VLM Benchmark Release Report

VLM Benchmark for Smart City Governance:

Kaohsiung Local Sovereign AI

Item	Description
Document Title	VLM Benchmark for Smart City Governance: Kaohsiung Local Sovereign AI
Authors	Kaohsiung City Government & Linker Vision Co., Ltd.
Release Version	Version 1.0
Release Date	July 2025
Document Type	VLM Benchmark Methodology Guidelines
Citation	Kaohsiung City Government & Linker Vision. (2025). VLM Benchmark for Smart City Governance – Kaohsiung Local Sovereign AI. Version 1.0
Image Sources	Some images are provided by Kaohsiung City Government and are authorized for use solely within this report. Unauthorized use or redistribution is prohibited.
Contact	https://www.linkervision.com

1. Introduction

With the continual and rapid advancement of Vision-Language Models (VLMs) within the broader field of artificial intelligence, their potential applications have significantly expanded to include a diverse range of complex real-world domains, notably encompassing smart city development, industrial governance frameworks, and data-driven community management initiatives. Despite this technological growth, most contemporary evaluation frameworks for VLMs remain predominantly academic and research-oriented. They often lack comprehensive validation mechanisms and practical testing methodologies that are truly aligned with real-world governance challenges and operational constraints.

Drawing from a variety of cross-departmental, scenario-based use cases provided by the Kaohsiung City Government, this research proposes the establishment of a VLM Benchmark specifically tailored to smart city applications. This benchmark, developed with high fidelity to practical urban governance, integrates representative event scenarios, rigorous question type design, and a standardized evaluation protocol. The overarching goal is to provide a reference system for evaluating model performance that maintains both academic integrity and practical relevance.

This proposed benchmark rests upon three fundamental design principles:

- **Alignment with localized application logic**: All assessment items are formulated based on real-world urban scenarios and governance needs identified in Kaohsiung City, ensuring the evaluation content is not only contextually relevant but also grounded in policy and operational logic.
- **Capacity for multi-variable testing**: The benchmark accounts for diverse scene variations and environmental conditions—ranging across diurnal cycles (day and night), spatial contexts (indoor and outdoor), and meteorological variables (such as clear, cloudy, and rainy weather)—to

simulate the multifaceted and dynamic nature of urban environments.

- **Scalability and potential for global interoperability**: The structure of the benchmark was designed with compatibility in mind, referencing internationally recognized evaluation paradigms such as MMBench. This allows the framework to support both local adaptation and comparative international benchmarking.

To realize this goal, the benchmark construction process incorporated governance requirements and visual data contributions from eight distinct departments and affiliated public agencies, covering domains such as transportation planning, environmental supervision, emergency service coordination, infrastructure management, and urban management. A multi-stakeholder VLM Benchmark Committee was assembled to guide the project, composed of end-user representatives from municipal government, system integrators and AI developers, as well as academic researchers. Through cross-sector collaboration, this committee jointly defined data selection criteria, question development workflows, and review protocols to uphold authenticity, representativeness, and fairness across all benchmark components.

2. Current Status and Related Work

2.1 Existing Benchmarks

In the field of multimodal model evaluation, numerous datasets and benchmarking frameworks have been widely circulated and cited within academia, including COCO Caption, VQAv2, ScienceQA, GQA, and more recently, MMBench. While these benchmark resources offer certain effectiveness and reference value in verifying the capabilities of general-purpose VLM models under controlled conditions, they still face three major limitations when extended to the context of smart city governance.

First, most existing benchmarks focus heavily on “static scenarios,” which primarily involve object recognition of specific items or scenes within an image. Such tasks are relatively mature within the traditional object detection framework, supported by well-established datasets. However, in the context of urban governance, “dynamic scenarios” are far more common—for instance, determining whether a traffic accident has occurred or whether flooding has caused road closures. These situations not only require multi-layered semantic understanding but also involve inference about the progression and potential impact of events. VLM tasks designed for city governance aim to enable models to judge the occurrence of accidents, assess their influence on specific areas, and even observe public behavioral responses using only image-based inputs, thus surpassing the interpretive limitations of static models.

Second, although many international benchmarks have accumulated substantial resources, their design is often centered on general-purpose applications, lacking support for governance-specific semantics in diverse environments. In contrast, the “City Governance VLM Benchmark” is grounded in actual urban scenarios from Kaohsiung, constructed using localized street surveillance footage and emphasizing the alignment between images and textual descriptions. This design enables models to more accurately reflect the reasoning logic and contextual characteristics required for urban governance. For example, distinguishing between different levels of flooding and how they affect traffic flow on major roads can directly inform municipal response strategies.

