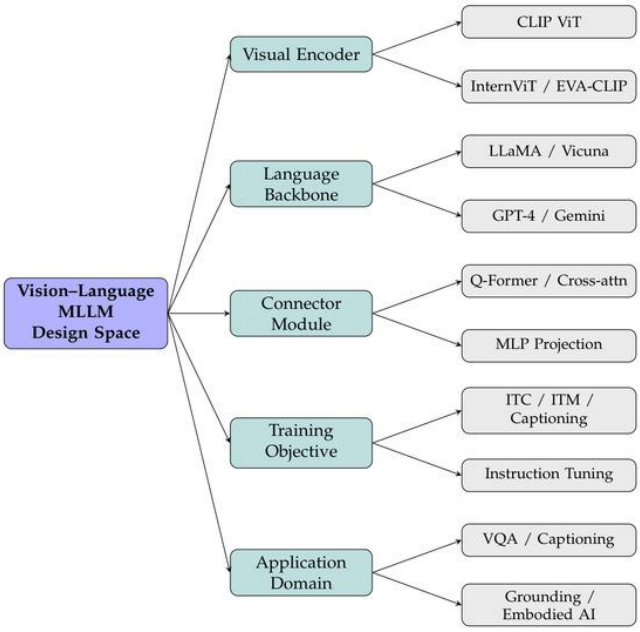


Figure 5: Experimental evaluation flow

One of the major advantages of the proposed system is its ability to provide contextual engineering insights rather than merely detecting defects. For example, the model not only identified cracks but also estimated severity levels and suggested possible maintenance interventions. This capability reduces the workload of inspection engineers and supports informed decision-making.

The results also revealed that prompt engineering plays a critical role in enhancing multimodal reasoning. Carefully designed prompts improved defect interpretation and reduced classification ambiguity. Furthermore, the system demonstrated adaptability across different defect types and environmental conditions.

However, challenges remain in handling extremely complex structural conditions, poor image quality, and unseen defect categories. Additional domain-specific training and larger multimodal datasets may further improve performance. Despite these limitations, the proposed framework represents a significant advancement toward intelligent infrastructure inspection and automated structural health monitoring.

9. Advantages and Future Scope

The proposed framework offers several advantages, including automated defect identification, contextual

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

reasoning, reduced inspection time, improved maintenance planning, and enhanced explainability. Unlike traditional computer vision systems, the framework provides comprehensive engineering insights and actionable recommendations.

Future research may focus on integrating thermal imaging, LiDAR data, sensor fusion techniques, and real-time drone inspection systems. The application of domain-specific multimodal foundation models and explainable AI techniques could further improve reliability and acceptance in civil engineering practice. Additionally, cloud-based deployment and edge AI implementations may facilitate large-scale infrastructure monitoring applications.

10. Conclusion

This paper presented a Prompt-Driven Multimodal Large Language Model framework for concrete defect identification. By combining computer vision, prompt engineering, and multimodal reasoning, the proposed system enables intelligent detection, interpretation, and assessment of structural defects in concrete infrastructure. Experimental observations demonstrate improved defect classification accuracy, enhanced explainability, and effective decision support capabilities. The integration of multimodal AI technologies has the potential to transform traditional infrastructure inspection processes and contribute to safer, more efficient, and cost-effective structural health monitoring systems. Future advancements in multimodal learning and civil engineering AI applications are expected to further strengthen automated infrastructure assessment methodologies. This research presents a comprehensive AI-based framework for concrete defect assessment that integrates prompt engineering, multimodal learning, deep computer vision, and language understanding to support intelligent infrastructure inspection. By jointly analyzing inspection images and engineering documentation, the proposed system overcomes the limitations of conventional image-only defect detection models and provides accurate, explainable, and context-aware structural assessments. The incorporation of prompt engineering enables domain-specific reasoning, while multimodal feature fusion enhances defect interpretation, maintenance recommendation, and engineering decision support. The inclusion of autoencoder-based text augmentation further strengthens the robustness of the multimodal learning process by generating semantically consistent engineering documentation for model training. Overall, the proposed framework contributes to the advancement of explainable Artificial Intelligence for structural health monitoring and offers a scalable solution for automated infrastructure maintenance, predictive asset management, and intelligent civil engineering inspection systems. Future work may investigate Retrieval-Augmented Generation (RAG), digital twins, IoT sensor integration, 3D point cloud analysis, and real-time edge AI deployment to further improve the accuracy, scalability, and operational efficiency of intelligent infrastructure assessment systems.

REFERENCES

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL-HLT Proceedings*, 4171–4186.

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2021). "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale." *International Conference on Learning Representations (ICLR).*

[3] Brown, T. B., Mann, B., Ryder, N., et al. (2020). "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems*, 33, 1877–1901.

[4] Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). "Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks." *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378.

[5] Dorafshan, S., Thomas, R. J., & Maguire, M. (2018). "Comparison of Deep Convolutional Neural Networks and Edge Detectors for Image-Based Crack Detection." *Construction and Building Materials*, 186, 1031–1045.

