Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

A Theoretical Framework for AI-Assisted Civic Issue Reporting and Validation in Smart Cities

Naveen. V
Student, B. Tech IoT 4th Year
Holy Mary Inst. Of Tech. and Science
Hyderabad, Telangana, India

Vaishnavi. B
Student, B. Tech IoT 4th Year
Holy Mary Inst. Of Tech. and Science
Hyderabad, Telangana, India

Umesh. I
Student, B. Tech IoT 4th Year
Holy Mary Inst. Of Tech. and Science
Hyderabad, Telangana, India

Devi. G
Asst. prof, CSE
Holy Mary Inst. Of Tech. and Science
Hyderabad, Telangana, India

Raghavender. M
Student, B. Tech IoT 4th Year
Holy Mary Inst. Of Tech. and Science
Hyderabad, Telangana, India

Dr. Venkataramana. B
Assoc. prof, CSE
Holy Mary Inst. Of Tech. and Science
Hyderabad, Telangana, India

Abstract - **Urban administrations face persistent challenges in reporting, prioritizing, and validating civic infrastructure issues due to manual workflows, inconsistent citizen inputs, and limited resolution verification mechanisms. Existing digital grievance systems remain largely text-centric and require substantial human intervention. This paper presents a multimodal AI–assisted, crowdsourced civic issue management framework that leverages Multimodal Large Language Models (MLLMs) for image-based issue understanding and workflow support. Citizens submit geotagged images through a mobile-friendly interface, enabling automated generation of standardized descriptions, semantic categorization, and severity-aware prioritization to assist municipal decision-making. A key contribution is an AI-assisted before–after image validation mechanism that formulates issue resolution as a semantic change detection task, allowing meaningful repairs to be distinguished from irrelevant visual variations. Rather than proposing new model architectures, the framework demonstrates system-level integration of existing multimodal AI capabilities to improve transparency, consistency, and accountability in smart city governance while maintaining human oversight.**

Keywords - Crowdsourcing, Civic Issue Reporting, Artificial Intelligence, Image Processing, Next.js, MERN Stack, WebView Android App, Municipal Governance, Issue Prioritization, Work Validation, Computer Vision, Smart City Automation

I. INTRODUCTION

The rapid growth of urban and rural populations has led to an increasing number of civic infrastructure challenges, including damaged streetlights, water leakages, potholes, and waste accumulation. Efficient identification and resolution of such issues are essential for maintaining functional and safe communities. However, existing grievance redressal systems continue to face significant limitations. Conventional reporting mechanisms rely heavily on manual text entry, which is often slow, inconsistent, and prone to incomplete or ambiguous information. Traditional channels such as phone calls, paper-based forms, or basic online portals frequently lack transparency, provide limited mechanisms for verifying resolution, and offer minimal support for prioritizing urgent or high-risk issues. As a result, municipal authorities often struggle to identify critical hazards in a timely manner, leading to delayed responses and fragmented coordination across departments.

Addressing these challenges requires a combination of active citizen participation and intelligent automation. While crowdsourcing provides a scalable means of collecting civic issue data, the resulting volume and diversity of reports necessitate automated interpretation and decision support to be operationally useful. Recent advances in multimodal large language models (MLLMs) have demonstrated the ability to jointly reason over visual and linguistic information, enabling open-ended semantic interpretation of images without reliance on narrowly defined object categories or task-specific training pipelines. Prior work in multimodal urban analysis has shown that such models can extract meaningful semantic patterns from image collections by emphasizing contextual reasoning over low-level visual matching. These capabilities motivate the application of multimodal reasoning to civic infrastructure reporting, where visual evidence plays a central role and environmental variability is common.

This paper introduces an AI-driven crowdsourced civic issue reporting and resolution system designed to bridge the gap between citizens and municipal administrations. The proposed framework enables users to report civic issues by submitting geotagged images through a mobile-friendly interface, reducing dependence on detailed textual input. Multimodal reasoning is employed to support image-based issue understanding, automated description generation, severity-aware prioritization, and image-based resolution assessment. Rather than claiming perfect accuracy, the system is designed as a decision-support and accountability-enhancing tool that assists municipal workflows by improving consistency, transparency, and evidence-based processing.

The framework is implemented as a scalable, end-to-end system using a modern web technology stack, demonstrating the practical feasibility of integrating existing multimodal AI models into real-world governance processes. Unlike prior multimodal research that

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

focuses on large-scale historical or longitudinal urban analysis, the proposed system applies multimodal reasoning at an operational, micro-temporal level, targeting individual civic issues and short-term resolution cycles. This distinction allows the system to function as a practical governance tool while leveraging general principles established in recent multimodal AI research.

