Received: 17 July 2025 | Accepted: 27 October 2025

DOI: 10.1111/mice.70129

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

RESEARCH ARTICLE

Training-free few-shot construction tool and material detection using pre-trained vision-language model

Zhaoxin Zhang[1] | **Yantao Yu**[1,2] | **Zaolin Pan**[1] | **Maxwell Fordjour Antwi-Afari**[3]

[1]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

[2]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Shenzhen, China

[3]Department of Civil Engineering, College of Engineering and Physical Sciences, Aston University, Birmingham, UK

Correspondence
Yantao Yu, Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China.
Email: ceyantao@ust.hk

Funding information
National Natural Science Foundation of China, Grant/Award Number: 72201226; Research Grants Council, Grant/Award Numbers: 26208323, C6044-23GF

Abstract

Direct visual understanding of construction entities, such as tools and materials (T&M), underpin construction management and resource scheduling. Traditional supervised learning methods suffer from high annotation cost, severe computational demands, and limited datasets. In contrast, training-free approaches offer an effective alternative well-suited for construction scenarios constrained by data scarcity and limited resources. Besides, vision-language models (VLMs) can directly learn image semantics through natural language supervision and also demonstrate strong zero-shot detection capabilities without requiring retraining. Existing methods often exhibit limited image–text semantic alignment in construction scenarios, which restricts their effectiveness in construction tasks. Therefore, there is an urgent need for approaches that can enhance cross-modal understanding in such domain-specific contexts. To address this challenge, this paper proposes a training-free, knowledge-enhanced VLM to recognize T&M in construction tasks. The proposed approach leverages image matching and image–text knowledge alignment strategies, thereby utilizing the training-free nature of existing VLMs while benefiting from enhanced performance brought by knowledge integration. This method offers a novel solution for construction management and robotic collaboration tasks that are traditionally constrained by data and computational resource dependencies.

1 | INTRODUCTION

On dynamic and complex construction sites, various entities such as tools, materials, and equipment constitute the fundamental units supporting construction activities (K. Chen et al., 2025; Wu et al., 2025). Accurate recognition and understanding of these entities are not only critical for enabling digital perception of construction sites but also essential for facilitating intelligent construction management (Bavelos et al., 2025; Jiang et al., 2024). In subsequent project life-cycle stages, including resource scheduling, task allocation, and progress control, real-time monitoring and precise assessment of entity states are indispensable (Fei et al., 2024). Therefore, enhancing the perception and semantic understanding capabilities of visual systems for construction entities is of great significance to achieve efficient, safe, and sustainable automated construction.

Current deep learning technologies provide powerful tools for object detection in construction, autonomously extracting image features to advance visual

TABLE 1 Typical construction tasks.

Construction task	Tools and materials
Scaffolding erection (Pan & Yu, 2024)	Steel pipes, planks, fasteners
Structural assembly (Cai et al., 2024)	Hammers, wooden boards, wrenches
Mechanical construction (Jeoung et al., 2025)	Concrete formwork, mixer, transport vehicles

understanding (Mitterberger et al., 2022; M. Sun & Gao, 2022). For example, visual perception methods can enable construction robots to analyze and safely navigate the environments (T. Ren & Jebelli, 2025), perform collaborative assembly (Xi Wang et al., 2023) and object handover (Yu et al., 2023). However, several limitations remain. First, by relying on numerical labels instead of semantic text, traditional visual methods may fail to effectively capture construction-specific objects (Radford et al., 2021). Second, the dynamic nature of construction sites causes frequent changes in spatial and temporal contexts (Jeoung et al., 2025). Most public datasets focus on common objects like machinery and workers, lacking essential tools such as couplers and footboards for scaffold assembly, or tape measures for distance evaluation under uncertain conditions. Third, existing models not only require large annotated datasets, but also manual labeling is considered time-consuming, costly, and labor-intensive (Kim & Chi, 2017; Yanyu Wang et al., 2019). The wide variety of tools and materials (T&M) across construction project life-cycle phases further complicates dataset creation (Kim, 2020; K. Liu & Golparvar-Fard, 2015). The typical construction tasks are listed in Table 1. Therefore, developing vision algorithms that require fewer samples and a shorter training time is crucial for object detection in construction.

