Graph Retrieval Augmented Generation for Multimodal Construction and Demolition Waste Data Integration Toward Traceable Disposal Recommendations

Yu Gao[1], Tak Wing Yiu[1], Xuesong Shen[2], Vivian W.Y. Tam[3]

[1]School of Built Environment, University of New South Wales, Australia
[2]School of Civil and Environmental Engineering, University of New South Wales, Australia
[3]School of Engineering, Design and Built Environment, Western Sydney University, Australia
yu.gao4@unsw.edu.au, kenneth.yiu@unsw.edu.au, x.shen@unsw.edu.au, v.tam@westernsydney.edu.au

Abstract – Construction and demolition waste (C&DW) management is central to urban renewal and sustainable construction, yet current intelligent solutions often suffer from fragmented multimodal data, weak knowledge linkage, and limited policy interpretability. Image recognition models can classify C&DW materials but cannot integrate regulatory clauses, disposal processes for explainable reasoning, while text centered large language models (LLM) retrieval lacks visual inputs and knowledge graph based path reasoning. This paper proposes a multimodal framework that integrates Contrastive Language-Image Pre-training (CLIP) ViT-B-32 for zero-shot visual evidence extraction, a lightweight Neo4j based C&DW knowledge graph for structured domain modeling, a hybrid Graph Retrieval Augmented Generation (GraphRAG) workflow for joint subgraph and vector retrieval, and GPT-4o for evidence-anchored recommendation generation. The framework links on-site images with regulatory clauses and disposal pathways to generate disposal recommendations, compliance alerts, and clause-level citations. Experimental results show that the proposed framework achieves 71% end-to-end accuracy, 100% citation coverage, and a 0% hallucination rate, outperforming single modality baselines, while remaining scalable and easy to deploy.

Keywords
GraphRAG; Multimodal; Data Integration; Recommendation; Waste disposal; Construction and demolition waste;

1 Introduction

With the expansion of urban renewal and building demolition, C&DW has become a major component of municipal solid waste [1]. Achieving accurate identification, rational sorting, and compliant disposal are therefore critical to low-carbon construction practice. In real engineering scenarios, however, C&DW information is inherently multimodal: images capture visible material characteristics, regulatory documents specify classification and disposal requirements, and structured data record key relationships such as material categories and disposal pathways. However, most existing studies focus on single modality analysis, making it difficult to integrate these highly heterogeneous data within a unified semantic space. This semantic gap leads to a disconnect between visual perception and compliance oriented decision making, which in turn hinders the practical deployment of automated management systems.

Existing vision based deep learning methods, such as convolutional neural networks (CNN) [2],YOLO [3] and pretrained visual models [4]，can classify typical waste types including concrete, timber, metal, and bricks, thereby supporting on site quantification and preliminary sorting [5]. Nevertheless, these approaches primarily address the question of what the waste is, but cannot further incorporate regulatory knowledge or disposal logic to determine how the waste should be handled or whether compliance requirements are satisfied. In contrast, natural language processing technologies represented by LLMs have demonstrated strong capabilities in policy interpretation and text based question answering [6]. However, general purpose LLMs lack domain specific structural constraints, are prone to factual hallucination, and cannot directly associate visual information, which makes it difficult for them to provide reliable and traceable decision support in strongly regulated C&DW scenarios. Knowledge graphs (KGs) provide a structured means of expressing the relationships between materials, policies, and disposal methods in C&DW [7]. However, traditional KG reasoning methods are difficult to combine with the language understanding capabilities of LLMs and are insufficient to support the unified processing of cross modal data.

Recently, the GraphRAG framework has combined graph retrieval with text retrieval, enabling a unified retrieval mechanism that integrates structured knowledge with unstructured documents. This allows LLMs to generate more precise responses grounded in relational structures [8]. However, the key limitation in current multimodal C&DW research is not merely the absence of system-level integration. More fundamentally, existing methods still lack a structure-aware multimodal reasoning mechanism that can connect visual evidence, regulatory constraints, and traceable recommendation generation within one interpretable workflow. In particular, prior multimodal RAG-based systems generally use images as supplementary context, but do not explicitly transform visual evidence into graph-grounded entities that can participate in downstream reasoning over disposal rules, facility constraints, and policy clauses.

