DOI: 10.6653/MoCICHE.202502_52(1).0007

營建工地影像生成文字摘要系統之開發與應用：以工地安全缺失摘要生成為例

蔡明修* ／淡江大學土木工程學系 助理教授
張傳育／國立雲林科技大學資訊工程學系 特聘教授
許輝煌／淡江大學資訊工程學系 教授
盧祥偉／台灣世曦工程顧問股份有限公司BIM中心 副理
陳炳宏／台灣世曦工程顧問股份有限公司營建部 計畫工程師

本研究開發了一套基於生成式人工智慧的「營建工地影像生成文字摘要系統」，透過整合多模態模型、大語言模型和圖像檢索 RAG 技術，實現工地安全影像的智能分析與管理。研究團隊建立了包含 1,373 筆的工安缺失影像資料集，並開發出能自動分析工地照片並生成專業摘要的 AI 引擎，該引擎可產生包含場景描述、工安缺失、造成原因以及違反法規等完整資訊。系統採用台灣本土優化的 Llama-3-Taiwan 作為大語言模型，結合圖像檢索 RAG 技術提升專業知識的準確性。實測結果顯示，系統在場景描述正確性及缺失內容正確性方面表現優異，特別是在法規檢索方面較其他 AI 模型展現出顯著優勢。研究成果不僅提供了便捷的工地影像管理與分析工具，更為工程知識的累積與傳承提供創新解決方案。透過 API 的佈署，本系統可與既有的工地管理系統整合，為推動智慧工地的發展奠定重要基礎。

關鍵詞：多模態模型、大語言模型、RAG、工地安全、法規、工地影像

前言

隨著人工智慧（AI）技術的快速發展，各行各業都在積極探索其應用潛力，營建工程也不例外。特別是在工地安全管理方面，AI 技術的應用變得越來越重要。根據研究，全球營建業每九分鐘就發生一起死亡事故 [1]，顯示出傳統安全管理方法的侷限性。尤其是在工地現場管理方面，影像資料的處理和分析扮演著至關重要的角色。傳統上，工地影像的分析主要依靠人工方式，但這種方式耗時費力，且容易受到主觀因素影響。因此，發展自動化的工地影像分析技術已成為提升工地管理效率和安全性的關鍵。

近年來，深度學習技術的突破為工地影像分析帶來了新的契機。深度學習模型，特別是多模態模型，具備同時處理影像和文本資料的能力，能夠從工地影像中提取關鍵資訊，並生成易於理解的文字摘要。結合大型語言模型（LLM）的強大推理能力和檢索增強生成（RAG）技術的知識整合能力，可以進一步提升工地影像分析的準確性和專業性。最新研究表明，在

* 通訊作者，mht@gms.tku.edu.tw

營建安全管理領域，RAG 技術可提升 21.5% 的效能，而 fine-tuned LLM 則可提升 26% 的效能 [2]，這證實了 AI 技術在提升安全管理效率方面的巨大潛力。

本研究旨在開發一套基於多模態生成式人工智慧的「營建工地影像生成文字摘要系統」，重點關注工地安全缺失的辨識與描述，以提升工地安全管理的效率與準確性。近期研究指出，整合文字、視覺和聲音等多模態數據的 AI 系統，是未來營建安全管理的重要發展方向 [1]。本系統的研究成果與此趨勢相互呼應：從最容易取得的照片影像出發，整合了多模態模型、中文大型語言模型和專業知識檢索技術，能夠自動分析工地影像，生成包含場景描述、工安缺失、造成原因和違反法規等資訊的專業摘要，並提供便捷的影像管理、查詢和評估功能。本系統的開發將有助於提升工地安全管理效率，促進工程知識與經驗的有效學習傳承，為營建產業的 AI 輔助安全管理提供創新解決方案。

