A Multimodal Visual Retrieval Framework for Construction Site Management via High-Dimensional Vector Embeddings

Qiao Zheng[1], Zhiqiang Wei[1], Minjin Fang[1], Rui Jin[1*] and Xia Lei[1]

[1] Zhejiang Construction Investment Group Co., Ltd., China.

[*] Corresponding author

zhengqiao@u.nus.edu, hzrui@126.com

Abstract –

The exponential growth of visual data in construction management has rendered manual image tagging inefficient and error-prone. While traditional computer vision methods offer automation, their reliance on closed-set supervised learning lacks the semantic flexibility required for dynamic, open-vocabulary site querying. To address this limitation, this study proposes a multimodal visual retrieval framework leveraging the Doubao-embedding-vision model. The system maps visual and textual features into a shared 2048-dimensional continuous vector space, utilizing a Hierarchical Navigable Small World (HNSW) index for efficient approximate nearest neighbor search. Validated on a dataset of 1,000 construction site images, the framework demonstrated robust open-vocabulary capabilities, effectively handling both specific object-attribute combinations and complex compositional queries. Quantitative evaluations revealed high retrieval precision (93.80%) and exceptional computational efficiency, achieving an average query latency of 0.0899 seconds. Although challenges remain in precise attribute binding and logical negation handling, the proposed retrieval framework proves highly effective for retrospective analysis. By complementing real-time automated detection systems, this framework transforms passive CCTV archives into an actively searchable semantic database, providing significant practical value for post-incident investigations, dynamic compliance auditing, and intelligent site monitoring.

Keywords –

Multimodal Retrieval; Vector Embeddings; Construction Site Monitoring; Semantic Search; Attribute Binding

1 Introduction

The continuous influx of Closed-Circuit Television (CCTV) footage on modern construction sites offers unprecedented potential for real-time monitoring [1,2]. However, this data richness often creates an information bottleneck; raw visual data remains unstructured and inaccessible without efficient processing mechanisms. Traditional supervision strategies, heavily reliant on manual observation, are labor-intensive, susceptible to cognitive fatigue, and inherently unscalable in the face of massive video datasets [3,4].

To automate this process, Computer Vision (CV) solutions, particularly those utilizing Convolutional Neural Networks (CNNs), have been widely deployed [5]. However, their supervised paradigm struggles with the immense, heterogeneous nature of construction environments. A typical site comprises vast taxonomies of personnel, diverse materials, and specialized machinery [3,6-8]. Recognizing this extreme diversity under traditional frameworks requires exhaustive datasets and rigid "class-by-class" training, imposing prohibitive annotation workloads [9] and lacking the flexibility required for dynamic site management [10]. Recently, the emergence of Vision-Language Models (VLMs) and advanced multimodal embedding techniques has introduced a paradigm shift from closed-set classification to open-vocabulary semantic understanding [11]. However, deploying these generalized models directly into the highly specialized, dynamic, and unstructured domain of construction sites remains a significant challenge, particularly concerning the accurate retrieval of compositional attributes without extensive fine-tuning [12].

To overcome these limitations, this study proposes a multimodal retrieval framework leveraging high-dimensional vector embedding technology. Diverging from the fixed taxonomy of traditional classifiers, this approach maps visual and textual features into a shared continuous vector space, reformulating recognition as a semantic similarity calculation [13,14]. This methodology facilitates flexible, open-vocabulary retrieval without the need for task-specific retraining. By

integrating frame extraction from existing video streams with an efficient indexing vector database, the proposed system seamlessly augments legacy digital infrastructure, offering a cost-effective pathway to intelligent monitoring. The specific contributions and the contextual background of our chosen embedding approach are detailed in the subsequent sections.

2 Literature Review

This section reviews related work on visual data processing and multimodal embedding technologies in the context of construction site management. Section 2.1 outlines the evolution of visual embedding techniques, from traditional Convolutional Neural Networks (CNNs) to modern Vision-Language Models (VLMs). Section 2.2 compares state-of-the-art multimodal models, justifying the selection of the Doubao-embedding-vision model for this study. Finally, Section 2.3 identifies key research gaps in current intelligent monitoring systems and highlights the advantages of the proposed retrieval framework.