Contributions

This paper makes the following key contributions:

- Proposes an AI-powered crowdsourced civic issue reporting framework that leverages multimodal large language models to support image-based issue understanding and processing, reducing reliance on manual and text-centric reporting mechanisms.
- Demonstrates how general multimodal reasoning principles can be applied to civic infrastructure monitoring to enable automated description generation, semantic issue categorization, and severity-aware prioritization without task-specific model training.
- Introduces an AI-assisted before–after image comparison mechanism designed to support resolution validation by assessing semantic changes between baseline and post-resolution images, thereby improving accountability in municipal workflows.
- Presents a scalable end-to-end system implementation using the MERN technology stack with a Next.js frontend, illustrating practical integration of multimodal AI into smart city governance environments.

II. RELATED WORK

2.1 Civic Issue Reporting Systems

The transition from traditional to digital civic management has been driven by the increasing volume and diversity of infrastructure-related complaints associated with rapid urbanization. In response, numerous government-led and municipal digital platforms have been introduced to modernize grievance redressal processes. While these systems have improved accessibility and record-keeping compared to purely manual approaches, many continue to face structural and technological limitations that constrain their overall effectiveness in large-scale urban environments.

Existing Government Applications

Current civic issue reporting frameworks in India typically operate through a combination of manual and early-stage digital channels, including phone calls, paper-based forms, in-person visits to municipal offices, and basic web or mobile portals. Prominent national initiatives include the Swachhata MoHUA App, which enables citizens to report sanitation-related concerns, and the UPYOG platform developed under the National Urban Digital Mission, which provides a centralized Municipal Grievance Redressal (MGR) module for multiple urban services. Despite these efforts, many city-level deployments remain functionally siloed, with grievance data fragmented across departments. This fragmentation often results in limited inter-departmental coordination, duplicated effort, and delayed issue resolution.

Limitations: Manual Verification and Limited Automation

A common limitation across existing civic reporting systems is their heavy reliance on manual workflows and human intervention, which can be slow, inconsistent, and difficult to scale as report volumes increase. Several recurring challenges have been identified in prior deployments:

- Manual Redressal Bottlenecks: Many platforms require structural engineers, inspectors, or municipal staff to manually review reported issues before action can be taken. While necessary in certain cases, this process is time-consuming, resource-intensive, and subject to variability in human judgment.
- Data Management and Consistency Challenges: In practice, grievance tracking is often supported by spreadsheets or paper-based records alongside digital portals. Such hybrid data management practices can lead to inconsistencies, errors, missed deadlines, and the absence of a single reliable source of truth for issue status.
- Limited Resolution Verification: Most existing systems lack automated mechanisms to assess whether reported issues have been physically resolved on the ground. Issues may be marked as "resolved" in administrative systems without systematic validation using visual or sensor-based evidence, creating potential accountability gaps in service delivery.
- Absence of Intelligent Prioritization: Traditional reporting platforms generally rely on rule-based or manual triaging methods, which makes it difficult to consistently distinguish between low-impact complaints and critical safety or public health hazards. Without AI-assisted severity assessment, resource allocation may be inefficient, and high-risk issues can experience delayed intervention.

2.2 Computer Vision in Urban Monitoring

The integration of computer vision techniques into urban monitoring has significantly influenced the way civic infrastructure is inspected and maintained. Traditional inspection workflows, which relied heavily on manual visual assessment by certified engineers or field staff, have increasingly been supplemented by automated approaches based on deep learning. In particular, Convolutional Neural Networks (CNNs) have become widely used for detecting and localizing specific types of infrastructure defects in urban environments.

CNN-Based Pothole Detection

Pothole detection is a commonly studied problem in road maintenance due to its implications for traffic safety, vehicle damage, and economic cost. Research in this area has explored a range of deep learning architectures aimed at automating road condition assessment:

- Object Detection Frameworks: Models based on region proposal networks, such as Faster R-CNN with Inception-V2 backbones, have been applied to pothole detection tasks, with reported classification and detection performance varying across datasets and experimental settings.
- Real-Time YOLO Variants: The YOLO (You Only Look Once) family of models, including versions from YOLOv3 through YOLOv8, has gained attention for its ability to perform real-time inference. Variants such as YOLOv5 are commonly explored in urban road monitoring scenarios due to their balance between detection performance and computational efficiency.
- Siamese and Advanced Network Architectures: More recent approaches, including frameworks such as RoadScan, combine CNN-based feature extractors (e.g., VGG16) with Siamese network designs and metric learning objectives. These methods aim to reduce model complexity while maintaining competitive performance, making them potentially suitable for larger-scale deployments.

Waste Detection and Segregation Models

Computer vision has also been applied to automated waste detection and classification in the context of smart waste management systems, where accurate material identification can support improved operational efficiency and sustainability:

- Deep CNN-Based Classifiers: Systems such as ConvoWaste employ deep convolutional architectures, including Inception-ResNet-V2, to classify waste into categories such as plastic, metal, glass, organic waste, and electronic waste. Reported performance in controlled settings indicates high classification accuracy on curated datasets.
- Specialized Classification Models: Alternative approaches, including Capsule Networks and other specialized architectures, have

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

been explored for distinguishing between specific waste types (e.g., plastic versus non-plastic). These models typically require large, carefully labeled datasets and controlled imaging conditions to achieve reliable performance.