問題與對策

問題定義

將生成式人工智慧技術應用於工地影像分析，特別是工地安全缺失摘要的生成，仍面臨著許多挑戰：

1. **工地場景多樣性**：工地現場環境複雜多變，不同類型的設備、多樣的工人活動和施工流程，以及天氣、光線、拍攝角度等外部因素，都為影像分析帶來了挑戰。如何讓模型適應這種多樣性並準確識別安全隱患是首要問題。
2. **文字描述的專業性與嚴謹性**：生成的文字描述必須符合工地安全專業需求，正確描述安全缺失，並對應相關法規。如何確保文字描述的專業性、正確性、一致性，以及避免語言模糊、重複或不符合專業語境等問題是另一項挑戰。
3. **影像中細節與背景資訊的辨識能力**：工地影像中經常包含大量細節和模糊的背景資訊，模型需要具備精確識別細節和篩除干擾資訊的能力，才能避免錯誤歸類或無法準確識別安全缺失。
4. **法規的精準性**：工地安全相關法規繁雜且多層次，不同地區和工程類型可能適用不同的規範 [2]。模型需要準確匹配具體法規條文，避免引用不當導致錯誤解讀或法律責任問題。
5. **多目標優化的矛盾**：在保證摘要專業性、準確性的同時，還需考慮生成效率和模型佈署資源等限制。如何在專業性、正確性與效率之間取得平衡，並確保系統在實際工地環境中的可用性是另一個重要挑戰。

研究方法

為解決營建工地影像分析中的關鍵技術挑戰，本研究提出了一種基於多模態模型、大語言模型和檢索增強生成（RAG）技術的創新解決方案。我們的方法旨在整合視覺理解、語言推理和知識檢索，以實現高效且正確的工地影像安全摘要生成。

1. **多模態模型**：負責視覺的處理，主要功能是識別影像中多樣且複雜的工地場景、設備和人員活動，並生成詳細的場景描述 [3,4]。在本研究中，我們採用測試後符合實務需求之圖文多模態模型作為生成引擎之「眼睛」，以處理工地圖像和文本的轉換與整合 [5]。
2. **大語言模型**：作為系統的「推理大腦」，負責「推理」與「判斷」的任務，主要功能是根據多模態模型生成的場景描述，以及圖像檢索 RAG 提供的工安缺失和違反法規資訊，進行綜合邏輯推理，生成工安缺失、造成原因和違反法規等文字內容。本研究採用 Llama-3-Taiwan 模型 作為大語言模型，因其在繁體中文的理解和生成能力上表現出色 [6,7]。
3. **圖像檢索 RAG**：用於知識擴展，負責提供「專業知識」的補充。但不同於傳統的文字 RAG，本研究所建立之圖像檢索 RAG 可根據輸入影像的特徵，從知識庫中快速檢索出相似的影像，並提供相應的工安缺失和違反法規資訊，作為大語言模型推理的依據。本研究使用深度學習模型進行特徵提取，並建立包含工安缺失影像、場景描述、安全缺失說明及違反法規條款等內容的知識庫 [8-10]。

「工地安全摘要生成智慧引擎」開發與評估

為了解決工地影像分析所面臨的挑戰，本研究結合多模態模型的影像解析能力、大語言模型的語言推理能力，以及 RAG 的知識擴展能力開發了「工地安全摘要生成引擎」。此方法的優點是能夠綜合利用各模型的優勢，生成更準確和專業的摘要。以下說明工地安全摘要生成引擎之開發內容。

資料收集與處理

本研究的資料收集主要包含兩個面向：工安缺失

資料和工安法規資料。工安缺失資料由工程師提供的一萬張工地影像和對其中數千張工安缺失影像標註的文字描述組成，經過 AI 工程師進行資料清洗，包括錯字修正、冗字移除和同義詞統一等處理。工安法規資料則由人工從國家法規、技術標準和行業規範中篩選整理，建立結構化的資料庫。

多模態模型與大語言模型選擇

在模型選擇方面，本研究測試了多種多模態模型的工安缺失生成效能。評估採用人工評分方式，若生成的缺失描述與圖像相符則得 1 分，透過大量測試選出最適合且資源需求合理的模型。為確保系統具備專業性和在地化能力，本研究選用 Llama-3-Taiwan 作為大語言模型，其在繁體中文的理解和生成能力上表現優異，並經過台灣在地資料的優化，能更準確地理解和表達符合本土工程專業用語，這對於工地安全管理的專業性和實用性極為重要。