2.1 Evolution of Visual Embedding Techniques

The processing of visual data in construction management has historically relied on task-specific Convolutional Neural Networks (CNNs) [5]. While effective for predefined tasks, such as detecting standard personal protective equipment (PPE) [10], these models lack semantic flexibility. The advent of Contrastive Language-Image Pre-training (CLIP) [15] marked a significant breakthrough by aligning images and text within a shared high-dimensional vector space, enabling zero-shot recognition capabilities. Following CLIP, the field of Vision-Language Models (VLMs) has expanded rapidly. Advanced frameworks, such as BLIP-2 [16] and the Qwen-VL series [17], have integrated large language models with frozen image encoders, significantly enhancing multimodal comprehension and complex visual reasoning [18].

2.2 Model Selection

While CLIP variants established the foundation for cross-modal retrieval, recent evaluations indicate limitations in their ability to handle fine-grained attribute binding and complex compositional reasoning [19]. More recent state-of-the-art models, particularly the Qwen-VL series (including the latest iterations like Qwen3-VL) [20] and the Doubao-embedding-vision model [21], have demonstrated superior performance in capturing detailed visual semantics. Specifically, quantitative evaluations on the recent MMEB-V2 benchmark illustrate these performance capabilities [22]. In the overall assessment for image-based tasks, mainstream open-source models typically achieve scores ranging from 59.7 to 75.9. In contrast, the Doubao-embedding-vision model (evaluated as Seed-1.6-embedding-0615) attains a superior overall score of 77.8, effectively outperforming the Qwen3-VL-Embedding-2B model, which scores 75.0. While the larger Qwen3-VL-Embedding-8B model reaches a higher score of 80.1, the substantial increase in its parameter size inherently introduces significantly higher computational overhead and computing costs. Therefore, to optimally balance state-of-the-art retrieval accuracy with the computational efficiency required for continuous construction site monitoring, the Doubao-embedding-vision model was selected as the representative engine for this validation.

2.3 Research Gaps and Advantages of the Proposed Framework

Despite the rapid advancements in visual embeddings, a distinct research gap remains in their practical application to construction site management. Most existing intelligent monitoring systems either rely on rigid CNNs that fail in open-set scenarios or explore generalized VLMs that still face critical challenges in zero-shot construction site deployment without task-specific fine-tuning [23]. Furthermore, generalized embedding models often struggle with attribute binding (e.g., accurately linking a specific color to a specific object in a crowded scene) and logical negation [19]. This research bridges this gap by proposing a streamlined, retrieval-centric framework. By strictly utilizing the embedding layer and coupling it with a HNSW index, the proposed system circumvents the high inference latency of generative models. The primary advantage of this research lies in offering a highly efficient, open-vocabulary semantic search mechanism that integrates directly with legacy CCTV infrastructure. It provides an honest assessment of current vector space limitations (such as attribute binding and negation handling) while establishing a practical baseline for zero-shot construction site monitoring.

3 Methodology and System Framework

To address the challenges of unstructured visual data in construction management, we designed a four-layer architecture that transforms raw CCTV footage into a semantic retrieval system. This section details the system architecture, the data processing workflow, and the core retrieval mechanisms.

3.1 Framework Overview

The proposed framework is structured into four distinct layers: the data source and ingestion layer, the data management layer, the logic and indexing layer, and the interaction and application layer. This hierarchical design ensures that the system is compatible with legacy infrastructure while providing advanced semantic capabilities. As illustrated in Figure 1, the architecture supports two primary workflows: the Preprocessing Workflow (shown in solid blue arrows), which handles data ingestion and indexing, and the Querying Workflow (shown in dashed orange arrows), which manages real-time semantic retrieval.

