VLM Benchmark 發布報告書

智慧城市治理 VLM 評測基準：高雄在地主權 AI

項目	說明
文件標題	智慧城市治理 VLM Benchmark：高雄在地主權 AI
作者	高雄市政府、鑫蘊林科股份有限公司
發佈版本	1.0 版本
發佈時間	2025 年 7 月
文件類型	VLM 評測基準指引
引用建議格式	高雄市政府 & Linker Vision（2025）。《智慧城市治理 VLM Benchmark：高雄在地主權 AI》。第 1.0 版。
圖片來源說明	本報告內使用部分圖像素材由高雄市政府提供，僅限於本報告撰擬，未經授權不得翻拍、重製或外流。
聯絡方式	https://www.linkervision.com

1. 前言

隨著視覺語言模型（Visual Language Models, 簡稱 VLMs）在人工智慧應用領域中的迅速發展，其潛在應用價值與實際落地能力日益受到重視。VLM 技術結合視覺處理與語意理解能力，近年來廣泛應用於多模態任務，包括圖像敘述、跨模態搜尋、即時問答等。尤其在智慧城市推動過程中，AI 系統若能具備強大的情境理解力與事件判讀能力，將成為城市治理效能提升的重要技術支撐。

儘管技術面發展快速，當前主流之 VLM 評測基準大多仍偏重於學術實驗場景，缺乏對於真實治理需求與操作環境的深度回應。評測資料常來自靜態圖像與標準化任務，與實際城市中多樣、即時與不確定性的挑戰仍存落差。為回應此一問題，並推動技術落地實證，本報告提出專為智慧城市應用情境設計的 VLM 評測基準，作為 AI 系統部署前之能力驗證工具。

該基準藉由高雄市政府提供的跨局處、跨場域之治理案例，整合視覺資料與場景敘述，設計具代表性、具挑戰性且能模擬真實任務的多模態測試題型，並配合結構化的評分流程與彈性調整機制，以建立一套兼顧學術嚴謹性與在地實用性的評估框架。

此一評測基準的開發，依循下列三大核心原則：

- 根植在地應用邏輯：所有測試資料與任務設計皆來自高雄市實際治理場景與政策需求，確保題材具備實用脈絡與在地對應性，提升技術評估的真實性與可行性。
- 涵蓋複合環境與多元變因：測試題型設計包含日夜時段、天候條件（晴、陰、雨）、空間類型（室內與室外）等不同變因，模擬城市環境中常見但具挑戰性的動態場景，增強模型適應性驗證強度。
- 符合國際標準，具全球擴散潛力：本評測基準之結構設計參照國際指標如 MMBench，強調系統兼容性，使其得以支援本地化應用之餘，亦利於國際橫向比較與指標整合。

為實現上述目標，本評測基準之建構歷程整合來自八個政府局處與附屬公共單位之治理需求與視覺資料，範圍涵蓋交通規劃、環境監理、緊急應變、基礎建設維管與城市管理等多元領域。同時成立跨領域之「VLM 評測基準工作小組」，由市府端使用者代表、系統整合及 AI 開發單位，以及學術研究人員共同組成。透過此多方合作平台，共同訂定資料選擇原則、題型設計流程與審查機制，以確保整體基準體系之真實性、代表性與公平性。

2. 現況分析與研究參考

2.1 現有評測架構概述

在多模態模型評估領域中，現今已有多套資料集與評測架構於學術界中廣泛流通並被引用，包含 COCO Caption、VQAv2、ScienceQA、GQA，以及近年所推出的 MMBench 等。儘管上述基準資源對於驗證一般性 VLM 模型在控制性條件下之能力表現具有一定效能與參考價值，然在擴展至智慧城市治理情境時，仍普遍面臨以下三項主要侷限：