圖像檢索 RAG 技術

圖像檢索增強生成（Retrieval-Augmented Generation, RAG）技術是本研究中提升工地安全影像分析准確性的關鍵所在。該技術通過將深度學習模型與專業知識庫相結合，能夠顯著增強模型對特定領域的理解和生成能力。在本引擎中，我們結合了深度學習、傳統影像特徵其取技術及圖像文字知識庫，來增強多模態模型對工地安全狀況的理解和評估能力。實現步驟如圖 1。

系統的運作可分為兩個階段。在預處理階段，系統會將收集到的工地影像進行標準化處理，並建立圖像檢索知識庫。在實際使用階段，當新的工地照片輸入後，系統會在知識庫中搜尋最相似的歷史案例，並提取這些案例對應的安全缺陷描述和違反法規資訊。這些資訊會輸入到多模態模型中進行分析，最終生成包含工地安全缺陷、造成原因及違反法規等內容的專業評估報告，協助工地管理人員更好地了解和處理現場的安全隱患。圖 2 為本研究透過深度學習模型所得到之相似度圖像索引結果，此結果符合本研究之要求，可供生成引擎整合之用。

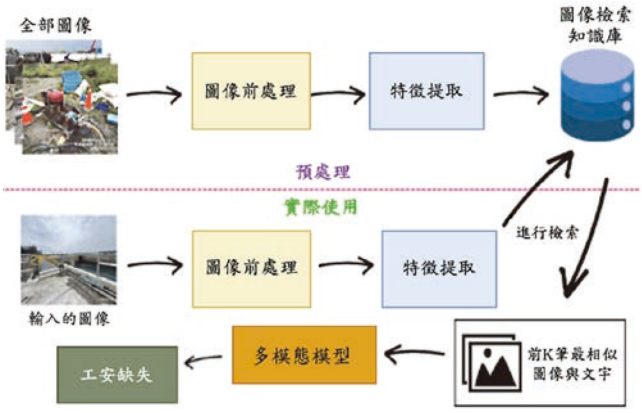

圖 1　圖像檢索 RAG 技術

工地安全摘要生成引擎整合架構

整合上述的工地安全缺失資料、多模態模型、大型語言模型和圖像檢索增強生成（RAG）等關鍵技術，本研究開發了一套創新的工地安全摘要生成引擎演算機制（圖 3），此機制可動態調整生成策略，確保輸出的摘要具有高精確度與實用性，適用於工地安全監控與決策支持。

當引擎接收上傳帶有工地安全缺失的照片後，將分為 3 步驟生成該照片的安全缺失摘要及違反法規內容。此機制可動態調整生成策略，確保輸出的摘要具有高精確度與實用性，適用於工地安全監控與決策支持。