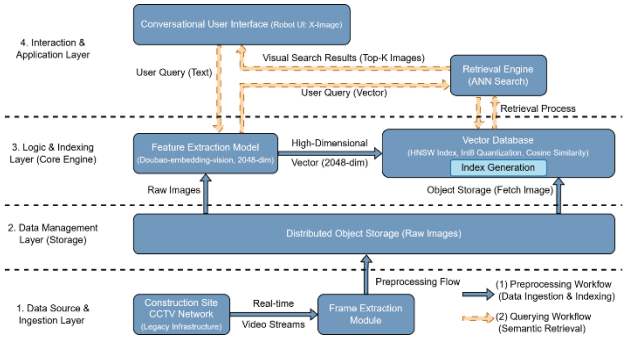

Figure 1. The proposed multimodal visual retrieval framework

Layer 1: Data Source & Ingestion Layer

This layer interfaces directly with the existing Construction Site CCTV Network. Unlike systems requiring specialized hardware, our framework integrates with legacy digital infrastructure by capturing real-time video streams. A Frame Extraction Module then samples these streams to generate discrete images for analysis.

Layer 2: Data Management Layer

This storage layer utilizes Distributed Object Storage to handle the massive volume of raw images generated by the extraction module. It acts as the central repository, ensuring data persistence and accessibility for the upper layers.

Layer 3: Logic & Indexing Layer

This is the computational heart of the system. It contains the Feature Extraction Model (Doubao-embedding-vision) which converts images and text into 2048-dimensional vectors. It also hosts the Vector Database, which manages the HNSW index, Int8 quantization, and cosine similarity calculations.

Layer 4: Interaction & Application Layer

The top layer provides the interface for construction managers. We developed a conversational user interface named X-Image (Robot UI), which facilitates natural language interaction and visualizes search results.

3.2 Data Preprocessing and Vectorization

The foundation of the retrieval system lies in the effective transformation of visual data into mathematical vectors. The preprocessing workflow begins at the bottom layer, where the frame extraction module captures images from the CCTV streams. These raw images are uploaded to the Distributed Object Storage.

Once stored, the images are fed into the feature extraction model. we utilized the doubao-embedding-vision model for this task. Unlike traditional models that output class probabilities, this model maps the semantic content of an image (e.g., objects, colors, spatial relationships) into a high-dimensional dense vector of 2048 dimensions. This high dimensionality is crucial for capturing the nuanced details of construction site elements, distinguishing between visually similar objects like different types of heavy machinery.

3.3 Vector Database and Index Generation

To enable real-time searching within a large dataset, linear scanning of vectors is inefficient. Therefore, the generated image vectors are stored in a specialized Vector Database. Within this database, we employ Index Generation techniques to structure the data for rapid retrieval. The proposed method utilizes the HNSW algorithm to generate the index, which creates a multi-layer graph structure that allows for logarithmic time complexity in approximate nearest neighbor searches. To further optimize storage and retrieval speed, Int8 Quantization is applied, compressing the vectors with minimal loss in precision.

3.4 Semantic Retrieval Mechanism

The retrieval workflow is initiated by the user through the X-Image interface. This process follows a specific path designed to bridge the semantic gap between natural language and visual data:

User Query (Text): The user inputs a natural language description (e.g., “Red Truck”) into the X-Image conversational interface.

Vectorization: This text query is sent to the Feature Extraction Model (the same model used for images) to be encoded into a user query (vector). This ensures that both text and images reside in the same 2048-dimensional vector space.

Retrieval Engine: The vector query is passed to the retrieval engine. This engine interacts with the vector database through a bi-directional retrieval process. The similarity between the query vector (T) and the image vector (I) is calculated using cosine similarity (as shown

in Equation (1)).

$$similarity(T, I) = \frac{T \cdot I}{\|T\| \cdot \|I\|} = \frac{\sum_{u=1}^{n} T_u I_u}{\sqrt{\sum_{u=1}^{n} T_u^2} \cdot \sqrt{\sum_{u=1}^{n} I_u^2}} \quad (1)$$

where n=2048, and $T \cdot I$ denotes the dot product of the two vectors.