Limitations of Traditional Computer Vision Approaches

Despite promising results in experimental evaluations, conventional CNN-based computer vision systems face several challenges when deployed in real-world urban environments:

- Task-Specific Design: Most vision models are trained on fixed, labeled datasets for narrowly defined detection or classification tasks. This task specificity limits their ability to generalize to novel object categories, unforeseen scenarios, or open-ended semantic queries without retraining or redesign.
- Generalization and Domain Shift: Performance degradation is commonly observed when models trained on one geographic region or imaging setup are applied to new environments. Variations in camera sensors, image resolution, lighting, weather conditions, and urban layout can negatively affect model reliability.
- Sensitivity to Environmental Noise: Vision-based models may produce false positives due to shadows, reflections, road markings, or surface irregularities, which can visually resemble defects such as potholes or debris under certain conditions.
- Computational and Data Constraints: Many high-performing deep learning models require substantial computational resources and large volumes of annotated training data. These requirements can limit scalability and complicate deployment on low-resource devices or in city-wide civic reporting systems.

2.3 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) extend traditional language models by jointly processing visual and textual inputs within a shared semantic representation space. By learning cross-modal associations, these models enable open-ended interpretation of complex visual scenes without reliance on rigid, predefined object categories or narrowly scoped task-specific training pipelines. This represents a shift from conventional computer vision approaches, which are typically optimized for specific detection or classification tasks.

Recent work in multimodal reasoning suggests that MLLMs can generate context-aware textual descriptions from images, support high-level semantic comparison, and reason about visual differences across multiple observations. In contrast to convolutional neural network–based systems that often depend on pixel-level feature matching, multimodal models emphasize semantic interpretation, which can help mitigate sensitivity to variations in lighting, weather, and viewpoint. These characteristics make MLLMs well suited for complex real-world environments, including urban settings where visual conditions are highly variable.

In the context of civic infrastructure monitoring, such capabilities support image-based issue interpretation, standardized description generation, and semantic comparison of before–after visual evidence for resolution assessment. Rather than proposing a new multimodal architecture, this work focuses on the system-level integration of existing MLLMs into municipal governance workflows. The contribution lies in demonstrating how general multimodal reasoning principles can be adapted to operational civic processes—such as issue reporting, prioritization, and resolution assessment—thereby highlighting the practical applicability of multimodal AI within smart city governance systems.

III. SYSTEM ARCHITECTURE

The proposed Crowdsourced Civic Issue Reporting and Resolution System adopts a modular and scalable architecture designed to support multimodal AI-assisted analysis, real-time citizen interaction, and structured municipal workflows. The system is implemented using the MERN technology stack (MongoDB, Express.js, React, Node.js) with a Next.js frontend, and integrates pre-trained Multimodal Large Language Models (MLLMs) to support image-based interpretation, description generation, prioritization, and resolution assessment.

3.1 High-Level Architectural Overview

At a high level, the architecture is organized into five loosely coupled layers:

1. Client Layer
2. Application & API Layer
3. AI & Multimodal Reasoning Layer
4. Data & Storage Layer
5. Administration & Monitoring Layer

This layered design supports scalability, fault isolation, and flexible deployment across different municipal contexts.

3.2 Client Layer

The client layer provides a mobile-friendly interface through which citizens can report civic issues by capturing images and optionally supplying minimal textual input. The frontend is implemented using Next.js to support efficient rendering and optimized performance across devices. To improve accessibility and reduce deployment complexity, the web application can be packaged as an Android WebView, enabling use on low-end smartphones without requiring a dedicated native application.
Captured images are automatically augmented with geospatial metadata using device-level GPS services before being forwarded to the backend for further processing.

3.3 Application & API Layer

The application layer is implemented using Node.js with Express.js and exposes RESTful APIs responsible for:

- User authentication and request validation
- Image upload and metadata management
- Issue lifecycle tracking (reported → assigned → resolved)
- Coordination between AI services, databases, and administrative interfaces

This layer functions as the orchestration component of the system, managing data flow and ensuring consistent interaction between frontend clients, AI-assisted analysis services, and municipal dashboards.

3.4 AI & Multimodal Reasoning Layer

The AI layer provides analytical support for interpreting reported civic issues and assisting downstream decision-making. It integrates pre-trained Multimodal Large Language Models (MLLMs) to perform semantic reasoning over visual and textual inputs. Rather than relying on rigid predefined labels or task-specific classifiers, this layer emphasizes open-ended interpretation guided by visual context and metadata. The primary functions of this layer include:

1. **Issue Understanding:** Uploaded images are analyzed to infer the likely nature of the reported civic issue using multimodal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

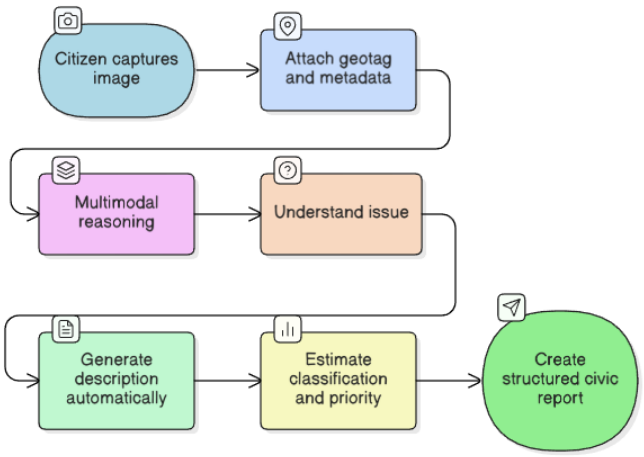
reasoning. This process is intended to support consistent interpretation rather than replace human judgment.