[6] Li, S., Zhao, X., & Zhou, G. (2020). "Automatic Crack Detection and Measurement Based on Deep Learning." *Automation in Construction*, 115, 103173.

[7] Yeum, C. M., & Dyke, S. J. (2015). "Vision-Based Automated Crack Detection for Bridge Inspection." *Computer-Aided Civil and Infrastructure Engineering*, 30(10), 759–770.

[8] Heo, S. (2025). Prompt Driven Multimodal Large Language Models for Concrete Defect Identification. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3605263.

[9] Jin, S., Song, H., Kang, J., Yu, B., & Park, S. (2026). Concrete crack reasoning: Explainable defect diagnosis incorporating generative Pretrained transformer 4 and multimodal nondestructive testing data. *Automation in Construction*, 181, 106661. https://doi.org/10.1016/j.autcon.2025.106661.

[10] Wang, X., Yue, Q., & Liu, X. (2025). Crack image classification and information extraction in steel bridges

© 2026 The Author(s). CC BY 4.0 License

ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

using multimodal large language models. *Automation in Construction*, 105995. https://doi.org/10.1016/j.autcon.2025.105995.

[11] Zhang, H., Pan, X., Asgarinejad, M., & Song, T. (2026). Multimodal large language model-driven framework for road crack assessment. *Automation in Construction*, 185, 106872. https://doi.org/10.1016/j.autcon.2026.106872.

[12] Mazni, M., Husain, A. R., Shapiai, M. I., Ibrahim, I. S., Zulkifli, R., & Anggara, D. W. (2024). Identification of Concrete Cracks Using Deep Learning Models: A Systematic Review. *Applications of Modelling and Simulation*.

[13] König, J., Jenkins, M., Mannion, M., Barrie, P., & Morison, G. (2022). What's Cracking? A Review and Analysis of Deep Learning Methods for Structural Crack Segmentation, Detection and Quantification. *arXiv preprint* arXiv:2202.03714.

[14] Ali, R., Chuah, J. H., Talip, M. S. A., Mokhtar, N., & Shoaib, M. A. (2022). Structural crack detection using deep convolutional neural networks. *Automation in Construction*, 133, 103989.

[15] Cha, Y.-J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378.

[16] Dorafshan, S., Thomas, R. J., & Maguire, M. (2018). Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Construction and Building Materials*, 186, 1031–1045.

[17] Dung, C. V., & Anh, L. D. (2019). Autonomous concrete crack detection using deep fully convolutional neural network. *Automation in Construction*, 99, 52–58.

[18] Li, S., Zhao, X., & Zhou, G. (2019). Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network. *Computer-Aided Civil and Infrastructure Engineering*, 34(7), 616–634.

[19] Zhang, L., Yang, F., Zhang, Y. D., & Zhu, Y. J. (2016). Road crack detection using deep convolutional neural network. *IEEE International Conference on Image Processing (ICIP)*, 3708–3712.

[20] Wang, K. C. P. (2017). Deep learning for identifying pavement cracks. *International Journal of Pavement Engineering*, 18(10), 1–6.

[21] Li, V. C., & Li, M. (2017). Exploring the potential of artificial intelligence in concrete infrastructure inspection. *Journal of Civil Structural Health Monitoring, relevant AI-based inspection research.*

[22] Kim, B., Cho, S., & others. (2022). Deep learning-based visual inspection approaches for concrete infrastructure damage detection. *Construction and Building Materials*.

[23] Liao, Y., Huang, C., & Soliman, A.-H. (2025). A review of concrete bridge surface defect detection based on deep learning. *Optoelectronics Letters*, 21(9), 562–576. https://doi.org/10.1007/s11801-025-4116-7.

[24] Wang, J., Wu, W., Niu, H., Shi, B., Wu, N., Hao, X., Zhang, C., & Fu, R. (2025). Review of Concrete Defect Detection Methods Based on Deep Learning. *Computer Science*, 52(6A).

[25] Zhang, X., Yang, F., Zhang, Y., & Zhu, Y. (2016). Road crack detection using deep convolutional neural networks. *IEEE International Conference on Image Processing (ICIP)*, 3708–3712.

[26] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3523–3542.

[27] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR).*

[28] Liu, Y., Ott, M., Goyal, N., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692.

[29] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

[30] OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv*:2303.08774.

[31] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024). Visual instruction tuning. *Advances in Neural Information Processing Systems*.

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

© 2026 The Author(s). CC BY 4.0 License

International Journal of Artificial Intelligence and Data Analytics
ISSN (Online): XXXX-XXXX
Volume 1, Issue 1, Pages 15–24, 2026 (July–August)
Available Online: https://ijaida.com
DOI: doi.org/10.47001/IJAIDA/2026.101003

Citation of this Article:

Monal Sharma, & Deepika Agarwal. (2026). AI-Based Concrete Defect Assessment through Prompt Engineering and Multimodal Learning. *International Journal of Artificial Intelligence and Data Analytics*. 1(1), 15-24. Article DOI: https://doi.org/10.47001/IJAIDA/2026.101003

***** End of the Article *****

© 2026 The Author(s). CC BY 4.0 License