To address these challenges, this study proposes a multimodal GraphRAG framework for C&DW compliance support. Rather than simply combining CLIP, KG, GraphRAG, the framework builds a unified multimodal reasoning pipeline in which CLIP grounds waste images into semantic entities in a lightweight C&DW KG, hybrid retrieval combines subgraph traversal with vector-based clause retrieval, and the LLM generates disposal recommendations with explicit clause-level traceability. In this way, the framework connects visual perception, structured knowledge, and evidence-grounded reasoning within a unified decision workflow.

The main objectives of this study are as follows:

(1) To construct a lightweight C&DW KG that organizes heterogeneous information on materials and policies, providing structured support for compliance reasoning.

(2) To design a hybrid retrieval mechanism that combines vector retrieval with subgraph topology so that multimodal evidence can be fused for disposal recommendation and policy clause traceability.

(3) To validate the feasibility of an end-to-end multimodal workflow that links CLIP-based visual grounding, graph-based regulatory constraints, and LLM-based evidence-anchored generation for C&DW management.

This study contributes a unified framework for visual-to-graph grounding, hybrid retrieval, and evidence-anchored compliance recommendation in multimodal C&DW management. The framework is intended as a prototype pathway for linking waste recognition with traceable disposal recommendation. Rather than claiming broad real-world generalizability at this stage, the study focuses on demonstrating the feasibility, interpretability, and auditability of the proposed workflow under a controlled experimental setting.

2 Related Work

2.1 Vision-based C&DW identification

Computer vision plays a fundamental role in automated C&DW identification. Early studies mainly relied on handcrafted image features, such as color histograms and texture descriptors, but these methods showed limited robustness in complex construction site environments [5]. With the development of deep learning, more advanced models have been adopted for C&DW recognition. For example, Lu et al. [9] used DeepLabv3+ to classify nine common C&DW categories, including concrete, debris, and wood. Fang et al. [10] proposed an improved YOLO-based method for real-time object detection and counting of mixed waste on construction sites. Other studies have further explored multi-view classification models based on ResNet-50 to recognize C&DW clusters and estimate their volume using Delaunay triangulation [11].

Although these methods have shown strong performance for perception tasks, their outputs are usually limited to predefined labels, detection boxes, or quantity estimates. As a result, they mainly answer what the waste is, but provide little support for determining how the waste should be handled or whether disposal decisions satisfy regulatory requirements. More importantly, most vision-based methods do not establish explicit semantic links between visual evidence and downstream policy clauses, facility constraints, or disposal pathways. This limitation makes them insufficient for compliance-oriented decision support, where material recognition must be connected with interpretable regulatory reasoning.

2.2 Knowledge modeling and LLMs in construction

To address the heterogeneity and unstructured nature of engineering data, KGs have been widely introduced into the construction domain. Existing studies have developed ontologies and graph-based representations to organize building codes, material properties, and safety rules, thereby improving interoperability across heterogeneous data sources [12]. However, conventional KG systems typically depend on manually constructed schemas and structured query languages such as SPARQL, which reduces their accessibility and limits their direct use in practical site-level decision scenarios.

In parallel, LLMs have shown strong capabilities in text understanding and generation, and have been applied to tasks such as contract analysis, safety checklist generation, and design consultation in construction. Their key advantage lies in flexible human-machine interaction and natural language reasoning. However, LLMs remain probabilistic models without explicit domain constraints, making them prone to factual hallucination and unsupported responses [13], especially in applications involving environmental regulations or precise engineering standards.

From the perspective of C&DW compliance management, neither KGs nor LLMs alone are sufficient. KGs provide structural precision but lack flexible interaction and generative reasoning, whereas LLMs provide semantic flexibility but lack reliable logical constraints and traceable evidence support. Therefore, the central challenge is not simply to use both techniques together, but to combine the structural reliability of KGs with the language generation capability of LLMs in a way that remains explainable and auditable for compliance-critical decisions.