Third, most existing datasets emphasize isolated, single-question formats, lacking the coherence and hierarchical structure that mirrors real-world city decision-making logic. To address this limitation, the project introduces a layered question framework divided into L1 and L2 levels. L1 is used to categorize event scenarios, while L2 addresses detailed follow-up questions based on L1 responses. For instance, if the first L2 question is “Has a traffic accident occurred?” and the answer is “Yes,” subsequent questions will explore the type of accident, the number of involved vehicles, and whether specific traffic lanes are affected. However, if the answer is “No,” further questions are skipped. This structure not only conserves computational resources but also aligns closely with real governance logic. Such a layered design enhances the integrity of the model’s reasoning chain and mirrors the actual response and decision-making processes of municipal governments, ensuring that the AI model’s operational logic is highly consistent with the needs of governance..

2.2 Gaps in Smart City Application

AI deployments in the smart city domain—especially those intended for frontline, citizen-facing tasks—demand sophisticated comprehension of diverse information sources and robust cross-modal reasoning. This is particularly evident in scenarios involving environmental ambiguity or infrastructural uncertainty. For example, accurately interpreting a photographic image depicting a temporary construction zone on a cloudy evening requires the model to conduct visual object recognition (identifying elements such as barricades or caution signs), infer spatial relationships (such as the proximity between pedestrians and hazard zones), and apply contextual governance rules (e.g., identifying violations or triggering alert mechanisms).

Although these tasks may appear conceptually simple, the execution requires a high degree of contextual awareness, multi-step inference, and domain-specific knowledge. Without a dedicated benchmark framework that reflects these complexities, it is exceedingly difficult to evaluate whether an AI model is adequately prepared for deployment in high-stakes, real-time urban governance settings. Such a gap emphasizes the critical need for new evaluation structures that go beyond generic accuracy measures and incorporate deeper dimensions of practical performance.

3. Data Sources and Construction Principles

3.1 Participating Departments and Dataset Scope

This benchmark was co-developed through the joint efforts of eight governmental departments and state-owned enterprises based in Kaohsiung City. These include the Transportation Bureau, Sports Development Bureau, Mass Rapid Transit Bureau, Water Resources Bureau, Public Works Bureau, and three major public sector enterprises—Taiwan International Ports Corporation, China Steel Corporation and Taiwan Power Company (Taipower). Each participating unit contributed a set of domain-relevant governance scenarios based on their daily operational experiences. These scenarios were later formalized into test items with the support of Information Technology Office, which served as the coordinating lead.

In summary:

- **Total participating units**: 8 (comprising 5 city government departments and 3 state-owned enterprises)
- **Core governance scenarios compiled**: 108
- **Finalized test items**: 609 questions derived from practical operational contexts (dynamic adjustment based on the requirements.)
- **Question types**: Binary classification, category recognition, ordinal ranking tasks
- **Design approach**: Scenario-driven, governance-informed, and aimed at multi-dimensional evaluation diversity

3.2 Question Types and Structure

To systematically assess different aspects of model capability, the benchmark questions were categorized into three primary types:

- **Binary Classification**: Designed to determine the presence or absence of specific events (e.g., traffic violations, obstruction).
- **Category Recognition**: Aims to classify the situation or object type, such as identifying the category of an accident (collision, rollover, scratch, etc.).

- **Ordinal/Ranked Scenarios**: Intended to gauge the model's ability to assess severity levels, such as rating the degree of traffic obstruction from level A to F.

Each of these categories reflects a key cognitive skill required for governance reasoning and real-time decision-making. In the final scoring system, appropriate weighting and accuracy thresholds were defined per type.

The illustrative examples presented below feature representative camera footage provided by Water Resources Bureau, Economic Development Bureau, and Public Works Bureau. These demonstrate that every captured scenario can be formulated into three clearly defined categories of questions.