Visual Search Results: The retrieval engine identifies the top-k images with the highest similarity scores, where the k value is customized. These results, along with their specific similarity scores and rankings, are returned to the X-Image interface for display.

4 Experimental Results and Discussion

To evaluate the efficacy of the proposed framework, a preliminary validation was conducted using real-world data collected from active construction sites.

4.1 Dataset Statistics

A pilot dataset of N=1,000 images was curated for this study. The dataset reflects the heterogeneous and unstructured nature of construction environments. It is important to note that while the dataset consists of 1,000 distinct source images, the total count of annotated object instances (N_{obj}=1,110) exceeds the image count. This discrepancy arises because construction site footage frequently captures multi-object scenes, where a single frame contains multiple co-existing entities (e.g., a worker standing next to a concrete mixer).

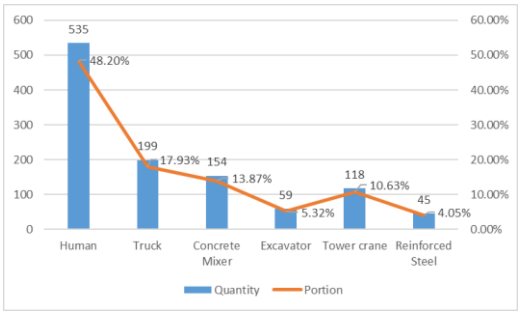

Figure 2. Statistical distribution of object categories within the validation dataset

Figure 2 illustrates the class distribution of the dataset using a dual-axis combination chart. The blue bars (left axis) represent the absolute quantity of each category, while the orange line (right axis) indicates their proportional distribution. "Humans" constitute the most frequent category (535 instances, 48.20%), followed by machinery such as "Trucks" (199 instances, 17.93%) and "Concrete Mixers" (154 instances, 13.87%).

To ensure the system's robustness across different video feed qualities, the dataset includes images of varying resolutions, consistent with the output of standard onsite monitoring equipment. As shown in Figure 3, the majority of the data (75.1%) consists of standard Full HD (1920×1080) images, which provides a representative baseline for validating the vectorization accuracy.

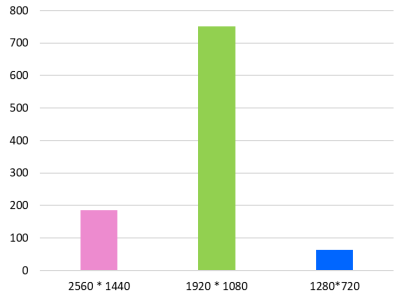

Figure 3. Distribution of image resolutions in the pilot dataset

4.2 Retrieval Performance Verification

System retrieval performance was evaluated via the X-Image conversational UI, where users retrieved relevant images by specifying a top-k parameter based on cosine similarity. Benchmark tests were deployed on a single-core 6th Generation Intel® Xeon® Scalable Processor with a 2.8 GHz base frequency and 8 GB of RAM, without dedicated GPU acceleration. Across 10 independent trials, the average index build time for the 1,000-image dataset was 268.6 seconds, while the average query latency was merely 0.0899 seconds.

4.2.1 Quantitative Evaluation

To facilitate a rigorous performance evaluation, a two-step manual verification methodology was implemented. Initially, the entire dataset was exhaustively annotated to determine the absolute number of relevant images (Total Ground Truth) for four specific compositional queries: "White Concrete Mixer", "Red Truck", "Blue Truck", and "Orange Excavator". Subsequently, a global cosine similarity threshold of 0.35 was applied to filter the vector search results. The subset of images exceeding this threshold (Returned Images) was then manually reviewed to identify the true positives (Accurately Retrieved Images).