2.3 Retrieval-Augmented Generation and Multimodal Fusion

Retrieval-Augmented Generation (RAG) has been proposed to reduce hallucination and improve the timeliness of LLM responses by retrieving external knowledge before generation [14]. However, conventional RAG frameworks mainly rely on vector similarity over unstructured text chunks and often overlook the explicit logical relations and hierarchical dependencies among knowledge elements. When applied to highly interconnected engineering regulations, this limitation can lead to fragmented interpretation and weak consistency across retrieved evidence.

To address this issue, GraphRAG has recently been introduced as an extension of RAG that incorporates graph topology into the retrieval process [15]. By leveraging entity relations and graph structure, GraphRAG improves the logical consistency of downstream reasoning. At the same time, multimodal foundation models such as CLIP provide a new paradigm for zero-shot visual recognition by aligning vision and text in a shared feature space [4]. These advances offer important technical foundations for multimodal compliance support.

Nevertheless, existing studies still lack a structure-aware multimodal reasoning workflow for C&DW management. Most current systems either focus on a single modality or treat multimodal inputs as loosely connected information sources. They rarely ground visual evidence into graph entities for downstream retrieval and reasoning, nor do they integrate image perception, regulatory constraints, and clause-traceable recommendation generation within one auditable pipeline. In this respect, the present study addresses these limitations through visual-to-graph grounding, hybrid retrieval, and evidence-anchored compliance recommendation within a unified framework.

3 Research Method

This section presents the proposed Multimodal GraphRAG Framework in detail. As illustrated in Fig. 1, the framework is designed to address the disconnect between visual perception and compliance knowledge in C&DW management. The overall workflow consists of three tightly coupled stages.

(1) Visual Semantic Perception, in which the CLIP model is used to map on site images into semantic queries.

(2) Hybrid Knowledge Retrieval, in which structured KGs and unstructured vector databases are jointly integrated through the GraphRAG workflow.

(3) Evidence Anchored Generation, in which interpretable decisions with explicit policy clause traceability are generated.

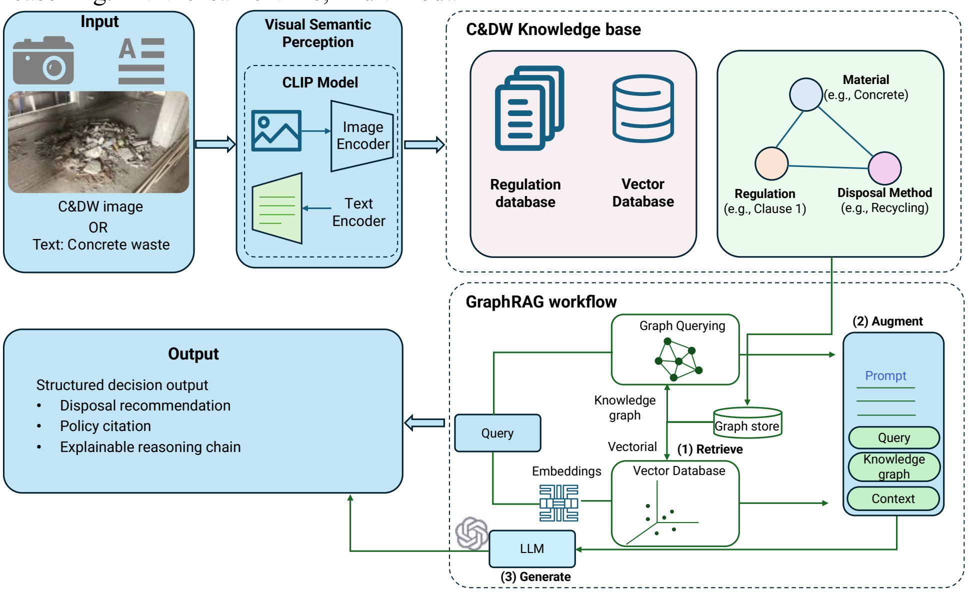

Fig.1 Multimodal C&DW data integration framework based on GraphRAG.