As vision and language typically convey complementary information, multimodal learning has proven highly effective across various construction tasks (Yong et al., 2023), such as visual question answering (Areerob et al., 2025; Kunlamai et al., 2024), image captioning (Chun et al., 2022), and image classification (Yong et al., 2023) for construction images. Unlike prior approaches, which independently trained visual and linguistic features on separate construction datasets, recent vision-language multimodal methods, such as contrastive language-image pre-training (CLIP; Radford et al., 2021), acquire generalized visual representations guided by natural language supervision, enabling impressive zero-shot detection capabilities (Cheng et al., 2024). Encoded visual features, through language-vision interaction, facilitate open-vocabulary recognition without the need for additional training (C. Chen et al., 2023), as demonstrated by zero-shot personal protective equipment (PPE) monitoring (Gil & Lee, 2024) and excavator activity recognition methods (C. Chen et al., 2023).

Despite significant progress made by large-scale pre-trained multimodal models in visual-language recognition (Fan et al., 2025), their semantic alignment between image and text remains insufficient in construction scenarios, leading to poor performance in direct applications such as object localization and recognition (Gil & Lee, 2024). To enhance model adaptability to construction tasks, previous studies have focused on fine-tuning models with few-shot data, such as construction report generation (Yoonhwa Jung et al., 2024) and automatic hazard identification (Yong et al., 2023). However, fine-tuning approaches pose significant challenges for construction scenarios that lack large-scale data and computational resources. To address these limitations, researchers have explored training-free detection methods to reduce the computational burden of adapting models to construction tasks (Liang et al., 2024). Although such methods perform well in classification tasks, effective application of object detection in construction scenarios is still lacking, as the problem of localizing targets in a training-free manner remains unexplored.

Existing vision-language models (VLMs) show promising zero-shot detection ability in general domains, but their application to construction remains limited. Their reliance on Internet-based data often leads to semantic mismatches in construction contexts, and their predictions may be ambiguous. Moreover, available datasets suffer from imbalance and weak textual annotations, while fine-tuning large-scale models with limited domain data are unstable and costly.

To address these gaps, this study contributes a training-free framework for construction object recognition. It aims to achieve training-free object localization and recognition in construction scenarios by integrating domain-specific image and text knowledge, resulting in a multimodal detection approach referred to as training-free knowledge-enhanced automatic detection. Specifically, it proposes a training-free method based on a knowledge-enhanced VLM for the localization and recognition of construction T&M. This study has two primary contributions. First, it introduces construction T&M localization method based on template matching. By leveraging on its powerful feature extraction and similarity computation capabilities of existing self-supervised vision models, the method matches reference images of construction objects with real-world scene images to localize objects. This process not only requires no retraining of network parameters but also relies solely on pre-trained vision foundation models. Second, this study proposes a classification method for construction objects that enhances the alignment between visual and textual semantics by constructing a similarity

distribution matrix enriched with domain knowledge. This facilitates the alignment of image–text representations for construction equipment. The results could demonstrate the potential of training-free VLMs for object recognition on construction sites. The proposed approach offers a promising solution to recognize construction objects using only open-source textual prompts, without incurring the computational cost of additional training.

The remainder of this paper is structured as follows. Section 2 discusses current methodologies for identifying and classifying objects in construction applications, as well as vision-language foundation models while identifying research gaps and objectives. Section 3 details the proposed method. Section 4 presents a comprehensive experimental evaluation on both a self-constructed construction T&M dataset and a publicly available construction machinery dataset. Section 5 discusses the benefits of the proposed approach and presents the visualization results. Section 6 outlines the limitations, while Section 7 concludes this study.

2 | LITERATURE REVIEW

The development of multimodal models has transformed traditional approaches of pure visual methods to understand construction site images and pure textual methods to interpret construction text data. Here, this research delves into the following three areas.