首先，現有 benchmark 多數側重於「靜態情境」，即以物件辨識畫面中特定物件或場景為主，這類任務於傳統物件偵測架構中已有相對成熟的應用與資料集。然而，在城市治理場域中，常見的則為「動態情境」，例如是否發生車禍、是否因淹水造成交通中斷等，不僅需進行多層次語意理解，更需推論事件的發展脈絡與潛在影響。城市治理所設計的 VLM 任務，即嘗試讓模型僅透過影像畫面，即能判斷事故是否發生、是否影響特定區域，乃至觀察民眾反應行為，進一步突破靜態模型判讀的限制。第二，目前多數國際 benchmark 雖已累積龐大資源，惟其應用語境多聚焦於通用性，缺乏對不同環境之治理語意的支持。「城市治理 VLM benchmark」則以高雄實際場景為基礎，從在地化的街景監視影像出發，並強調影像與文字之對應關係建構。

此設計有助於模型更貼合城市治理所需之判讀邏輯與語境特徵，例如辨識不同淹水深度對車道通行性的差異，即可作為城市應變處置之依據。第三，多數既有資料集在問題設計上，偏重單一、片段式問答，未能對應實務中城市決策邏輯之連貫性與層次性。為解決此一限制，本專案提出 L1/L2 分層設計架構，L1 用來分類事件樣態情境，L2 為對應 L1 的細節問題。舉例來說，L2 第一題問題若為「是否發生車禍」，當回答為「是」，再 L2 延續第二題以後的分析，包含車禍類型、肇事車輛數量、是否影響特定車道等；但若第一題回答為「否」，則後續題目不再啟動，如此可節省計算資源並符合真實治理邏輯。此種分層式設計不僅可強化模型推論鏈的完整性，也更貼近市府實際在面對事件時的應變與決策流程，讓 AI 模型的運作邏輯與治理需求高度一致。

2.2 智慧城市應用場域之落差

於智慧城市應用領域中，特別是面向市民端之第一線任務場景，AI 模型需具備能整合異質資訊來源並進行跨模態推理之能力。此一需求在具高環境模糊性或基礎建設不確定性的場景中更為明顯。例如，一張拍攝於陰天傍晚的臨時施工區照片，其理解過程需包含視覺物件辨識（如辨認圍欄、警告標誌等）、空間關係推論（如行人與施工區

之距離），以及治理規則的應用（如判斷是否違規、是否觸發告警機制）。

儘管上述任務在概念上看似單純，實際執行時卻需高度的情境感知、多步驟的邏輯推演與結合領域知識之複合能力。倘若欠缺一套專門反映上述特性之評測基準架構，將難以準確評估 AI 模型是否已具備足以部署於即時且高風險之城市治理場域中的實務能力。此種能力缺口突顯出傳統評估方式所無法涵蓋的實務驗證需求，亦反映出當前急需建立一套超越一般準確率量測，能深入評估模型真實執行效能之創新評測架構。

---

3. 資料來源與建構原則

3.1 單位參與與資料規模說明

本評測基準之開發係由高雄市政府轄下八個局處及國營事業單位共同協力完成。參與單位包含高雄市交通局、運發局、捷運局、水利局、工務局，以及三家國公營事業單位：臺灣港務公司、中鋼公司與台灣電力公司。各參與單位依據其日常業務執行經驗，提供具治理代表性之應用場景，作為評測資料來源。後續由高雄市政府資訊處統

籌統整並轉換為結構化測試題項，確保整體資料體系具備一致性與可操作性。

整體概述如下：

- 單位參與總數：8 個單位（5 個局處和 3 個國公營單位）
- 核心事件情境數：108 個具代表性情境
- 事件情境題目數：609 題問題內容
- 涵蓋問題類型：二元分類、類別辨識、有序/層級類別等題型
- 問題設計原則：以場景驅動為核心，融合治理邏輯導向，並強調多維度能力評估的多樣性