3.1 Visual Semantic Perception via CLIP

To bridge the semantic gap between image pixels and regulatory text, this paper introduces the CLIP model as a visual encoder. Unlike traditional classification models that only output discrete labels, CLIP can map images into a continuous semantic vector space, achieving zero-shot entity linking. Let the C&DW image collected on-site be I, and the set of material entities in the knowledge graph be defined as $M = \{m_1, m_2, ..., m_n\}$. The system first uses Prompt Engineering to convert each entity name m_i into a natural language description template T_i (e.g., "A photo of construction waste containing [Entity]"). Using CLIP's dual-stream encoder, we calculate the image embedding V_I and the text embedding V_{Ti} respectively, and then calculate their cosine similarity.

The system selects the top-k entities with the highest similarity scores S as semantic anchors. This step not only completes the initial identification but also generates structured query entry points for subsequent GraphRAG retrieval.

3.2 Construction of Hybrid Knowledge Base

To support complex reasoning, a hybrid knowledge base containing both structured and unstructured data was constructed. C&DW Knowledge Graph: We formalize domain knowledge as a graph $G = (E, R)$.

Entities (E): Contains core classes such as Material (e.g., Concrete, Asbestos), Regulation (e.g., Clause 5.1 of NSW_W1), and Disposal Method (e.g., Recycling, Landfill). Relations (R): Defines logical constraints between entities, such as (Concrete, requires, Crushing). This topological structure clarifies the deterministic relationships between "material-regulation-disposal".

Vector & Regulation Database: To compensate for the lack of detailed text descriptions in the graph, we chunk detailed regulatory documents and operation manuals, transforming them into high-dimensional vectors using an embedding model (e.g., text-embedding-3) and storing them in the Vector Database. This provides extensive contextual details to the system.

3.3 GraphRAG Inference Workflow

This is the core module of this framework. To address the hallucination issue in LLM, this study employs a dual-channel evidence injection mechanism combining retrieval-augmented generation, simultaneously incorporating structured knowledge constraints and unstructured textual details. The process comprises the following three steps:

(1) Subgraph Retrieval (Structure-Aware Path): Based on the anchor entities identified in Section 3.1 (such as "Concrete"), the system performs an h-hop subgraph traversal in the knowledge graph. The retrieval algorithm extracts a subgraph structure G_{sub} centered on the anchor with a radius of h.

$$G_{sub} = \{(e_i, r_{ij}, e_j) \mid e_i \in A, dist(e_i, e_j) \leq h\} \quad (1)$$

This subgraph explicitly captures the regulatory nodes that the material must follow and the allowed disposal nodes, forming hard logical constraints.

(2) Vector Retrieval (Semantic-Aware Path): Simultaneously, the system uses the text description output by CLIP as a query to retrieve the Top-N most relevant regulatory text chunks from the vector database. This step captures the detailed descriptions in the regulations, such as specific landfill requirements, forming a soft semantic supplement.

(3) Context fusion and evidence anchored generation: The system serializes the structured subgraph triples into natural language context C_{graph}, and fuses them with the retrieved textual context C_{text} to construct the final Augmented Prompt:

$$Prompt_{final} = Instruction \oplus C_{graph} \oplus C_{text} \oplus UserQuery \quad (2)$$

Here, $\oplus$ denotes concatenation and organization according to a predefined prompt template. The LLM generates disposal recommendations based on this augmented prompt and preserves key policy clause references and relational evidence in the output, thereby enabling interpretable decision making and policy traceability for C&DW management scenarios.

3.4 Evidence-Anchored Decision Generation

Finally, the LLM receives the fused prompt and performs chain-of-thought reasoning. To ensure engineering usability of the outputs, we embed strict formatting constraints in the prompt, requiring the model to produce three mandatory sections, as illustrated in the output of Fig. 1:

Identification Verification: confirm the identified waste type and its key attributes.