表 1 呈現了二張不同測試案例工地場景照片的生成結果。其中，表 1 左圖可發現生成引擎能甚至能發現

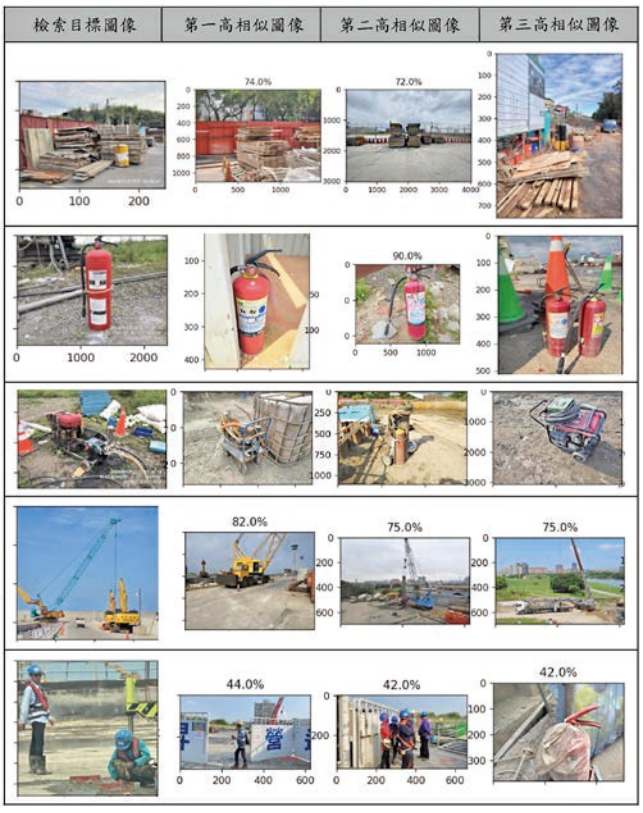

圖 2　圖像 RAG 相似度檢索結果範例

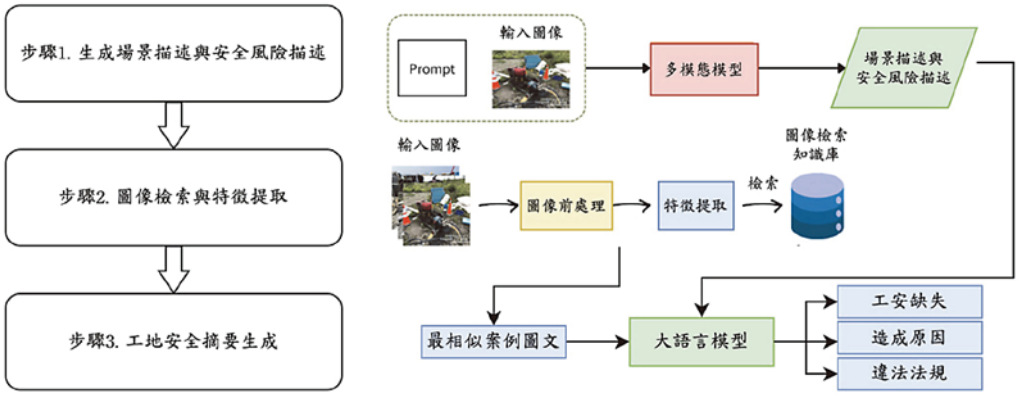

圖 3　整合多模態模型、大語言模型及圖像檢索 RAG 技術之工地安全摘要生成引擎演算架構

表 1　生成引擎生成內容範例

欄位名稱	測試案例 1	測試案例 2
輸入圖像		
場景描述與安全風險描述（AI 生成內容）	1. 場景描述：這張影像顯示了一個室外的工地場景。在影像的左側，可以看到一名身穿白色襯衫、藍色安全帽和紅色反光背心的工人，他正在觀察周圍的環境。工人身旁有一瓶水。在影像的右側，另一名工人身穿藍色上衣、藍色安全帽和紅色反光背心，正蹲在地上，似乎在進行某些工作。背景中可以看到一些建築結構和起重機。 2. 安全風險描述：在影像中，可以看到兩名工人都佩戴了安全帽和反光背心，這表明他們採取了基本的個人防護措施。然而，其中一名工人沒有穿著合適的工作靴，這可能會增加滑倒或踩到尖銳物體的風險。此外，地面上似乎有一些雜物或碎屑，這可能會導致工人滑倒或被刮傷。工人的活動區域也沒有明顯的標記或屏障，這可能會增加意外碰撞的風險。	1. 場景描述：這張工地影像顯示了一個港口或碼頭的場景。一輛黃色的 KOMATSU 挖掘機正吊起一個輪胎，輪胎被一條黃色的繩索吊起，並且已經接近地面。在挖掘機的右側，可以看到一艘船的一部分，船上堆放著一些物品，包括繩索和一個大型的金屬結構。在挖掘機和船的周圍，有幾名工人，他們都穿著安全帽和反光背心，有些人站在挖掘機旁邊，有些人則站在船旁。地面上還有一些其他的裝置和物品，包括一個黑色桶和一些雜物。 2. 安全風險描述： 影像中顯示的工地存在以下安全風險： (1) 工人未穿著完整的個人防護裝備，例如手套、耳罩等，僅穿著安全帽和反光背心。 (2) 現場有大量雜物和裝置，可能導致工人跌倒或被物品擊中。 (3) 吊起輪胎的過程中，如果繩索斷裂或吊裝過程中出現失誤，可能會對工人造成傷害。 (4) 工人站的位置和行動可能會影響挖掘機的操作，存在被挖掘機吊臂或其他部件撞擊的風險。
工安缺失、造成原因與違反法規（AI 生成內容）	● 工安缺失：作業人員安全帽帽扣未扣緊。 ● 造成原因：影像中左側的工人並未佩戴安全帽，違反了提供適當安全帽並使其正確戴用的規定。 ● 違反法規：營造安全衛生設施標準第 11-1 條 1 僱主對於進入營繕工程工作場所作業人員，應提供適當安全帽，並使其正確戴用。	● 工安缺失：使用挖土機吊掛物件 ● 造成原因：挖土機未經改裝或加裝吊掛裝置，直接使用挖斗作為吊掛工具，存在吊掛物掉落或挖斗損壞的風險，可能導致工人被掉落物砸傷或物件掉落造成的二次傷害。 ● 違反法規：職業安全衛生設施規則第 116 條九、不得使車輛機械供為主要用途以外之用途。