2. **Automatic Description Generation:** The system generates standardized textual descriptions based on visual input to improve consistency and completeness of issue documentation, particularly when user-provided text is minimal or ambiguous.

3. **Classification and Prioritization:** Visual evidence, generated descriptions, and contextual metadata are combined to support assignment of issues to relevant municipal departments and to estimate relative priority levels. These outputs are designed to assist triaging rather than serve as definitive classifications.

4. **Resolution Assessment (Before–After Comparison):** Baseline and post-resolution images submitted at different stages of the issue lifecycle are compared using semantic reasoning to assess whether the reported condition appears to have been addressed. The analysis focuses on meaningful visual changes while reducing sensitivity to non-essential variations such as lighting or viewpoint differences.

To improve computational efficiency at scale, embedding-based retrieval techniques may be used to shortlist relevant image pairs or records prior to detailed multimodal reasoning.



3.5 Data & Storage Layer

The data layer employs MongoDB as the primary datastore for managing:

- Civic reports and associated metadata
- Image references and embeddings
- AI-generated descriptions and routing information
- Issue status, service-level agreement (SLA) timelines, and validation outcomes

Images are stored in scalable object storage services, while structured metadata and embeddings are maintained within the database to support efficient querying and retrieval.

3.6 Administration & Monitoring Layer

Municipal administrators access a centralized dashboard that provides tools for:

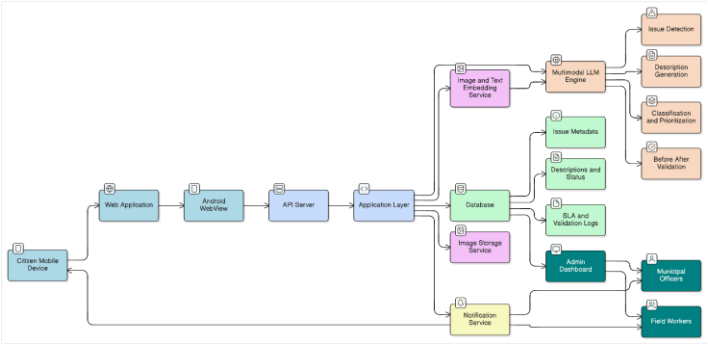
- Monitoring reported issues with geospatial visualization
- Assigning tasks based on priority indicators
- Tracking SLA adherence and operational metrics
- Reviewing AI-assisted resolution assessment outputs

Automated notifications can be triggered at different stages of the issue lifecycle to support transparency and timely intervention.

3.7 Architectural Considerations

The proposed architecture is designed to support:

- An end-to-end workflow from issue reporting to resolution assessment
- Scalable integration of AI-assisted analysis without task-specific retraining
- Improved accountability through image-supported resolution review
- Flexible integration with existing municipal information systems



IV. MULTIMODAL AI FRAMEWORK

The proposed framework aims to transition civic issue management from predominantly manual, text-centric reporting pipelines toward an AI-assisted, vision-first workflow supported by Multimodal Large Language Models (MLLMs). By leveraging multimodal reasoning, the system is designed to support the interpretation of unstructured visual inputs and convert them into structured representations that can assist municipal authorities in issue documentation, prioritization, and resolution assessment.
In contrast to traditional computer vision pipelines that rely on narrowly defined object categories and task-specific training, the proposed approach emphasizes semantic interpretation of visual content. This design allows the system to operate in diverse urban environments without requiring exhaustive retraining or rigid label definitions, while remaining dependent on human oversight for final decision-making.

4.1 Image-Based Issue Understanding

The Image-Based Issue Understanding module serves as the primary analytical entry point for all reported civic issues. Its objective is to support consistent and context-aware documentation of civic conditions, independent of a user's technical expertise or linguistic proficiency.

Input
The system accepts an image captured through a mobile-friendly interface, optionally accompanied by minimal textual input. Each image is automatically augmented with geospatial metadata,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

including GPS coordinates and timestamp information, to provide spatial and temporal context for the reported issue.

Multimodal Reasoning Process

The uploaded image is analyzed using a pre-trained Multimodal Large Language Model (e.g., Gemini-1.5 Pro) to perform semantic reasoning over visual and contextual information. Rather than relying on a fixed set of predefined object categories, the model is used to infer the likely nature and context of the reported issue through open-ended interpretation of the scene.