上述設計流程不僅重視資料真實性與實務關聯性，亦兼顧題型多樣性與語意層次，為模型能力之系統性檢測奠定堅實基礎。

3.2 問題設計類型與結構

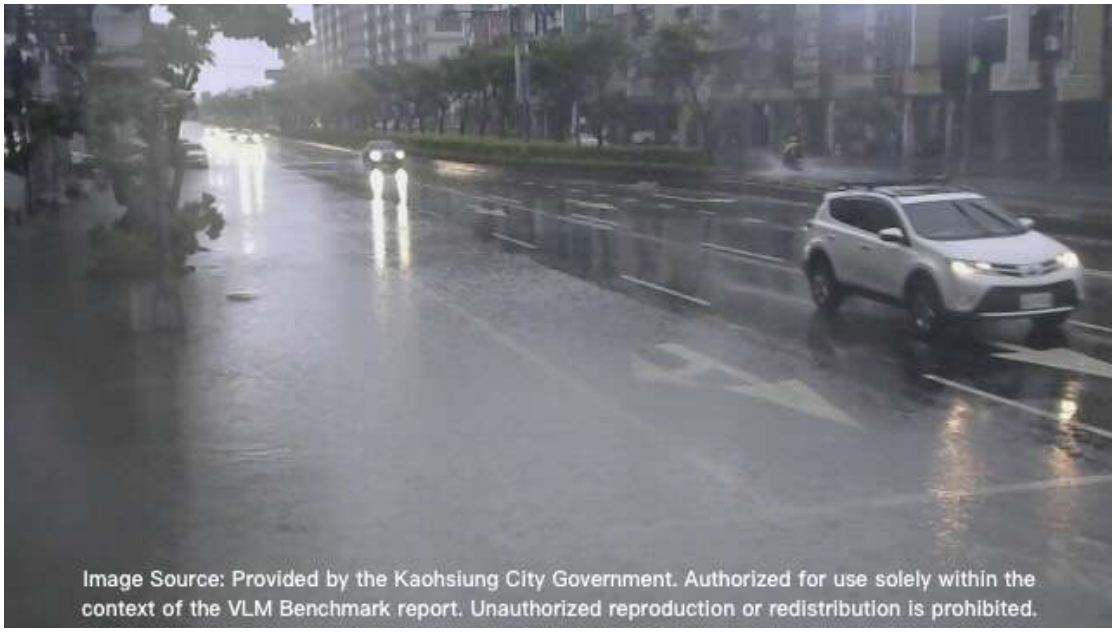
為全面且系統性地評估模型於城市治理任務中所需的多面向能力，本評測基準將所有題型劃分為三大主要類別，分別對應不同層級之推論與判讀需求：

- 二元分類題（Binary Classification）：針對特定事件是否發生進行判定，例如判斷畫面中是否存在違停、阻塞或其他不當使用情形。
- 類別辨識題（Category Recognition）：要求模型辨別事件或物件的具體類型，例如判定交通事故為碰撞、翻車或擦撞等分類。
- 等級排序題（Ordinal/Ranked Scenarios）：考驗模型對情境嚴重程度進行排序或等級判斷的能力，例如將交通壅塞情形依 A 至 F 六級等級評估。

此三類題型分別對應治理任務中所需之三種基本認知與決策技能，涵蓋事件察覺、分類判斷與情境推估等能力，並在最終評分系統中依據題型性質設定不同之權重與正確率門檻，以強化測評結果之準確性與可比較性。

此外，以下所展示的測試範例，選自由水利局、運發局及工務局所提供的實景監視畫面資料，藉此說明所有治理場景均可清楚對應轉化為上述三類題型，並具備可標準化處理之潛力，進一步確保評測內容的代表性與一致性。

水利局



1. 是否有積淹水：是 / 否
2. 受積淹水影響路段：無影響 / 單一(單向)或部分車道 / 整段(雙向來回)道路或路口
3. 積淹水深度：淺（< 30 cm）/ 中等（30~50 cm）/ 深（> 50 cm）
4. 積淹水道路封閉狀況：封閉 / 未封閉
5. 是否有車輛在積淹水區域：是 / 否
6. 積淹水狀況下車輛是否能通行：是 / 否
7. 是否有行人在積淹水區域：是 / 否
8. 積淹水狀況下行人是否能通行：是 / 否
9. 積淹水區域是否有警示標誌：是 / 否
10. 排水孔是否阻塞：是 / 否