左邊工人的安全帽未正確配戴。故從生成內容來看，無論是場景描述、安全風險評估，還是工安缺失判斷與法規引用，系統都能正確識別人員、機具在工地場景中的狀況，並能正確引用安全衛生法規。此一致性證實了生成引擎具有一定良好的穩定性，能在不同場景下維持可靠的分析水準。這種特性不僅確保了系統判斷的一致性和可信度，對工地安全管理的實務應用也極具價值。更重要的是隨著知識庫資料的持續擴充和更強大模型的導入，引擎對工地安全缺失判斷與陳述的正確性還能進一步的提升。

工地缺失摘要生成引擎成效評估

為了驗證工地缺失摘要生成引擎的效能與實用性，我們設計了一套專家評分機制，針對生成結果的多面向表現進行嚴謹的評估。題庫設計、評分構面設定與評分結果統計的具體說明如下：

1. 題庫設計：評估的基礎資料來自訓練資料集以外的「台北市勞動檢查處職災案例」，透過從中挑選 26 個實際案例作為測試考題。這些案例涵蓋多種常見的工地安全缺失類型，包括施工架安全（34.6%）、開口墜落（30.8%）、物料堆放（19.2%）與個人防護具（15.4%）等，確保評估具有多樣性與代表性（圖 4）。

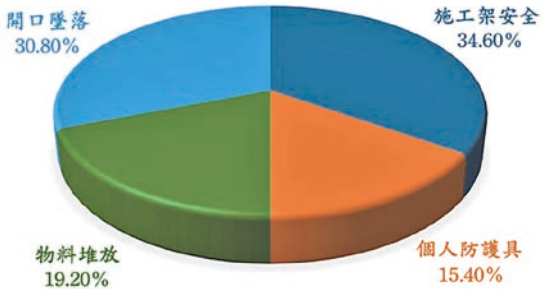

圖 4　工安缺失考題分類及比例

2. 專家評分構面：針對工安缺失摘要的生成需求，擬定五個評分構面（表 2）。專家依據此五大構面，針對每個影像的生成結果進行 1 至 5 分的評分。此評估方法不僅著眼於生成摘要的技術正確性，還考量實際應用於工地安全教育與管理的價值，從而為引擎的持續優化提供依據。
3. 評分結果分析：針對工地缺失摘要生成引擎的生成內容進行綜合評估，分析多項指標與不同缺失類型的生成表現，並比較模型與 ChatGPT-4o 的比較，提供以下深入觀察與結論。

- 場景描述正確性（表 2）：在場景描述正確性構面上，生成引擎的表現尤為突出，獲得平均 4.3 分，能準確還原影像中的人物、物件及場景細節，並避免產生錯誤或虛構資訊。
- 缺失內容正確性（表 2）：引擎在缺失內容正確性構面評分為 4.1，能有效反應影像中實際存在的安全問題，提供準確的缺失描述，對工地安全管理的應用具有高度參考價值。
- 法規適合性的專業性（圖 5）：法規適合性雖然是本生成引擎模型中評分最低的一項（3.5），但相較於僅使用 ChatGPT-4o 的 1.2 分，仍展現出顯著優勢。這表明本生成引擎在結合法規條文與本地安全標準時具有更高的準確性與專業性，有助於在實際應用中提供更可靠的安全建議。然未來仍可增加法規知識庫之案例，能進一步提昇正確性。
- 潛在缺失辨識能力不足（表 2）：在潛在缺失辨識構面，系統的表現仍有待提升，評分僅為 3.8。這反應了模型在捕捉影像中隱性安全風險的能力有限，未能充分發揮對隱性問題的檢測作用。