Through joint reasoning over visual features, metadata, and linguistic representations, the model evaluates the broader urban context depicted in the image. This process is intended to emphasize semantically meaningful elements relevant to civic infrastructure while reducing sensitivity to non-essential visual variations.

Structured Output Generation

Based on the multimodal interpretation, the framework produces a structured representation of the reported issue consisting of the following components:

1. **Issue Type:** A semantic categorization suggesting the relevant municipal service domain (e.g., sanitation, road infrastructure, electricity). This categorization is intended to support routing and triaging rather than act as a definitive classification.

2. **Location Context:** In addition to GPS metadata, the system may infer contextual cues from the image—such as nearby pedestrian areas, roadways, or residential surroundings—to enhance situational awareness for municipal staff.

3. **Severity Indicators:** A relative priority level (e.g., High, Medium, or Low) is estimated based on visual evidence and contextual interpretation, such as apparent obstruction, potential safety risks, or scale of impact. These indicators are designed to assist prioritization decisions.

4. **Auto-Generated Description:** The model generates a standardized textual description summarizing the observed issue. This description aims to improve consistency and completeness of documentation, particularly when user-provided input is limited or ambiguous.

Relation to Prior Multimodal Reasoning Approaches

The proposed framework is informed by recent research in multimodal semantic reasoning, which treats images as contextual representations rather than isolated detection tasks. Traditional vision-based systems often exhibit sensitivity to environmental variations such as lighting, shadows, weather conditions, or seasonal changes, which can introduce noise into automated analysis.

By emphasizing semantic interpretation over pixel-level comparison, the multimodal reasoning approach adopted in this work is intended to reduce the influence of such non-essential variations. This design supports the generation of interpretable, evidence-based outputs that explain *what* issue may be present and *why* it may require attention. These structured outputs form the basis for downstream prioritization, routing, and resolution assessment processes within the overall system.

4.2 Automatic Description Generation

The proposed system incorporates an AI-assisted description generation component to reduce reliance on user-dependent textual reporting. By leveraging the multimodal reasoning capabilities of Multimodal Large Language Models (MLLMs), the framework is designed to support consistent and structured documentation of civic issues based on visual input, even when user-provided descriptions are minimal or incomplete.

Limitations of User-Provided Textual Reports

Conventional grievance redressal platforms typically depend on manual text entry by citizens to describe reported issues. In practice, such descriptions are often brief, vague, or incomplete, which can limit their usefulness for downstream assessment and routing. Citizens may lack technical familiarity with infrastructure terminology or may omit important contextual details related to severity, scale, or surrounding conditions.

These variations in reporting quality can result in fragmented and non-standardized data, complicating automated processing, delaying inter-departmental coordination, and increasing the burden on municipal staff to interpret reports manually. As a result, textual input alone may act as a bottleneck rather than a reliable source of actionable information.

Multimodal Captioning Using Large Language Models

To mitigate these limitations, the proposed framework employs a vision-based multimodal captioning module using pre-trained MLLMs. The model performs joint reasoning over visual content and associated metadata to infer a descriptive summary of the reported issue.

Unlike traditional computer vision pipelines that operate within a fixed set of predefined object categories, the multimodal model is used to generate open-ended textual descriptions of observed civic conditions without task-specific retraining. By emphasizing semantic interpretation rather than pixel-level matching, the approach is intended to reduce sensitivity to non-essential visual variations, such as changes in lighting, weather, or viewpoint, which commonly affect image-based analysis in urban environments.

Standardized and Structured Output Generation

Based on the multimodal interpretation of the input image, the system produces a standardized textual representation that can be used to support downstream automation. The generated output typically includes:

- **Descriptive Summary:** A concise narrative describing the observed civic issue, grounded in visual evidence extracted from the uploaded image.

- **Semantic Indicators:** Identification of visually relevant elements associated with the issue (e.g., apparent waste accumulation or surface damage affecting traffic flow), expressed in natural language to support interpretability.

- **Workflow-Compatible Output:** Structured textual content designed to assist automated routing, prioritization, and administrative review, rather than serving as a definitive assessment.

An illustrative example of a generated description is shown below:

```
{
 "issue_type": "Pothole",
 "severity": "High",
 "description": "Large pothole visible on a roadway, potentially obstructing traffic and posing a safety risk to vehicles and pedestrians."
}
```

Role within the Multimodal Framework

By reducing dependence on user-generated text and supplementing reports with AI-assisted visual descriptions, the system aims to improve the consistency and completeness of civic issue

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

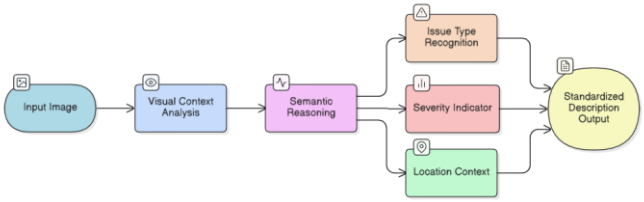
documentation. Automatic description generation functions as an intermediary layer between image-based issue interpretation (Section 4.1) and downstream prioritization and resolution assessment processes. Rather than replacing human judgment, this component is intended to enhance transparency and operational efficiency by providing structured, interpretable information to support municipal workflows.