運發局

1. 是否有緊急出口 ： 是 / 否

2. 緊急出口是否阻塞：是 / 否

3. 造成緊急出口阻塞的類型：人群 / 雜物 / 垃圾 / 其他 / 兩種以上

4. 緊急出口阻塞程度為何：部分阻塞/ 完全阻塞

5. 緊急出口的阻塞物是否可移動：是 / 否

工務局

1. 是否有佔用車道的道路施工：是 / 否
2. 施工類型：道路修補作業 / 道路刨鋪作業 / 人行道工程 / 道路挖掘作業 / 高空作業車工程(路燈維修、路樹修剪) / 道路拓寬 / 建築工地占用道路施工(施工機具及物料) / 其他
3. 是否有施工告示牌或施工指示物：是 / 否
4. 道路施工區域交維措施是否完備(交通錐加連桿、型鋼護欄、紐澤西護欄等)：是 / 否
5. 臨道路施工區域是否有架設圍籬：是/否
6. 道路施工區域交維措施是否具備夜間警示：是/否
7. 是否有施工人員在場：是 / 否
8. 畫面時間段：日間 / 夜晚 / 其他

3.3 測試資料量與設計目標

根據參考三組問題類型同為手動設計之選擇題，且常被知名 VLM（VILA、LLaVA）作為測試比較之 Benchmark：ScienceQA、MME 和 MMBench 。其測試資料集數量分別為 4,241、2,374 和 1,784。而目前統計所有問題的選項總數約為 1,400，故若單一選項平均挑選 4 張影像，使其具有足夠多樣性的 5,600 筆資料，在與同類型之 Benchmark 相比之下，其測試結果應已具有一定的代表性。

若根據測試資料為 20% 佔比的原則，在單一選項之測試資料為 4 張的情況下，總資料量需達 20 張，但在觀察目前的資料收集情況後，資料提供單位目前許多選項的可用影像數並未達此標準，故目前先以收集 4 張影像為目標。若後續在罕見選項的資料收集上更為順利，則可達到原計畫書設定的 20,000 張目標（單一選項平均 15 張影像）。

簡而言之，在總體資料架構下，每個題目裡的每個選項，必須配有 4 張圖像，所有所選擇之測試圖像皆經由人工審核確認，從而將建置約 5,600 張高品質測試圖像作為測試資料集基礎版本，並嚴格將這些資料排除於訓練資料之外，確保該資料集是未受污染或未被揭露過的資料。

設計目標如下：

- 基本覆蓋規模：每一答案選項對應 4 張圖片（作為最低標準）
- 擴充性目標：每一選項最多對應至 15 張圖片（作為理想設計上限）
- 初始資料集規模：預估約 5,600 組圖文題項組合
- 最大擴展上限：若圖像取得能力提升，整體資料筆數可望拓展至 20,000 筆以上

後續在使用該資料集時，為保障評測的公正性並避免模型預先接觸測試內容（即防止資料洩漏），需確保測試圖像完全未曾出現在受評估模型之訓練資料集中。此程序為提升模型測試可信度之關鍵步驟。

3.4 資料挑選原則與場景規劃依據

為確保測試資料具備高度的情境真實性與環境多樣性，本評測基準採行一套結構化之三層級資料選取策略，具體規劃如下：

- 場景類型區分：所有資料首先依據場域性質進行分類。對於室內場景，因其不受天候與時段變化之顯著影響，故不強制標註時間與氣候資訊；而戶外場景則須完整標註包含時間（如白天、

夜晚）與天氣狀態（如晴、陰、雨）等環境條件，以強化其在實際應用場域中的辨識與推理挑戰度。

- 依時間維度進行分布規劃：圖像資料依據實際取得過程中之比例進行分布設計，以模擬真實城市監控系統常見的時段條件。具體分布為：日間圖像占比 75%、黃昏圖像 2%、夜間圖像則占 23%。
- 依氣候變因進行分布設計：資料之天氣條件則參考高雄市近十年歷史氣象資料統計比例進行對應調配，具體為：晴天 52%、陰天 20%、雨天 28%。此分布模式亦有助於模擬模型於不同可視條件與環境狀態下的表現穩定性。

透過此三層式場景規劃策略，不僅可確保題材情境之真實性與在地性，更有助於構建一套能涵蓋多種操作條件變數的全面性測評系統，進而強化模型於不同環境條件下的適應力與推論穩健性之評估效度。