Water Resources Bureau:

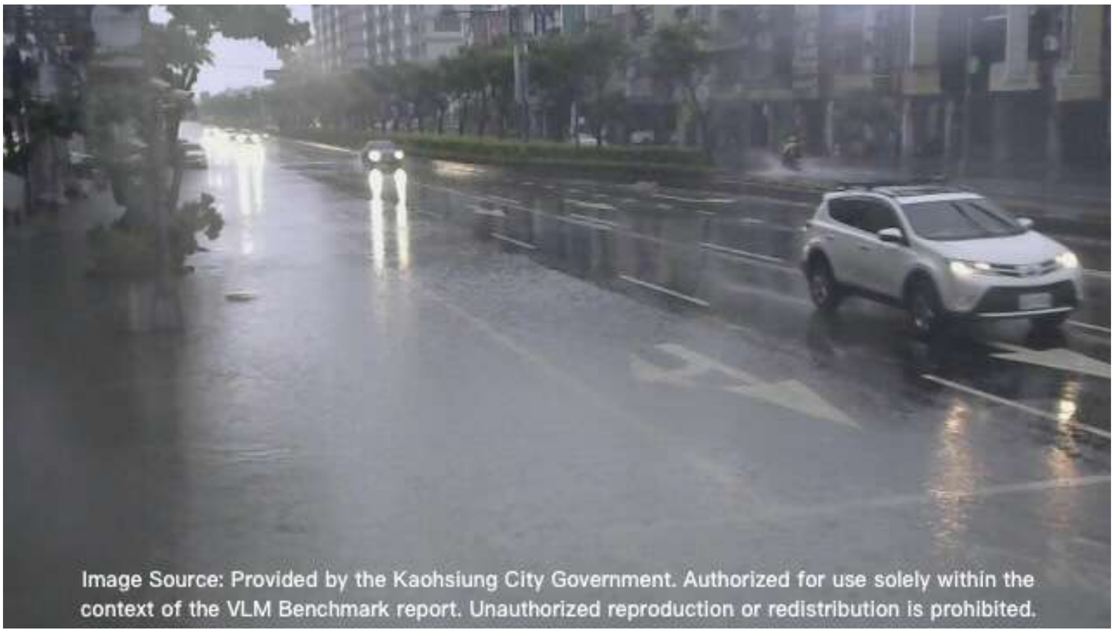


Image Source: Provided by the Kaohsiung City Government. Authorized for use solely within the context of the VLM Benchmark report. Unauthorized reproduction or redistribution is prohibited.

1. Is there flooding or water accumulation: Yes / No
2. Affected road section due to flooding: No impact / Single (one-way) or partial lane / Entire (both directions) road or intersection
3. Water accumulation depth: Shallow (< 30 cm) / Moderate (30–50 cm) / Deep (> 50 cm)
4. Is the flooded road closed off: Closed / Not closed
5. Are vehicles present in the flooded area: Yes / No

6. Is vehicle passage possible through the flooded area: Yes / No
7. Are pedestrians present in the flooded area: Yes / No
8. Is pedestrian passage possible through the flooded area: Yes / No
9. Are warning signs present in the flooded area: Yes / No
10. Is the drainage inlet blocked: Yes / No

Sports Development Bureau:

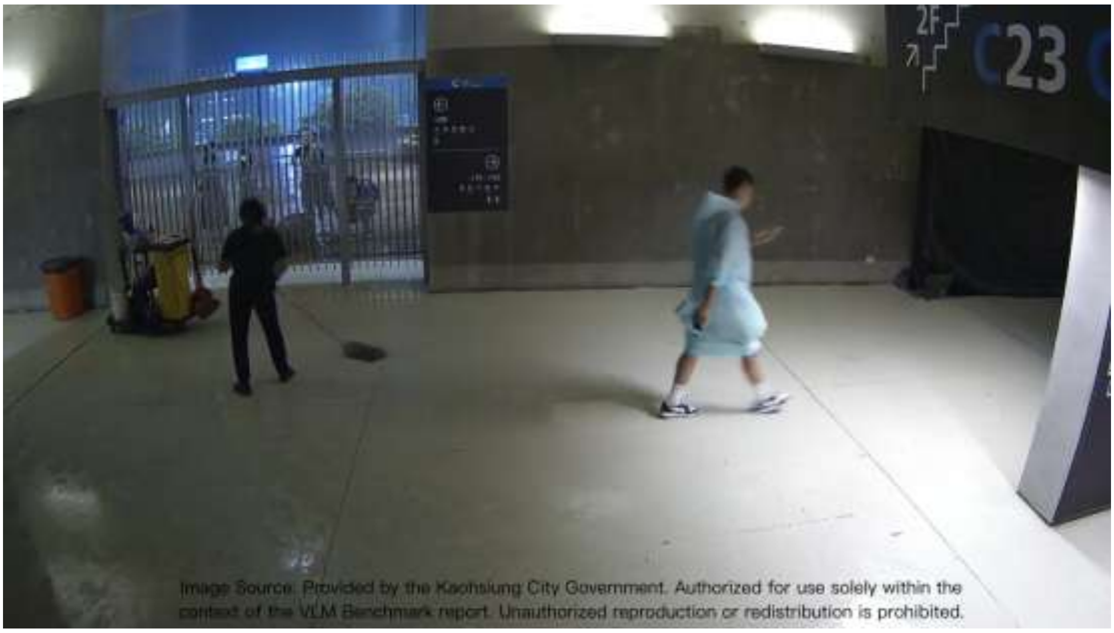


Image Source: Provided by the Kaohsiung City Government. Authorized for use solely within the context of the VLM Benchmark report. Unauthorized reproduction or redistribution is prohibited.

1. Is there an emergency exit: Yes / No
2. Is the emergency exit blocked: Yes / No
3. Type of blockage at the emergency exit: Crowd / Debris / Trash / Other / More than one
4. Degree of blockage at the emergency exit: Partially blocked (passable) / Completely blocked (impassable)
5. Are the obstructions at the emergency exit movable (including people, vehicles): Yes / No

Public Works Bureau:

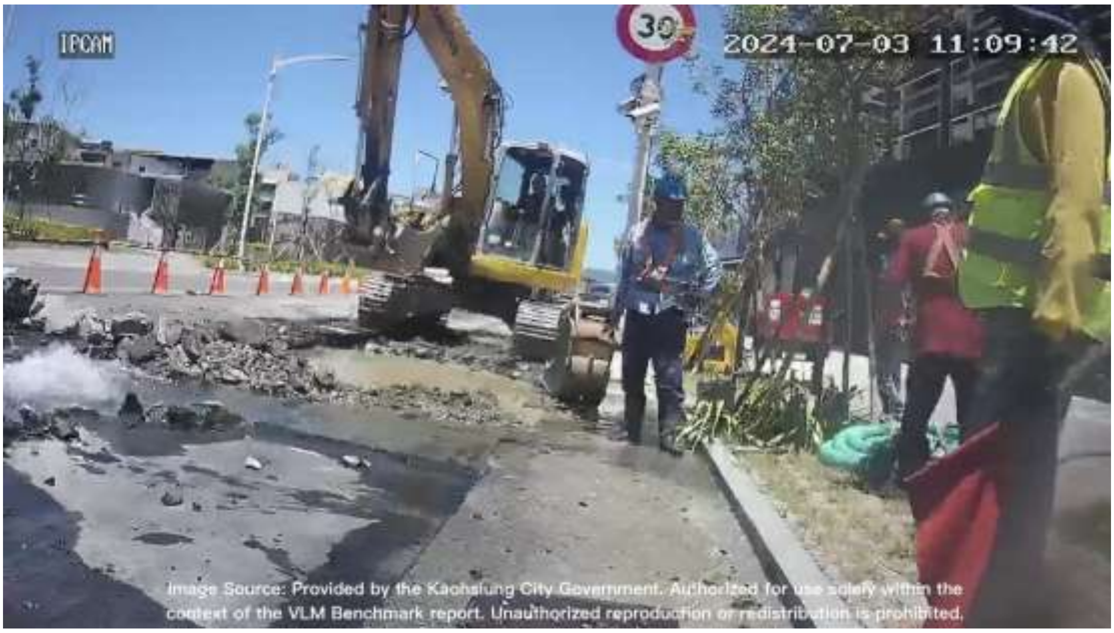


Image Source: Provided by the Kaohsiung City Government. Authorized for use solely within the context of the VLM Benchmark report. Unauthorized reproduction or redistribution is prohibited.

1. Is there road construction occupying lanes and affecting traffic: Yes / No
2. Type of construction: Road repair / Road milling and paving / Sidewalk construction / Road excavation / Aerial lift work (e.g., streetlight or tree maintenance) / Road widening / Bridge construction / Road occupation by building site (equipment and materials) / Other
3. Are there construction signs or indicators: Yes / No
4. Are traffic control measures in the construction area adequate (e.g., traffic cones with connectors, steel barriers, Jersey barriers): Yes / No
5. Is fencing installed around the roadside construction area: Yes / No
6. Do traffic control measures include nighttime warnings: Yes / No
7. Are construction personnel present: Yes / No
8. Time of scene: Daytime / Nighttime / Other

3.3 Dataset Volume and Design Targets

To ensure international comparability, the benchmark design referenced leading evaluation frameworks frequently used to assess VLMs such as VILA and LLaVA. For example, ScienceQA contains 4,241 labeled items, MME includes 2,374, and MMBench comprises 1,784. Based on these precedents, this benchmark includes approximately 1,400 unique answer options. Each option is paired with at least four distinct images, all of which have been manually reviewed, resulting in a total of approximately 5,600 high-quality images in the dataset.