System efficacy was quantified utilizing standard information retrieval metrics, specifically precision and recall. Precision delineates the proportion of accurately retrieved true positives relative to the total number of returned results. Recall represents the fraction of correctly classified actual positives out of all actual relevant images present in the dataset. The calculations

are defined by the following Equation (2) and (3):

$$Precision = \frac{Accurately\ Retrieved\ Images}{Returned\ Images} \quad (2)$$

$$Recall = \frac{Accurately\ Retrieved\ Images}{Total\ Ground\ Truth} \quad (3)$$

The quantitative findings are summarized in Table 1.

Table 1 Quantitative retrieval performance for specific queries ($Similarity\ Threshold\ \geq 0.35$)

Query	Returned Images	Accurately Retrieved	Total Ground Truth	Precision	Recall
White concrete mixer	129	121	134	93.80%	90.30%
Red truck	57	53	82	92.98%	64.63%
Blue truck	37	29	49	78.38%	59.18%
Orange excavator	151	116	118	76.82%	98.31%

As demonstrated in Table 1, the retrieval framework exhibits substantial variance in performance metrics depending on the specific object category. Under the uniform similarity threshold of 0.35, the "White Concrete Mixer" query achieves a highly balanced performance, with a precision of 93.80% and a recall of 90.30%. In contrast, the "Orange Excavator" query retrieves almost all relevant instances (98.31% recall) but suffers a relative decline in precision (76.82%), indicating an over-retrieval of false positives at this threshold. Conversely, the "Blue Truck" query experiences a severe drop in recall to 59.18%, meaning many true positives scored below the 0.35 threshold. This performance disparity exposes that the application of a uniform similarity threshold across diverse queries is fundamentally suboptimal. Distinct semantic categories exhibit varying degrees of feature dispersion within the 2048-dimensional embedding space, causing their underlying similarity score distributions to differ significantly. Consequently, for real-world construction site monitoring, the implementation of adaptive or category-specific similarity thresholds is strongly recommended.

4.2.2 Successful Attribute Identification

The system demonstrated high proficiency in retrieving objects with distinct color attributes. For the query "Red Truck", the X-Image interface successfully returned relevant images, as shown in Figure 4. Similarly, the query for "Blue Truck", "White Concrete Mixer", and "Orange Excavator" yielded accurate results (as shown in Figure 5, Figure 6, and Figure 7).

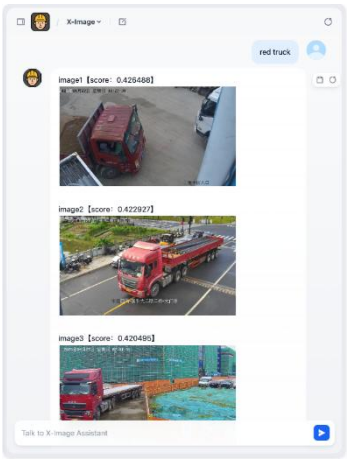

Figure 4. Retrieval results for the query "Red Truck"

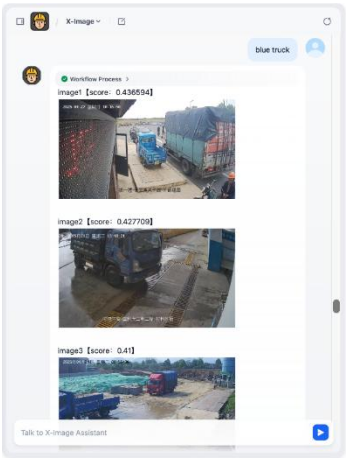

Figure 5. Retrieval results for the query "Blue Truck"

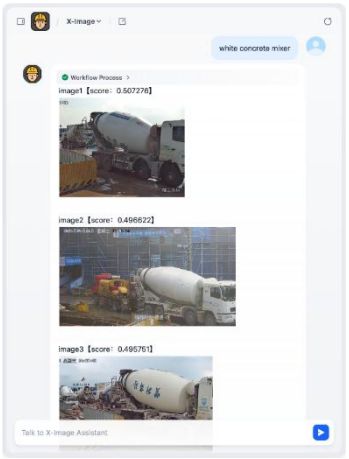

Figure 6. Retrieval results for the query "White Concrete Mixer"