3.5 城市治理領域分類與適用性

在實際資料蒐集與題目設計過程中，以局處劃分治理任務，雖有行政層級上的管理便利性，但在模型設計與任務分類上，容易產生權責交疊與任務模糊等問題。為提升資料系統性與未來跨域擴充的靈活

性，本評測基準特別將資料來源與場景任務，依據治理「領域」進行再分類，使之超越局處分工邊界，聚焦於任務本質及影像場景特徵。

領域分類對應實務現場場景與影像特徵，各領域即代表特定類型的城市監控影像或事件樣態，有助於題型設計更貼合視覺理解需求，且橫跨多局處可共享之任務型邏輯，例如「道路狀況」可能同時涉及工務局與水利局，而「災害應對」亦需整合運發局與交通局資料。具體而言，我們將治理場景劃分為以下城市應用領域，涵蓋交通設施、基礎建設、水資源管理、捷運運營、港區與工業場域等跨部門應用情境：

Domain	領域
Traffic Management	交通管理
Facility Anomaly Detection	設施異常
Disaster Response	災害應對
Station / Shelter / Parking Facility Management	車站/候車亭/停車場
Venue Management	場館管理
Metro Platform Monitoring	捷運車站月台
Metro Carriage Monitoring	捷運車廂
Light Rail Monitoring	輕軌偵測
Water Resource Management	水利管理
Drainage / Retention Basin Maintenance	排水渠道/滯洪池管理
Road Condition Monitoring	道路狀況
Construction Safety Monitoring	施工安全
Port Area Management	港區管理
Industrial Facility Management	廠區管理
Power Infrastructure Management	電力設施管理

3.6 隱私保護與去識別化處理原則

本評測基準所使用之影像資料，部分來源涉及實際街景、公共場域監視器等資料，為符合《個人資料保護法》第 6 條所規範之特種個資處理規定，並同步參照歐盟《GDPR》相關標準，所有影像資料在對外使用及介面展示階段，皆須完成人臉與車牌等資訊之去識別化處理。

本基準之去識別化處理程序，將依據各參與單位提供之原始資料進行逐筆確認與遮蔽，處理方式包含模糊化（blurring）、遮罩覆蓋（masking）等，並同步建立資料版控記錄，以確保隱私保護機制之透明性與可追溯性。透過上述設計，確保基準資料兼具實務效能與隱私合規，達成城市治理與倫理責任之平衡。

---

4. 評估設計與分數模型

4.1 評測流程架構

為真實模擬模型於城市治理現場部署之使用情境，本評測基準之評估流程導入重複測試週期（Repeated Evaluation Cycles）與語意標準化機制（Semantic Normalization），使得模型預測表現可在不同輸入條件下進行穩定性與準確性的多層次檢驗。

核心流程設計包含下列要素：

- 單一題目多圖輸入：每一測試題目對應多張視覺刺激影像，藉此觀察模型在變動視覺條件下的答案穩定性與一致性。
- 答案語意標準化：模型產生之開放式回答將對應至預先定義的標準答案集合中，以利計分與比較。由於題型涵蓋「二元分類」、「類別辨識」與「等級排序」三大類別，各自對應不同之評分機制與標準化邏輯（後續小節將分別詳述）。
- 綜合評分指標紀錄：評估結果將綜合紀錄以下指標：準確率（Accuracy）、召回率（Recall），並據此形成完整評估分析報告。

此一設計可提升評測過程之可信度與透明性，並更真實反映模型部署於實際治理場景下的可用性與穩健性。

4.2 題型設計與結構及得分邏輯

為了全面性檢視智慧城市治理脈絡下 VLM 模型的能力，本 Benchmark 針對所有問題進行了嚴謹的類型劃分，並搭配明確的得分邏輯，確保評測結果能真實反映模型的多層次能力。