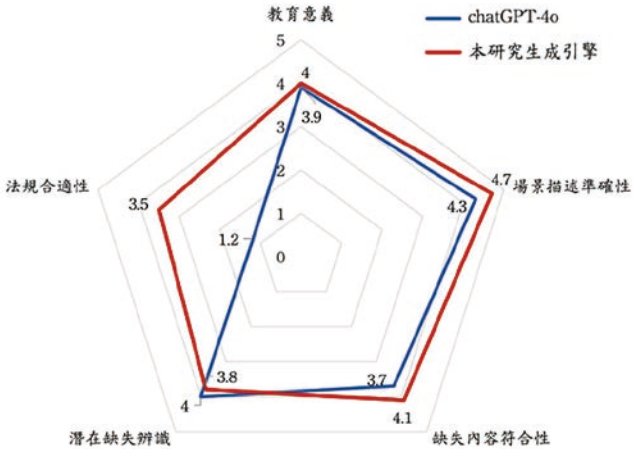

圖 5　本研究生成引擎（紅）與 chatGPT-4o（藍）生成結果評分比較

表 2　工安摘要專家構面說明及各構面評分分數

項目名稱	說明	平均分	標準差	最低分	最高分
場景描述準確性	除了應陳述的場景內容外，AI 的描述人事物正確，並沒有產生幻覺？	4.3	1.06	2.0	5.0
教育意義	此 AI 缺失描述有助於我對工地安全缺失的認識？	4.0	1.43	1.0	5.0
法規合適性	所引用法規與條款正確。	3.5	1.48	1.0	5.0
潛在缺失辨識	所描述的缺失內容包含了該照片可能潛在的工地缺失？	3.8	1.22	1.0	5.0
缺失內容正確性	所描述的缺失內容符合該照片所呈現的缺失？	4.1	1.32	1.0	5.0

「營建工地影像生成文字摘要系統」開發與實作

本研究基於上述工地安全缺失摘要生成引擎，進一步開發「營建工地影像生成文字摘要系統」，提供使用者上傳工安缺失照片，並自動生成摘要內容以方便查閱與教育訓練。使用者可以批次上傳工地照片（圖6），系統會自動進行分析並生成摘要（圖7）。系統內同時提供了專業評分功能，讓專家可以對AI生成的摘要進行評估，不僅有助於評估系統效能，也為後續模型優化提供了重要參考。此外，系統還整合了多關鍵字檢索功能（圖8），方便使用者快速查找所需的工地影像資料。