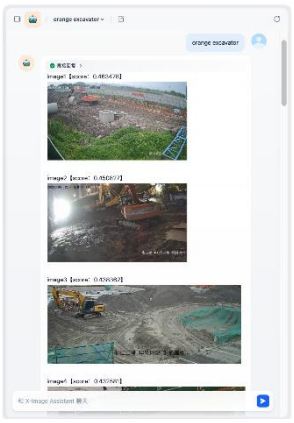

Figure 7 Retrieval results for the query “Orange Excavator”

4.2.3 Compositional and Open-Vocabulary Query Handling

While single-attribute queries demonstrate baseline feature extraction, the fundamental utility of a multimodal semantic system lies in its open-vocabulary flexibility to process complex, compositional queries involving multiple distinct entities, unannotated interactions, and spatial relationships. To explicitly verify this capability, retrieval tests utilizing descriptive, multi-subject sentences were conducted. As illustrated in Figure 8, the query “The worker is standing beside the concrete mixer” successfully retrieved contextual scenes capturing both the human entity and the specified heavy machinery in spatial proximity. As shown in Figure 9, the retrieval engine accurately identified the interactive, unannotated context between the worker and the vehicle, while strictly adhering to the designated color attribute. These findings validate that the proposed method effectively encode complex scene semantics, proving the framework's viability for open-vocabulary situational awareness on construction sites.

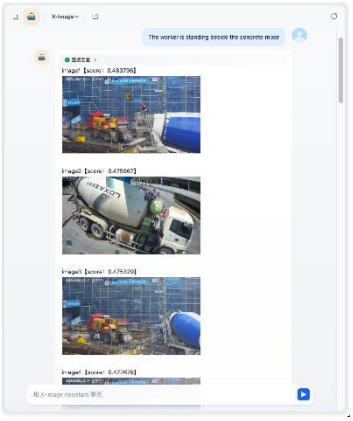

Figure 8. Retrieval results for the query “The worker is standing beside the concrete mixer”

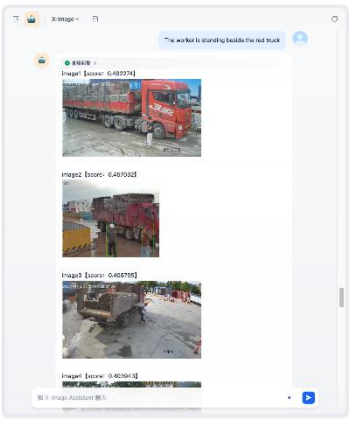

Figure 9. Retrieval results for the query “The worker is standing beside the red truck”

4.2.4 Challenge of Attribute Binding

While distinct colors were handled well, the system encountered challenges with the ability to correctly associate a specific attribute with a specific object when multiple similar objects exist in the vector space. For instance, when querying for “White Truck” (as shown in Figure 10), the system exhibited difficulty in distinguishing between white trucks and white passenger cars. This phenomenon suggests that within the current embedding space, the feature weighting for color attributes may be disproportionately high compared to object shape features, resulting in retrieval ambiguity within complex scenarios.

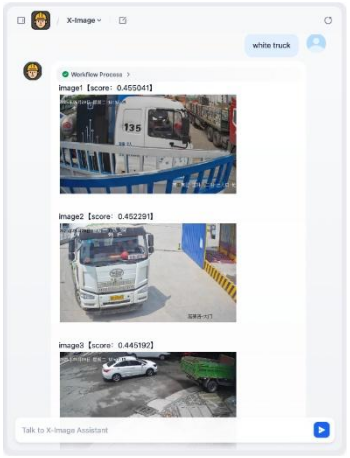

Figure 10. Retrieval results for the query “White Truck”

4.2.5 Challenge of Negation Handling

A critical requirement for safety monitoring is the identification of non-compliance, such as missing personal protective equipment. This study tested the system with the query “Human without helmet”. While

the system retrieved images of workers, it often failed to strictly adhere to the "without" negation, retrieving images of workers wearing helmets as well (e.g., image 3 in Figure 11 shows a worker with a yellow helmet). This highlights a limitation in current vector space representations regarding negation handling. The semantic embedding for "Helmet" and "without helmet" are spatially proximal because they share the same semantic feature. The vector algebra does not inherently capture the logical operation of negation, making it difficult to filter out safety gear purely through similarity search without additional negative prompts.