Design targets are as follows:

- **Baseline coverage**: 4 images per answer option
- **Expanded objective**: 15 images per option
- **Initial dataset scale**: ~5,600 image-question pairs
- **Maximum target**: 20,000 items if image sourcing capacity improves

To ensure fairness in evaluation and prevent the model from being exposed to test content in advance (i.e., to avoid data leakage), it is essential to verify that the test images have never appeared in the training dataset of the model being evaluated. This procedure is a critical step in enhancing the credibility of model testing.

3.4 Data Selection Rules and Scenario Planning

A structured, three-layer data selection strategy was implemented:

1. **Scene Type Differentiation**: Indoor scenes were excluded from time-of-day and weather constraints, while outdoor scenes were tagged with complete environmental parameters.
2. **Time-based Distribution**: Dataset images were distributed based on real-world collection ratios—daytime: 75%, dusk: 2%, nighttime: 23%.
3. **Weather-based Distribution**: Aligned with Kaohsiung's 10-year climatological data—sunny: 52%, cloudy: 20%, rainy: 28%.

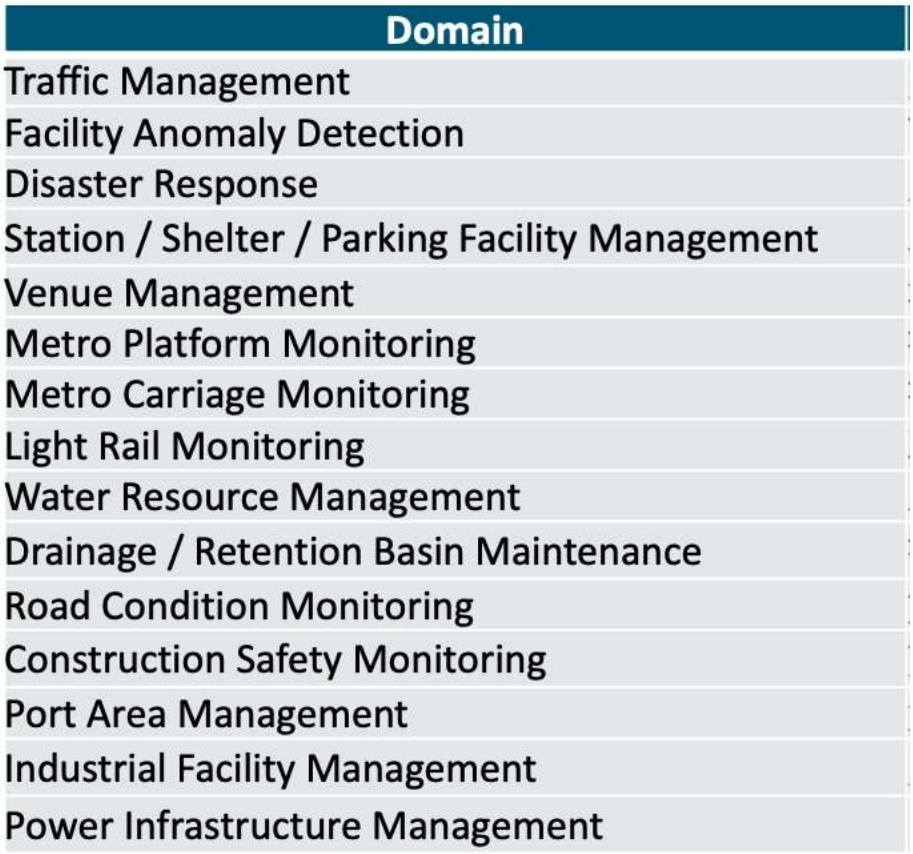
This approach ensures scenario realism and contributes to comprehensive evaluation across variable operational conditions.

3.5 Categorization of Urban Governance Domains

While categorizing governance tasks by government departments provides administrative clarity, it often leads to overlapping responsibilities and task ambiguity in model design and classification. To enhance data structure and enable cross-domain scalability, this benchmark adopts a domain-based categorization of data sources and scenario tasks—focusing on the nature of tasks and visual scene characteristics rather than organizational boundaries.