4.3 AI-Assisted Classification and Prioritization

The AI-assisted classification and prioritization component is designed to support the transformation of multimodal civic reports into structured inputs for municipal workflows. This module functions as a decision-support mechanism that assists administrators in organizing reported issues based on estimated severity and contextual relevance. By providing structured priority indicators, the system aims to support more informed resource allocation while retaining human oversight in final decision-making.

Implementation Logic

The system applies a multidimensional prioritization strategy that processes each reported issue to estimate a relative **Severity Score**, which is subsequently mapped to a priority level (e.g., High, Medium, or Low). This prioritization logic is integrated with the administrative dashboard to assist in task routing and review. Once an issue is analyzed, the system suggests an appropriate municipal department and an associated urgency level, with the objective of reducing manual triaging effort rather than fully automating decision-making.



Severity Score Estimation

The prioritization mechanism synthesizes multiple sources of information to estimate issue severity, including the following factors:

- **Issue Category:** The system infers the likely service domain associated with the reported issue—such as sanitation, electricity, road infrastructure, or water supply—and applies category-specific weighting to reflect general differences in potential impact. Issues associated with public safety or health concerns may be assigned higher relative importance than those with primarily aesthetic impact.

- **Visual Evidence:** Multimodal interpretation of the uploaded image is used to assess observable characteristics of the issue, such as apparent scale, obstruction, or potential risk. For example, the system may distinguish between minor surface irregularities and more substantial road damage that could affect traffic flow or pedestrian safety.

- **Location Context:** Geospatial metadata and contextual cues are used to infer the surrounding environment. Reports originating from high-traffic areas, major roadways, or locations near critical public facilities (e.g., hospitals or schools) may be assigned higher priority indicators compared to those in lower-impact areas.

- **Historical Occurrence:** The system references historical grievance records to identify locations with recurring or persistent issues. A higher frequency of similar reports in the same area may indicate underlying infrastructure problems and can inform escalation recommendations.

Operational Considerations

By aggregating these factors into a unified severity estimate, the proposed framework aims to support more consistent and transparent prioritization of civic issues. Rather than replacing manual judgment, the AI-assisted prioritization process provides structured inputs that can help reduce subjectivity and variability in triaging decisions. This approach is intended to support timely intervention and improved accountability within municipal infrastructure management, while remaining adaptable to policy constraints and human review.

V. AI-BASED RESOLUTION VALIDATION

The AI-based resolution validation component is designed to support accountability in civic issue management by introducing image-based evidence review into the issue closure process. Rather than relying solely on manual status updates or self-reported task completion, the proposed framework incorporates multimodal analysis of visual evidence to assist administrators in assessing whether reported issues appear to have been addressed on the ground.

This component is intended to reduce ambiguity in issue closure workflows by supplementing administrative records with structured visual context. It functions as a decision-support mechanism rather than an autonomous verification system, with final resolution decisions remaining under human oversight.

5.1 Before–After Image Comparison

Resolution assessment is formulated as a semantic comparison task involving visual evidence captured at two different stages of the issue lifecycle.

Implementation Workflow

- **Baseline Image (Before):** When a civic issue is reported by a citizen, a geotagged image is captured through the reporting interface and stored as baseline visual evidence representing the unresolved condition.

- **Post-Resolution Image (After):** Upon task completion, a field worker submits a second image of the same location using a worker-facing interface. This image represents the claimed post-resolution state.

- **Multimodal Analysis:** The baseline and post-resolution images are jointly analyzed using a pre-trained Multimodal Large Language Model (e.g., Gemini-1.5 Pro). Instead of relying on pixel-level differencing, the model is used to perform semantic reasoning over the image pair to assess whether meaningful visual changes relevant to the reported issue are present.

Semantic Evaluation Criteria

The multimodal analysis considers multiple semantic aspects of the before–after image pair:

- **Location Consistency:** The system assesses whether both images plausibly correspond to the same physical location by identifying persistent contextual elements such as road layout, surrounding structures, or nearby infrastructure. The analysis is designed to reduce sensitivity to non-essential variations such as lighting, weather, or viewpoint differences.

- **Relevant Visual Change:** The model examines whether visual differences between the images are consistent with the expected outcome of the reported repair activity (e.g., removal of accumulated waste, repair of visible road damage, restoration of public utilities).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

- **Resolution Adequacy:** Based on semantic interpretation, the system estimates whether the originally reported condition appears to be no longer present in the post-resolution image. Cases where changes appear partial, cosmetic, or ambiguous can be flagged for further human review.