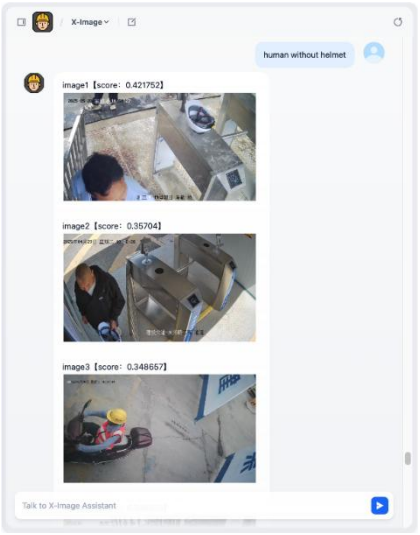

Figure 11. Retrieval results for the query " Human without helmet"

5 Conclusion

This study successfully developed and validated a multimodal visual retrieval framework tailored for the construction industry, integrating a high-dimensional vector embedding model with HNSW indexing. Empirical validation on a 1,000-image dataset confirmed the system's exceptional computational efficiency, achieving sub-second query latency (0.0899 seconds) on a cost-effective, CPU-only hardware environment. The framework demonstrated robust open-vocabulary capabilities, accurately processing not only single-attribute queries (e.g., "Red Truck") but also complex compositional queries involving unannotated spatial relationships between diverse site entities.

However, the empirical evaluation also identified critical boundaries within current vector space representations. Specifically, attribute binding remains a challenge, as dominant color features can occasionally obscure subtle object morphological distinctions. Furthermore, negation handling proved inherently difficult, as vector algebra struggles to semantically separate logical opposites (e.g., "Human without helmet" versus "Helmet") without additional constraints. Future work will focus on refining feature extraction to balance attribute weights and exploring hybrid search architectures that integrate continuous vector similarity with discrete logical filtering.

Crucially, this research clarifies the strategic role of vector-based retrieval in modern site management. While existing real-time automated detection systems (e.g., CNNs) are indispensable for immediate hazard warnings, their fully-supervised paradigms severely limit the generalizability. The proposed framework bridges this gap by acting as a semantic search engine. By transforming passive, unstructured legacy CCTV footage into an actively searchable database, the system provides profound practical value for complex management scenarios, such as post-incident investigations, dynamic compliance auditing, and the automated curation of safety training materials. Rather than relying on low-efficiency manual video scrubbing, this open-vocabulary approach empowers site managers to instantly locate unannotated events.

Acknowledgement

This research is supported by Zhejiang Provincial "Pioneer" and "Leading Goose" Science and Technology Plan Project (2026C04040); and Zhejiang Provincial Construction Scientific Research Project (2025K312).

References

[1] R. Sacks et al., Construction with digital twin information systems, Data-Centric Engineering, 1, (2020), p. e14. https://doi.org/10.1017/dce.2020.16

[2] J. Teizer, Status quo and open challenges in vision-based sensing and tracking of temporary resources on infrastructure construction sites, Advanced Engineering Informatics, 29, (2), (2015), pp. 225-238. https://doi.org/10.1016/j.aei.2015.03.006

[3] W. Fang et al., Falls from heights: A computer vision-based approach for safety harness detection, Automation in Construction, 91, (2018), pp. 53-61. https://doi.org/10.1016/j.autcon.2018.02.018

[4] C. Zeng et al., Self-supervised learning for point cloud data: A survey, Expert Systems with Applications, 237, (2024), p. 121354. https://doi.org/10.1016/j.eswa.2023.121354