題型劃分以及得分邏輯如下：

- 二元分類（Binary Classification）

- 測試模型對基本事實或規範的辨識能力，判斷事件是否發生，如：有無違規行為。
- 得分邏輯：答對給分，答錯不得分，適用簡明題型。

- 類別辨識（Category Recognition）
  - 檢驗模型對類別、物件、設施辨識的精準度，並能選擇正確的類別，如：事故類型：碰撞 / 翻覆 / 擦撞。
  - 得分邏輯：正確即得滿分，錯誤不得分，若答案接近可視情境給予部分分數。
- 有序／層級類別（Ordinal / Ranking Tasks）
  - 檢測模型對事件嚴重性、風險程度、緊急性的區分能力，處理具等級概念的問題，如：交通阻塞影響：A-F。
  - 得分邏輯：完全正確得滿分，若排序或層級誤差輕微，可依據比例給予應得之分數。

所有答案皆需對應真實治理任務邏輯，由領域專家進行雙重審核，以確保答案正確性與合理性，並防止語義模糊導致的偏誤。

4.3 目標分數門檻彈性配置邏輯

為更真實對應城市治理場景中不同題目的難易差異，本評測系統之計分門檻設有彈性調整機制，主要依據下列三項條件進行判定：

- 資料取得稀缺性：對於資料來源極為有限之情境題型，可適度放寬正確率門檻，以反映資料稀有特性。
- 推論邏輯複雜度：若題型需多階段推論或含抽象概念，則適用較低的嚴格度進行評核。
- 場景條件之罕見性或模糊性：若測試情境為少見或邊界不明確（如極端天候、夜間特殊情境），則門檻亦可調整，以鼓勵模型逐步克服難題。

透過此彈性調整設計，可針對不常發生問題或缺乏資料之情境進行適當調整，可被認定為風險及發生機率不高之情境。另外也能引導模型針對弱勢情境持續優化，亦有助於研究者辨識模型真實能力邊界與風險因子。

---

5. 初步驗證與階段成果

本評測基準已逐步完成多項初期評估任務，涵蓋智慧城市治理中數個核心場景，包括交通事故判讀、道路違規事件辨識、環境異常監測，以及社區公共空間之影像監控應用等。這些測試題項之設計均對應實際治理需求，具高度情境代表性。

在實際執行過程中，受評估之 VLM 系統所產生之回答內容，已有效驗證本評測架構在辨識模型常見錯誤與行為偏誤（behavioral biases）方面具高度效力。特別是針對訓練資料不足或情境複雜度較高之題型，模型表現波動顯著，進一步彰顯評測基準的鑑別能力與應用價值。具體觀察成果包括在夜間與雨天條件下的辨識能力明顯下降，特別是針對遠距模糊影像之解析度，模型準確率顯著低落；於分類任務中出現語意相近選項的混淆現象，如交通事故類別間(碰撞、翻車、擦撞等）之誤判；模型傾向偏誤回應過度訓練語料中之類型，顯示其泛化能力於不熟悉場景中仍有待提升。

上述問題的揭露，對於後續模型優化策略之制定具關鍵參考價值，並能有效指導未來之模型微調（fine-tuning）與資料集擴充策略（dataset augmentation），確保後續系統開發具備更高的實用性與穩定性。

---

6. 結論與後續方向

綜合而言，本專案所建構之「智慧城市視覺語言模型（VLM）評測基準」，其角色與定位已超越傳統學術導向之能力評估資料集。該基準現已轉化為一套具政策導向、多單位協作、本地自主、實務驗證與

全球擴展潛力之 AI 評估平台，展現高度應用價值與國際對接可能性。

在本計畫推動過程中，高雄市政府所屬治理單位與 AI 技術開發商、學術研究團隊密切合作，實現一套閉環式應用模式。此模式強調從實務需求出發進行資料與題項設計，並將模型之回應結果作為政策決策流程之回饋機制，有效建立數據—模型—治理之互動循環體系。

展望未來，我們將持續優化並在地化高雄版本之 VLM Benchmark，進一步支援資料集標準化與評估公平性提升工作。此外，將積極推動跨縣市間之合作與共同開發機制，期望藉由多方協力擴大應用場景與資料樣態，朝向國際輸出本標準之目標邁進，使其逐步成為全球智慧城市治理領域中 AI 能力驗證之重要評測依據。