2.1 | Detection and classification of objects in construction

In dynamic and complex construction environments, enhancing safety, improving progress monitoring, and optimizing resource management are essential, making the application of deep learning-based visual object detection on construction sites a rapidly growing research area. Numerous studies have focused on safety detection, particularly PPE compliance (Y. Zhang et al., 2025) and automatic worker trajectory detection (Cai et al., 2020). Some studies have also used drone imagery to extend the sensing range and detect hazardous areas from an aerial perspective (Bang et al., 2021). In the area of progress monitoring, research has been demonstrated on detecting construction components or materials and correlating their locations with planned schedules (Karim & Adeli, 1999a, 1999b) or building information modeling (BIM) data (Z. Wang et al., 2021). Practical pipelines have been developed for component detection, BIM registration, and progress quantification, with some systems enabling repeated measurements for temporal tracking (Deng et al., 2020; Pour Rahimian et al., 2020). In resource management, some studies have explored counting and dimensioning construction materials for on-site images (Li & Chen, 2022; Shen et al., 2019). While these studies have demonstrated the potential of computer vision in construction automation, challenges remain due to the limited training samples in traditional data-driven algorithms (J. Chen et al., 2024; Liang et al., 2024). Table 2 presents the dataset scale required for applying deep learning algorithms in construction scenarios. To enhance detection effectiveness, existing algorithms rely on extensive annotated data for training. However, object annotation in construction site images is a demanding, labor-intensive, and time-consuming task. Consequently, limited training images can significantly constrain the performance of deep learning-based computer vision algorithms.

2.2 | Vision foundation models in construction

In recent years, large-scale VLMs, represented by CLIP, have garnered widespread attention due to their outstanding zero-shot performance in image classification tasks. Pre-training on large datasets, followed by fine-tuning for downstream tasks, has further stimulated researchers' exploration of their applications in computer vision, such as image recognition (Liang et al., 2024) and image captioning (Yoonhwa Jung et al., 2024). Table 2 presents applications of VLMs for detection and management in construction scenarios. Existing algorithms enhance detection performance by fine-tuning the models. However, the extensive parameter updates required substantially limit their direct applicability in construction contexts.

In the area of image recognition, existing studies have leveraged the zero-shot capabilities of multimodal VLMs to significantly reduce reliance on labeled datasets. For example, detection-based methods have enabled automated PPE safety monitoring on construction sites (Xiyu Wang & El-Gohary, 2024), and CLIP-based models have been applied for temporary object classification in construction scenarios (Liang et al., 2024). In construction site caption generation, current methods have mainly focused on identifying and describing construction safety hazards (Yiheng Wang et al., 2024) and automatically generating textual reports such as construction progress (Jeoung et al., 2025) and work logs (Yoonhwa Jung et al., 2024). Although multimodal models have exhibited strong transferability and fine-tuning performance in general vision tasks, they still face limitations in construction scenarios. Specifically, construction sites feature numerous domain-specific entities and semantic concepts, such as specific tool types,

TABLE 2 Summary of the current applications across different domains of the construction industry.

Problem statement	Dataset scale	Use of vision-language model	Few-/zero-shot	Training required
Personal protective equipment detection (Kumar et al., 2022)	6 categories, 14,500 images	No	No	Yes
Construction equipment recognition (Kim & Chi, 2021)	2920 images	No	Few-shot	Yes
Worker and equipment detection (Gil & Lee, 2024)	100 images	Yes	Zero-shot	No
Construction site safety monitoring (Chan et al., 2025)	1,500 images	Yes	No	Yes
Temporary construction object recognition (Liang et al., 2024)	22 categories, 90 images	Yes	Few-shot	No

structural components, and vehicles, which are underrepresented in general pre-training corpora, making it difficult for models to accurately recognize and understand them (Jeoung et al., 2025). As a result, the cross-modal alignment between image and text has become a key bottleneck limiting its performance in multimodal tasks.

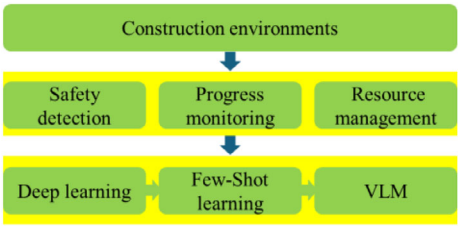


FIGURE 1 Summary of the literature review.