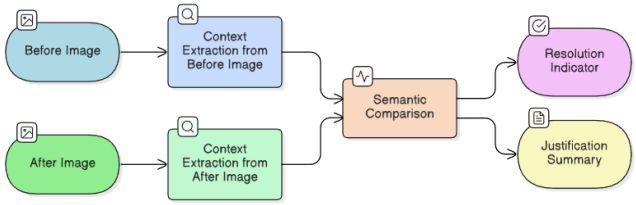
System Output

The resolution assessment process produces a structured output intended to support administrative decision-making:

- **Resolution Indicator:** A suggested status indicating whether the issue appears to be resolved or may require additional review.
- **Confidence Estimate:** A relative confidence value reflecting the model's internal assessment of its interpretation, intended to guide prioritization of manual inspection.
- **Justification Summary:** A concise natural language explanation grounded in visual evidence (e.g., "The waste accumulation visible in the baseline image is no longer present in the post-resolution image, and the surrounding area appears cleared.").

Operational Role

By incorporating AI-assisted analysis of before–after visual evidence, the proposed system aims to enhance transparency and consistency in issue closure workflows. Rather than eliminating errors or guaranteeing correctness, this mechanism provides structured, interpretable signals that can help identify potentially incomplete or ambiguous closures and support more informed administrative review. In doing so, the framework contributes to improved accountability and trust in digital civic governance processes while remaining adaptable to policy constraints and human oversight.



6. Experimental Setup and Evaluation

As the proposed system relies on pre-trained Multimodal Large Language Models (MLLMs) accessed through inference APIs and does not involve task-specific training or fine-tuning, evaluation is conducted at the system and workflow level rather than through conventional supervised learning benchmarks. The primary objective of this evaluation is to examine how multimodal reasoning can support civic issue reporting, prioritization, and resolution assessment within an operational governance workflow.

6.1 Dataset Description

The evaluation considers a representative collection of crowdsourced civic images obtained from citizen reports and field worker submissions. The dataset reflects real-world urban environments and includes visual variability arising from differences in lighting conditions, weather, camera viewpoints, and surrounding context.

For analysis purposes, the dataset is organized into broad civic issue categories, including:

- Road infrastructure issues (e.g., potholes)
- Waste management concerns (e.g., garbage accumulation)
- Public utility faults (e.g., damaged streetlights)
- Water infrastructure issues (e.g., leakages and drainage blockages)

Each reported issue is associated with:

- A geotagged **baseline (before)** image submitted by a citizen
- Optional user-provided textual input
- A corresponding **post-resolution (after)** image submitted by a field worker

This paired-image structure enables evaluation of image-based issue interpretation as well as semantic comparison for resolution assessment.

6.2 Evaluation Criteria and Metrics

System performance is evaluated using a combination of quantitative indicators and qualitative assessment methods commonly employed in civic analytics and applied multimodal AI systems. Since the system operates using pre-trained models, these metrics are used to assess **workflow effectiveness** rather than model training performance.

Task Component	Metric / Indicator	Evaluation Objective
Issue Routing Support	Agreement Rate / Accuracy	Consistency of AI-assisted routing with administrative review
Description Quality	BLEU Score / Human Review	Clarity, completeness, and semantic relevance of generated descriptions
Resolution Assessment	Precision / Recall (Sampled Review)	Alignment between AI-assisted resolution indicators and human judgment
Workflow Efficiency	SLA Trend Analysis	Observed changes in response and closure timelines

Where applicable, human evaluation is incorporated to account for the subjective and context-dependent nature of civic issue assessment.

6.3 Observational Outcomes

Based on the system design and prior findings in multimodal reasoning research, the evaluation focuses on observing the following system-level trends:

- **Workflow Streamlining:** AI-assisted description generation and prioritization may reduce manual triaging effort, potentially leading to faster issue routing and assignment.
- **Improved Closure Review:** Semantic comparison of before–after images provides structured visual evidence that can assist administrators in identifying ambiguous or potentially incomplete closures.
- **Enhanced Reporting Consistency:** Standardized AI-generated descriptions may improve uniformity in issue documentation across reports submitted by diverse users.
- **Administrative Support:** Municipal stakeholders may benefit from more structured and interpretable data to support decision-making, auditing, and performance monitoring.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

These observations are intended to evaluate the practical applicability of multimodal AI within civic governance workflows rather than to claim definitive performance gains.

7. Limitations and Challenges

While the proposed system demonstrates the potential of multimodal AI to support civic issue reporting and resolution workflows, several limitations and practical challenges must be acknowledged.

First, the effectiveness of image-based analysis depends heavily on the quality and relevance of the images provided by citizens and field workers. Poor lighting, occlusions, extreme camera angles, or insufficient visual coverage of the reported issue can limit the reliability of multimodal interpretation. In such cases, AI-generated descriptions or resolution assessments may be ambiguous and require additional human review.