Disposal Recommendation: generate detailed recommendations on disposal procedures.

Citation and Traceability: the system enforces that each recommendation must be accompanied by a citation to its source in the knowledge base, identified by a Source ID. For example, if the recommendation is to conduct crushing treatment, the relevant regulatory clause content must be explicitly referenced.

With this end-to-end closed loop design, the system transforms unstructured visual inputs into actionable, interpretable, and compliant engineering decision reports.

4 Experiments and Results

4.1 Experimental Setup

The experimental system was implemented in a Python environment. For visual evidence extraction, OpenCLIP ViT-B-32 (pretrained weights: laion2b_s34b_b79k) was used for zero-shot image classification. On the graph side, a lightweight C&DW knowledge graph was constructed and queried using Neo4j. The GraphRAG workflow was developed through secondary implementation based on an open-source codebase. GPT-4o was adopted to generate disposal recommendations, compliance alerts, and clause interpretations. For text retrieval, sentence-transformers was used to produce vector embeddings and perform similarity-based retrieval, while graph retrieval and evidence subgraph construction were executed in Neo4j.

All experiments were conducted in a CPU-only environment without GPU acceleration to demonstrate the feasibility of deployment under low hardware requirements.

4.1.1 Dataset Construction

To validate the proposed compliance recommendation workflow that connects visual evidence, structured knowledge, and regulation-aware reasoning, a multi-source heterogeneous dataset was constructed for C&DW disposal compliance scenarios. The dataset comprises four components: a regulatory clause corpus, a facility registry, transport load case records, and a visual evidence set. A unified ID scheme (e.g., load_id, facility_id, clause_id, and Waste Type codes) is adopted to enable cross modal linkage and traceable citation. The dataset covers six major C&DW categories, including concrete and masonry, metals, timber, plasterboard, mixed C&DW, and asbestos. In total, one hundred load samples are used throughout the experiments. These samples were generated in a scripted and controlled manner, rather than collected as a large-scale real-world dataset. Their purpose is to evaluate the feasibility, logical consistency, and auditability of the proposed workflow, rather than to support strong claims regarding large-scale robustness, scalability, or generalizability.

(1) Regulatory Clause Corpus

Clauses relevant to classification, disposal pathways and licensing are organized in a structured table. The corpus mainly draws on Australian guidance (e.g., Waste Classification Guidelines Part 1: Classifying waste). Each record includes clause_id, waste type/scope (W1–W6), and requirement. It serves as the evidence source for both baselines and GraphRAG based framework.

(2) Load Cases and Ground Truth

Transport loads are used as the basic decision unit. For each load_id, the dataset records the true waste type (true_type) and the intended receiving facility (facility_id). Load records are stored in loads.csv and serve as ground truth for compliance evaluation. Metrics such as decision correctness (is_correct) and citation consistency (e.g., top1_is_in_graph_clause) are computed. The case set is scripted at a fixed scale, with consistent mappings among load_id, waste type, facility, clause_id, enabling batch verification of the reasoning chain.

(3) Visual Evidence and Cross Modal Binding

Each load case is associated with an on-site image, with mappings stored in image_manifest.csv (load id, image path). Images are classified via CLIP in a zero-shot manner, and predictions are saved in visual_evidence.csv, including pred_type (W1–W6), pred_score, topk types/topk_scores, and model configuration fields (clip_model, clip_pretrained, prompt_set). The visual evidence table is imported into Neo4j and linked to load id, enabling end-to-end reasoning from images to graph retrieval, clause evidence, and explainable generation.

4.1.2 Knowledge Graph Construction & Indexing

A lightweight C&DW-KG is constructed and deployed in Neo4j. The graph uses transport loads (Load) as the core object nodes and links them to waste types (Waste Type), disposal facilities (Facility), and regulatory clauses (RegClause). RegClause nodes are uniquely identified by clause id and store the clause text and topic fields, while Facility nodes include eligibility attributes such as accepted treatment modes. The applicability scope between Waste Type and RegClause is represented via the APPLIES_TO relationship, and licensing requirements between RegClause and License Level are encoded through the REQUIRES_LICENSE relationship. To support retrieval augmented reasoning, clause texts are further embedded using sentence-transformers, and a FAISS nearest neighbor index is built to enable efficient top-k retrieval of textual evidence.