Each domain represents a specific type of urban surveillance imagery or event pattern, allowing for question designs better aligned with visual reasoning needs. This task-oriented logic also facilitates cross-departmental applicability. For example, "road condition monitoring" may involve both the Public Works Bureau and the Water Resources Bureau, while "disaster response" may draw on inputs from both the Transportation and Urban Development Bureaus.

Based on this framework, we define a set of urban governance domains covering traffic infrastructure, public utilities, water management, metro operations, port areas, and industrial facilities.



Domain
Traffic Management
Facility Anomaly Detection
Disaster Response
Station / Shelter / Parking Facility Management
Venue Management
Metro Platform Monitoring
Metro Carriage Monitoring
Light Rail Monitoring
Water Resource Management
Drainage / Retention Basin Maintenance
Road Condition Monitoring
Construction Safety Monitoring
Port Area Management
Industrial Facility Management
Power Infrastructure Management

3.6 Privacy Protection and De-Identification Procedures

Some of the image data used in this benchmark originate from real-world street scenes and public surveillance systems. To comply with Article 6 of Taiwan’s Personal Data Protection Act regarding the handling of sensitive personal information, and in alignment with the European Union’s General Data Protection Regulation (GDPR), all image data must undergo de-identification processing—such as the removal of facial features and license plate information—prior to public release or interface display.

The de-identification process for this benchmark is conducted on a case-by-case basis, based on the original materials submitted by participating agencies. Methods include blurring, masking, and other anonymization techniques. A version control log is maintained to ensure transparency and traceability of all data handling procedures. This approach balances practical model performance with legal privacy compliance, fulfilling both smart governance needs and ethical responsibilities.

4. Evaluation Design and Scoring Mechanism

4.1 Evaluation Workflow

To simulate field-level deployment conditions, the benchmark's evaluation procedure incorporates repeated testing cycles and semantic normalization. This allows for assessment of both predictive accuracy and output stability under variation.

Core workflow components:

- **Multiple inputs per test item**: Each question is linked to multiple visual stimuli to test consistency.
- **Answer Standardization:** Model-generated responses are mapped to the predefined answer options established by the benchmark for scoring purposes. The benchmark includes multiple question types—namely, binary classification, categorical identification, and ordinal/hierarchical classification—each associated with distinct scoring rubrics and evaluation criteria. (Detailed scoring methodologies for each type are outlined in the subsequent section.)
- **Performance metrics**: Accuracy and recall scores are captured and analyzed.

4.2 Scoring Rules by Question Type

Each question type is paired with tailored scoring rules:

- **Binary Classification**: Simple correct/incorrect scoring.
- **Category Recognition**: Full credit for precise matches, partial credit for semantically adjacent categories.
- **Ordinal/Ranking Tasks**: Proportional scores based on distance from the correct rank.

All answers are aligned with domain-specific governance logic and validated through expert review.

4.3 Flexible Threshold Adjustment

To reflect real-world difficulty, scoring thresholds are dynamically adjusted based on:

- **Data scarcity or collection difficulty**
- **Task-level inference complexity**
- **Scene condition rarity or ambiguity**

Such flexibility ensures fair benchmarking even in cases of rare scenarios, while also encouraging model improvement in traditionally underperforming domains.

---

5. Preliminary Validation and Interim Results

We have progressively completed early-stage evaluation tasks across a variety of test items that reflect key domains of urban governance. These include traffic accidents, roadway violations, environmental anomalies, and surveillance of community public spaces. The responses generated by VLM systems under evaluation have verified that this benchmarking framework is effective in uncovering typical model errors and behavioral biases—particularly those arising from insufficient training data or specific scene complexities. Notable examples include significant performance degradation under nighttime rain conditions and blurred, distant imagery, as well as confusion among semantically similar options in classification tasks. These analyses provide essential insights that inform targeted fine-tuning and dataset augmentation in future model development phases.

---

6. Future Directions

In summary, the Smart City VLM Benchmark established through this project has evolved beyond a conventional academic dataset for capability evaluation. It now serves as a policy-driven, multi-agency collaborative, locally sovereign, practically validated, and globally extendable AI evaluation platform. Through close cooperation among governance units of Kaohsiung City

Government, technology developers, and academic research communities, we have realized a closed-loop paradigm in which data and evaluation items are generated from real needs, and model outputs provide direct feedback for decision-making processes.

We will continue refining and localizing the Kaohsiung-specific VLM Benchmark to support dataset standardization and fairness efforts. This initiative aims to promote broader inter-municipal collaboration and co-development, driving progress toward exporting this standard internationally and making it a standard AI testing criterion for urban governance in various countries.