[5] L. Ding et al., A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory,

Automation in Construction, 86, (2018), pp. 118-124. https://doi.org/10.1016/j.autcon.2017.11.002

[6] A. Dimitrov et al., Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections, Advanced Engineering Informatics, 28, (1), (2014), pp. 37-49. https://doi.org/10.1016/j.aei.2013.11.002

[7] X. Yan et al., Computer Vision-Based Intelligent Monitoring of Disruptions due to Construction Machinery Arrival Delay, Journal of Computing in Civil Engineering, 39, (3), (2025), p. 04025011. https://doi.org/doi:10.1061/JCCEE5.CPENG-6178

[8] B. Xiao et al., Development of an Image Data Set of Construction Machines for Deep Learning Object Detection, Journal of Computing in Civil Engineering, 35, (2), (2021), p. 05020005. https://doi.org/doi:10.1061/(ASCE)CP.1943-5487.0000945

[9] D.Q. Tran et al., Leveraging Semisupervised Learning for Domain Adaptation: Enhancing Safety at Construction Sites through Long-Tailed Object Detection, Journal of Construction Engineering and Management, 151, (1), (2025), p. 04024190. https://doi.org/doi:10.1061/JCEMD4.COENG-15259

[10] N.D. Nath et al., Deep learning for site safety: Real-time detection of personal protective equipment, Automation in Construction, 112, (2020), p. 103085. https://doi.org/10.1016/j.autcon.2020.103085

[11] Y. Pan et al., Roles of artificial intelligence in construction engineering and management: A critical review and future trends, Automation in Construction, 122, (2021), p. 103517. https://doi.org/10.1016/j.autcon.2020.103517

[12] X. Chen et al., Are Large Pre-trained Vision Language Models Effective Construction Safety Inspectors?, 2025, p. arXiv:2508.11011. https://doi.org/10.48550/arXiv.2508.11011

[13] A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, in: M. Marina, Z. Tong (Eds.), Proceedings of the 38th International Conference on Machine Learning, Vol. 139, PMLR, Proceedings of Machine Learning Research, 2021, pp. 8748--8763. https://proceedings.mlr.press/v139/radford21a.html

[14] Y.A. Malkov et al., Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs, IEEE transactions on pattern analysis and machine intelligence, 42, (4), (2020), pp. 824-836. https://doi.org/10.1109/TPAMI.2018.2889473

[15] A. Radford et al., Learning Transferable Visual Models From Natural Language Supervision, 2021, p. arXiv:2103.00020. https://doi.org/10.48550/arXiv.2103.00020

[16] J. Li et al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, in: K. Andreas, B. Emma, C. Kyunghyun, E. Barbara, S. Sivan, S. Jonathan (Eds.), Proceedings of the 40th International Conference on Machine Learning, Vol. 202, PMLR, Proceedings of Machine Learning Research, 2023, pp. 19730--19742. https://proceedings.mlr.press/v202/li23q.html

[17] J. Bai et al., Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023, p. arXiv:2308.12966. https://doi.org/10.48550/arXiv.2308.12966

[18] S. Yin et al., A survey on multimodal large language models, National Science Review, 11, (12), (2024). https://doi.org/10.1093/nsr/nwae403

[19] M. Yuksekgonul et al., When and why vision-language models behave like bags-of-words, and what to do about it?, 2022, p. arXiv:2210.01936. https://doi.org/10.48550/arXiv.2210.01936

[20] S. Bai et al., Qwen3-VL Technical Report, 2025, p. arXiv:2511.21631. https://doi.org/10.48550/arXiv.2511.21631

[21] D. Guo et al., Seed1.5-VL Technical Report, 2025, p. arXiv:2505.07062. https://doi.org/10.48550/arXiv.2505.07062

[22] Z. Jiang et al., VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks, 2024, p. arXiv:2410.05160. https://doi.org/10.48550/arXiv.2410.05160

[23] C. Chen et al., Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition, Automation in Construction, 146, (2023), p. 104702. https://doi.org/10.1016/j.autcon.2022.104702