4.1.3 End-to-End GraphRAG Pipeline

The system takes as input either a single on-site photograph or an image record linked to a transport load. First, CLIP is applied for zero-shot classification to predict the waste type with confidence scores. The prediction is then recorded as “visual evidence” in the visual evidence table to enable traceable archiving. Next, the predicted type is used to locate the corresponding Waste Type node in the C&DW KG, followed by a two-hop neighborhood expansion to construct an evidence subgraph that captures candidate facilities, licensing constraints, and relevant regulatory clauses for the given load. In parallel, vector-based retrieval is performed over the clause corpus (top-k=three) to recall semantically similar clauses. Finally, the combined evidence from the graph subgraph and retrieved text is provided to the language model to generate disposal pathway recommendations, compliance alerts, and clause interpretations. The output includes structured citations in the form of clause id to support auditing and verification.

4.1.4 Baselines

To evaluate the gains of GraphRAG in interpretability and citation traceability, three baselines are considered: (1) Keyword-only: top-k evidence is retrieved from the clause corpus using keyword matching, and an LLM generates the explanation based on the retrieved evidence. (2) Vector-only: top-k clauses are retrieved using vector similarity search, and an LLM generates the explanation based on the retrieved evidence. (3) LLM-only: no external evidence is provided, and the LLM answers solely based on the user query.

All comparison methods use the same language model and the same output format constraints, including the requirement to report clause id, in order to isolate the impact of retrieval and graph based mechanisms.

4.1.5 Evaluation Metrics

System performance is evaluated along three dimensions: decision correctness, evidence traceability, and citation grounding. (1) End-to-end accuracy measures the overall correctness of the system by jointly evaluating visual recognition and compliance decision making. A prediction is considered correct only when both the waste type identification is correct and the final compliance decision is correct; otherwise, it is counted as incorrect. (2) Citation coverage: whether structured clause identifiers (clause_id) are included in the response. (3) Hallucination rate measures the proportion of responses containing unsupported claims. Hallucinations were detected through an automated consistency-checking mechanism against the scripted reference mappings and retrieved regulatory evidence. A response was counted as hallucinated if it cited a nonexistent clause_id or included unsupported regulatory claims not grounded in the retrieved clause evidence.

4.2 Results Analysis

4.2.1 Model Evaluation

To evaluate the practical applicability of the proposed end-to-end pipeline integrating visual recognition, graph retrieval, and LLM-based explanation generation, load case L0001 is analyzed using an on-site C&DW photograph (Fig. 2). The image exhibits typical visual heterogeneity, background interference, and unclear material boundaries. This description characterizes visual complexity only and does not presuppose the regulatory waste category. The case tests whether the system can identify the dominant waste type and produce traceable compliance recommendations under uncertain visual input by leveraging regulatory and facility constraints.

Step 1: CLIP zero-shot visual recognition.

The system applies CLIP in a zero-shot setting to produce top-k candidate waste types with similarity weights. For L0001, the top-1 prediction is W2 (metals), but several alternatives (e.g., W2/W3/W4) have close scores, indicating no decisive visual winner, consistent with mixed material appearances and motivating downstream evidence-based reasoning.

Step 2: GraphRAG evidence retrieval and fusion.

A compliance oriented query is constructed from (i) the candidate waste types and (ii) a disposal/licensing question template, and evidence is retrieved from two sources in parallel. Regulatory retrieval (text evidence): the clause corpus is queried via sentence-transformers vector retrieval to recall the most relevant clauses for the candidate types; the top match and similarity score are recorded.

Graph retrieval (structural constraints): in the C&DW-KG, the predicted type is mapped to a WasteType node and expanded to a two-hop local subgraph to collect related RegClause and feasible Facility nodes with their license level attributes. Facilities are filtered by license compatibility with the clause-implied requirements, removing structurally infeasible destinations even if they appear plausible from text alone.