Second, Multimodal Large Language Models are general-purpose systems and may occasionally produce incorrect or incomplete interpretations, particularly in visually complex or unfamiliar urban scenarios. Although the framework is designed to emphasize semantic reasoning, it cannot fully eliminate risks such as misinterpretation or hallucinated descriptions. For this reason, the system is intended to function as a decision-support tool rather than an autonomous authority, with human oversight remaining essential.
Third, the proposed approach relies on cloud-hosted, pre-trained multimodal models accessed through inference APIs. This introduces practical constraints related to inference latency, operational cost, and dependency on third-party services. Large-scale municipal deployment may require careful optimization, selective invocation of AI services, or hybrid strategies combining automated analysis with manual workflows.

Fourth, geospatial and contextual inference may be affected by inaccuracies in GPS data or limited availability of identifiable landmarks in certain environments. Dense urban areas, informal settlements, or rapidly changing construction zones can pose additional challenges for consistent location interpretation and historical comparison.

Finally, institutional adoption presents non-technical challenges. Integrating AI-assisted workflows into existing municipal governance structures requires policy alignment, staff training, and trust in AI-generated outputs. Resistance to automation, varying administrative practices across regions, and legal or regulatory constraints may influence the pace and extent of real-world deployment.

These limitations highlight the importance of human-in-the-loop design, incremental deployment, and continuous evaluation. Addressing these challenges through improved data collection practices, adaptive system tuning, and governance-aware deployment strategies represents an important direction for future work.

8. CONCLUSION

This paper presented an AI-assisted, crowdsourced civic issue reporting and resolution framework that leverages multimodal large language models to support image-based issue understanding, standardized documentation, prioritization, and resolution assessment. By shifting from text-centric reporting to a vision-first, multimodal workflow, the proposed system aims to improve consistency, transparency, and operational efficiency in municipal grievance redressal processes.

Rather than introducing new model architectures, the work focused on the system-level integration of existing multimodal AI capabilities into practical civic governance workflows. The framework demonstrates how semantic interpretation of visual evidence can assist municipal authorities in triaging issues, assessing resolution claims, and improving accountability, while maintaining human oversight in decision-making.

Through a modular architecture and workflow-level evaluation, the study highlights the feasibility of applying multimodal reasoning to real-world smart city contexts characterized by visual variability, limited user input, and operational constraints. The proposed approach emphasizes interpretability and evidence-based support over rigid automation, aligning with the needs of public-sector governance systems.

Future work may explore extended pilot deployments, deeper human–AI interaction studies, and integration with additional data sources such as sensor streams or historical maintenance records. As multimodal AI technologies continue to mature, their responsible and context-aware integration into civic infrastructure management holds significant promise for enhancing urban governance and citizen engagement.

9. REFERENCES

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A., You Only Look Once: Unified, Real-Time Object Detection, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[2] Ren, S., He, K., Girshick, R., & Sun, J., Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
[3] Maeda, H., Sekimoto, Y., Seto, T., Kashiyama, T., & Omata, H., Road Damage Detection Using Deep Neural Networks with Images Captured Through a Smartphone, *arXiv preprint arXiv:1801.09454*, 2018.
[4] Mandal, V., Uong, L., & Adu-Gyamfi, Y., Automated Road Crack Detection Using Deep Convolutional Neural Networks, *IEEE International Conference on Big Data*, 2018.
[5] Mittal, G., Yagnik, K. B., Garg, M., & Khare, D., SpotGarbage: Smartphone App to Detect Garbage Using Deep Learning, *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.
[6] Dosovitskiy, A., et al., An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale, *International Conference on Learning Representations (ICLR)*, 2021.
[7] Radford, A., Kim, J. W., Hallacy, C., et al., Learning Transferable Visual Models From Natural Language Supervision, *International Conference on Machine Learning (ICML)*, 2021.
[8] Li, J., Li, D., Xiong, C., & Hoi, S., BLIP: Bootstrapping Language–Image Pre-training for Unified Vision-Language Understanding and Generation, *International Conference on Machine Learning (ICML)*, 2022.
[9] Alayrac, J.-B., et al., Flamingo: A Visual Language Model for Few-Shot Learning, *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
[10] OpenAI, GPT-4 Technical Report, *arXiv preprint arXiv:2303.08774*, 2023.
[11] Google DeepMind, Gemini: A Family of Highly Capable Multimodal Models, *Technical Report*, 2023.
[12] Sakurada, K., & Okatani, T., Change Detection from a Street Image Pair Using CNN Features and Superpixel Segmentation, *British Machine Vision Conference (BMVC)*, 2015.
[13] Daudt, R. C., Le Saux, B., & Boulch, A., Fully Convolutional Siamese Networks for Change Detection, *IEEE International Conference on Image Processing (ICIP)*, 2018.
[14] Batty, M., Smart Cities, Big Data and Urban Analysis, *Journal of Urban Technology*, vol. 23, no. 2, pp. 3–21, 2016.
[15] Nam, T., & Pardo, T. A., Conceptualizing Smart City with Dimensions of Technology, People, and Institutions, *Proceedings of the 12th Annual International Digital Government Research Conference*, 2011.
[16] Kitchin, R., The Ethics of Smart Cities and Urban Science, *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2083, 2016.