圖6　上傳照片主頁面及功能說明（上傳後自動調用生成引擎生成工安缺失摘要）

圖7　顯示照片自動生成摘要資訊（查看AI生成摘要）

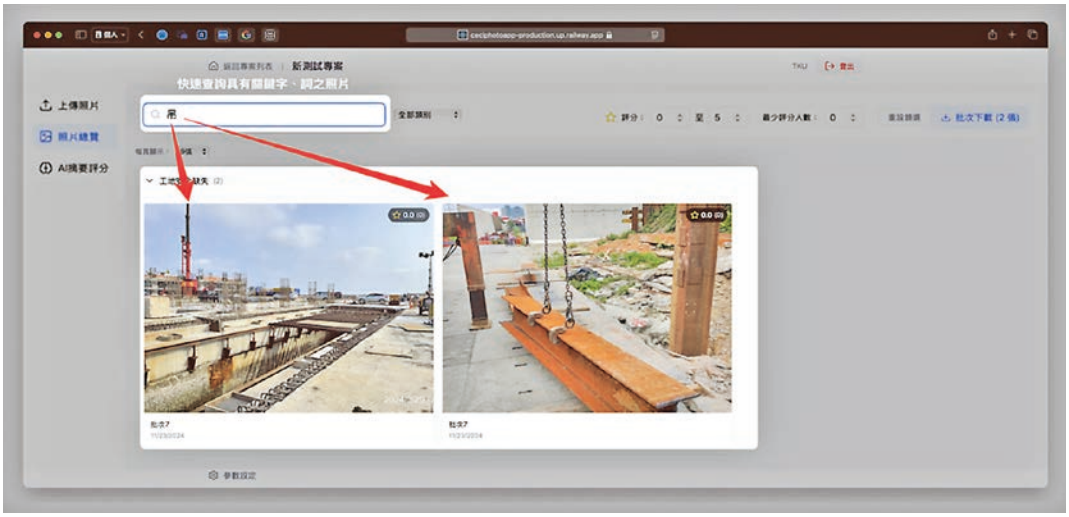

圖 8　上傳照片後不用另外輸入即可以在專案照片瀏覽頁面以關鍵字查詢圖片內容（ex. 輸入「吊」字查詢出剛上傳之吊車及吊掛照片）

結語

本研究成功創新整合多模態模型、大語言模型及圖像檢索 RAG 技術，完成「營建工地影像生成文字摘要系統」在工地安全上的開發與應用，為工地安全智慧化管理的願景向前邁進一步。重要結論如下：

1. 「工地缺失摘要生成引擎」驗證了生成式 AI 技術應用於工地影像分析的可行性。透過整合多模態模型的視覺理解、大語言模型的語言推理及專業知識檢索等技術，生成引擎高比率且正確地分析工地影像並生成專業的安全缺失摘要。而在工安題庫的測驗中，生成引擎在場景描述正確性及缺失內容正確性方面均獲得優異評分，更展現了對法規正確檢索的能力。
2. 「營建工地影像生成文字摘要系統」的開發為工程實務帶來潛在的顯著效益。系統不僅提供便捷的影像管理與摘要生成功能，更透過智慧化的分析，快速方便提供工程師明顯或潛在的安全缺失及潛在安全風險資訊。此外，系統的知識累積與傳承功能，為營建產業的永續發展提供了重要想像可能。
3. 本研究為 AI 微服務之應用建立可實現的技術基礎。本系統透過 API（Application Programming Interface）的佈署，能以微服務的形式與既有專案管理資訊系統、機械狗、空拍機等各類智慧工地影像設備整合，實現工地安全監控的自動化，為建立完整的智慧工地生態系統奠定重要基礎。

未來研究將持續擴充知識庫、強化多模態模型的視覺理解能力、優化圖像檢索準確度。同時，我們將深化與其他智慧工地系統的整合，探索更多元的應用場景，如工程進度追蹤、品質管理等領域。透過這些優化與擴展，本系統將在提升工地安全管理水準、促進工程知識傳承、推動營建產業數位轉型等方面發揮更大的價值。

參考文獻

1. A.B.K. Rabbi and I. Jeelani (2024). “AI integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications,” Autom. Constr., Vol. 164, pp. 105443.
2. J. Lee, S. Ahn, D. Kim, and D. Kim (2024). “Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval,” Autom. Constr., Vol. 168, pp. 105846.
3. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, et al. (2021). “An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale,” arXiv preprint arXiv:2010.11929.
4. D. Lowe (2004). “Distinctive image features from scale-invariant keypoints,” International journal of computer vision, Vol. 60, No. 2, pp. 91-110.
5. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., and Lin, J. (2024). Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. arXiv preprint arXiv:2409.12191.
6. P.-H. Chen, S. Cheng, W.-L. Chen, Y.-T. Lin, and Y.-N. Chen, “Measuring Taiwanese Mandarin Language Understanding,” arXiv preprint arXiv:2403.20180, 2024..
7. Y.-T. Lin, “Llama-3-Taiwan-70B-Instruct,” Hugging Face, [線上]. Available: https://huggingface.co/yentinglin/Llama-3-Taiwan-70B-Instruct. [存取日期：2025 年 2 月 10 日].
8. Lewis, P., Perez, E., Petroni, F., Karpukhin, V., Küttler, H., Lewis, M., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.
9. Soudani, H., Kanoulas, E., and Hasibi, F. (2024). Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge. arXiv preprint arXiv:2403.01432.
10. Balaguer, A., Benara, V., de Freitas Cunha, R.L., Estevão Filho, R.M., Hendry, T., Holstein, D., and Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv preprint arXiv:2401.08406.