By fusing retrieved clauses with graph-structured constraints, uncertainty from visually ambiguous recognition is converted into verifiable regulatory conditions and facility feasibility checks, reducing the risk of producing recommendations that are unsupported or operationally invalid.

Step 3: LLM explanation generation with clause-level traceability.

Retrieved clause and graph-derived feasibility context are organized into a structured prompt, requiring clause-level citations (via clause id) for each key recommendation. For load L0001, the fused evidence supports a recovery-oriented routing decision. The generated report recommends routing the load to a Material Recovery Facility and cites the matched regulatory clause (for example, NSW_W2_01) as the basis for the recommendation. Importantly, the explanation attributes the decision to a combined evidence chain, including (i) the candidate type set produced by CLIP, (ii) the retrieved regulatory clause content, and (iii) the facility license compatibility derived from the C&DW-KG, rather than relying on a single visual similarity score.

This case shows that, even under visually ambiguous conditions with closely distributed top-k scores, the hybrid GraphRAG pipeline can generate actionable and interpretable recommendations with clause-level traceability by combining regulatory text evidence and graph-based facility constraints.

Fig. 2. End-to-end visual compliance analysis report for a C&DW image.

4.2.2 Performance Analysis

Based on batch experiments on 100 load samples, the performance of Hybrid GraphRAG is compared with three baseline methods along three dimensions: compliance decision accuracy, evidence traceability, and hallucination rate.

Table 1 reports the comparative results in terms of decision correctness and auditability of citations. Hybrid GraphRAG achieves 71% end-to-end accuracy, while maintaining 100% citation coverage and a 0% hallucination rate. This indicates that the proposed approach not only produces stable compliance decisions, but also consistently provides traceable and verifiable clause-level support.

Table 1. Performance comparison.

Models	End-to-end Accuracy	Citation Coverage	Hallucination Rate
Baseline keyword	10%	91%	0
Baseline vector	59%	71%	0
Baseline LLM	18%	84%	84%
Proposed model	71%	100%	0

In contrast, Baseline_keyword attains a end-to-end accuracy of only 10%, suggesting that keyword-based retrieval often fails to surface the critical information required for licensing-level constraints. Baseline_vector improves end-to-end accuracy to 59%, but its citation coverage drops to 71%, indicating that semantic retrieval is more effective for evidence recall yet still suffers from unstable structured citation in the final output. Baseline LLM, which provides no retrieved evidence, achieves a decision accuracy of 18% and exhibits substantial hallucination risk (84%). Although it can generate plausible narrative suggestions, it fails to meet the minimum requirements for grounded citation and auditability in compliance critical scenarios.

The results show that the proposed hybrid GraphRAG model provides more reliable C&DW compliance recommendations than retrieval-only or generation-only baselines, consistently outperforming them in correctness, citation coverage, and hallucination avoidance.

5 Discussion

This study shows that integrating visual evidence extraction with structure-aware GraphRAG enhances the reliability and auditability of C&DW compliance recommendations. However, several technical, practical, and deployment-related limitations should be acknowledged before broader real-world use.

First, the current evaluation is based on a lightweight regulatory clause corpus, a limited facility registry, and one hundred scripted load samples. While this controlled setup is suitable for validating the feasibility, logical consistency, and auditability of the proposed workflow, it is not sufficient to support strong claims regarding robustness, scalability, or generalizability. Future work should therefore extend the evaluation to larger regulatory corpora, multiple jurisdictions, and more diverse disposal scenarios.

Second, end-to-end accuracy depends on the CLIP recognition output, and visual errors can propagate to downstream reasoning. CLIP-based zero-shot classification is sensitive to image quality and site conditions such as occlusion, blur, lighting variation, and mixed stacking, which may lead to ambiguous top-k predictions. Although the images used in this study were collected from real construction sites, we selected those in which the target materials were relatively easier to distinguish. Therefore, the current evaluation does not yet fully represent broader site-level conditions involving complex backgrounds and multiple-object interference. Future work should improve robustness through more realistic site-level images, prompt refinement, low-confidence filtering, and lightweight human-in-the-loop verification.

Third, the current mapping of regulatory clauses into graph entities and relations still relies on manual rule engineering for clause segmentation, scope binding, and license requirement extraction. This limits portability across jurisdictions and increases maintenance effort. In addition, the proposed method remains a prototype-level framework rather than a deployment-ready system, and issues such as continuous regulatory updates, operational integration, and large-scale field validation remain to be addressed.

6 Conclusions

This study proposed a multimodal framework for C&DW management that integrates visual semantic grounding, lightweight KG construction, and GraphRAG based evidence retrieval to support regulation-aware decision making. Unlike prior approaches that focus only on image classification or text-based reasoning, the proposed method connects perceptual input, structured domain knowledge, and clause-grounded reasoning within a unified workflow. Experimental results show that the proposed framework achieved 71% end-to-end accuracy, 100% citation coverage, and 0% hallucination rate under a controlled evaluation setting. Compared with the baselines, it demonstrated improved compliance consistency and evidence grounding in C&DW recommendation tasks. These findings support the feasibility of integrating multimodal perception, structured knowledge, and evidence-anchored reasoning for auditable C&DW compliance support.

However, the current study should be regarded as a prototype-level validation rather than evidence of broad robustness or generalizability. Future work should extend the evaluation to larger datasets, broader regulatory corpora, and more realistic site-level image scenarios.

References

[1] Y. Gao, T.W. Yiu, X. Shen, V.W.Y. Tam, Life cycle insights into construction and demolition waste management: Past, present and emerging futures, Journal of Building Engineering, 111 (2025), pp. 113441.

[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems, 25 (2012).

[3] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:.02767, (2018).

[4] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, et al., Learning Transferable Visual Models From Natural Language Supervision, Proceedings of the 38th International Conference on Machine Learning. Vol. 139, PMLR, Proceedings of Machine Learning Research, 2021 of Conference, pp. 8748--8763.

[5] Y. Gao, J. Wang, X. Xu, Machine learning in construction and demolition waste management: Progress, challenges, and future directions, Automation in Construction, 162 (2024), pp. 105380.

[6] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, et al., Large language models: A survey, arXiv preprint arXiv:.06196, (2024).

[7] S. Ji, S. Pan, E. Cambria, P. Marttinen, P.S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, IEEE Transactions on Neural Networks and Learning Systems, 33 (2) (2022), pp. 494-514.

[8] B. Peng, Y. Zhu, Y. Liu, X. Bo, H. Shi, et al., Graph Retrieval-Augmented Generation: A Survey, ACM Trans. Inf. Syst., (2025).

[9] W. Lu, J. Chen, F. Xue, Using computer vision to recognize composition of construction waste mixtures: A semantic segmentation approach, Resources, Conservation and Recycling, 178 (2022), pp. 106022.

[10] H. Fang, J. Chen, M. Wang, Q. Wu, Z. Wang, Real-time detection of construction and demolition waste impurities using the improved YOLO-V7 network, Journal of Material Cycles and Waste Management, 26 (4) (2024), pp. 2200-2213.

[11] A. Jaiswal, K. Jha, N. Bugalia, Q.P. Ha, Vision-based volumetric estimation of localized construction and demolition waste, Waste Management, 206 (2025), pp. 115046.

[12] B.T. Zhong, L.Y. Ding, H.B. Luo, Y. Zhou, Y.Z. Hu, et al., Ontology-based semantic modeling of regulation constraint for automated construction quality compliance checking, Automation in Construction, 28 (2012), pp. 58-70.

[13] H. Pu, X. Yang, J. Li, R. Guo, AutoRepo: A general framework for multimodal LLM-based automated construction reporting, Expert Systems with Applications, 255 (2024), pp. 124601.

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems, 33 (2020), pp. 9459-9474.

[15] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, et al., From local to global: A graph rag approach to query-focused summarization, arXiv preprint arXiv:.16130, (2024).