2.3 | Few-shot learning in construction

To address the issue of limited training samples, few-shot learning has been introduced in construction scenarios. For example, in tasks such as building defect detection (Zhao & Jebelli, 2025) and structural damage recognition (H. Sun et al., 2025), few-shot methods effectively alleviate the challenges of sample collection. In response to emerging needs for equipment recognition on construction sites, some studies have introduced few-shot construction equipment detection methods (Kim & Chi, 2021). Similarly, in construction worker safety management, few-shot learning has been used for the rapid identification of object attributes on construction sites (Xiyu Wang & El-Gohary, 2024). However, early research in few-shot object detection primarily has relied on extracting meta-knowledge from foundational datasets, a strategy that faces significant challenges in construction scenarios. Since existing studies often use construction-domain images as meta-knowledge for training, large annotated public datasets are rarely employed for meta-knowledge extraction (Kim & Chi, 2021). On the one hand, this raises the issue of cross-domain meta-learning, and there is currently a lack of systematic understanding of how to construct datasets best suited for construction tasks (Liang et al., 2024). On the other hand, the collection and annotation of construction foundational datasets are costly and often fail to comprehensively cover the complexity and diversity of construction sites, limiting the effectiveness of few-shot methods in real-world detection tasks. Figure 1 provides an overview of existing methodological advances in construction research.

In recent years, with the development of vision-language foundation models, researchers have proposed a novel few-shot learning paradigm based on the "pre-train adapter" model, enhancing few-shot performance by incorporating domain-specific knowledge. For example, Tip-Adapter can combine CLIP's zero-shot predictions with visual caching to improve the adaptability of few-shot classification (R. Zhang et al., 2022). Also, Tip-X introduces a framework for aligning image and text representations across both general and domain-specific datasets to further enhance few-shot recognition performance (Udandarao et al., 2023). Moreover, Adaptive Prior rEfinement can use a prior refinement module to selectively optimize critical feature channels for better few-shot prediction (Zhu et al., 2023). Building on these approaches, recent research has introduced training-free image classification methods, achieving a preliminary visual understanding of on-site temporary items (Liang et al., 2024). However, current research mainly focuses on the pattern of similarity measures between overall construction image–text label pairs, with a lack of effective modeling methods to match local feature regions and textual knowledge in construction scenarios. This limitation restricts the higher-precision detection and recognition capabilities of few-shot learning methods in construction contexts.

2.4 | Research gaps and objectives

By reviewing the extant literature, this study has identified the following research gaps and research objectives.

2.4.1 | Research gaps

In the construction domain, research on training-free construction object recognition based on multimodal models remains limited. Automated cross-modal alignment between construction objects and construction semantics still faces the following challenges.

Challenge 1: Construction sites involve several specialized terminologies while existing VLMs are primarily trained on general Internet-based text and images. This may limit their ability to understand text semantics in construction contexts, leading to semantic discrepancies when interpreting construction-related images and text. Although multimodal object detection models can demonstrate impressive zero-shot class-agnostic detection performance across various tasks, they tend to predict multiple ambiguous outputs, making it difficult to select the most appropriate detection result as an outcome. This raises a critical issue: *How can a more effective knowledge alignment mechanism between the construction domain and general domains be established to improve detection accuracy in construction environments?*

Challenge 2: Existing datasets in construction scenarios often suffer from imbalanced data distributions, ambiguous category definitions, and a lack of semantic textual knowledge, all of which hinder the generalization capability of models and limit their robustness in complex and dynamic environments. In practical applications, domain adaptation is often unavoidable. However, fine-tuning large-scale pre-trained models using only a small amount of domain-specific data tends to be unstable, prone to overfitting, and computationally expensive. Thus, an identified research question was: *how can the best for both worlds be achieved, by leveraging the training-free property of VLMs for object detection, while also utilizing construction text-image alignment knowledge to ensure the strong performance of few-shot training methods?*

2.4.2 | Research objectives

Based on the identified research gaps, this study has two main objectives: (1) Detection of unseen construction images using self-supervised models and image matching algorithms. To achieve this objective, this study utilizes a self-supervised model to encode domain-specific template images and construction images to be detected. By calculating the feature similarity of the object regions between the two images, the model performs feature matching between images, thus identifying the optimal regions in the object image for detection. (2) Construction object recognition through a training-free image–text similarity method to avoid model overfitting. To achieve this objective, the study leverages category-specific textual information from construction scenes and computes a similarity matrix between image and text features. This non-parametric, training-free approach enables effective alignment of multimodal features between construction object images and domain-specific textual knowledge.

3 | METHODOLOGY

Training-free construction object detection faces two primary challenges. First, the large search space between local features in construction images and actual objects makes accurate localization difficult. Second, the limited cross-modal alignment knowledge results in significant semantic ambiguity between images and text, thereby affecting detection accuracy. To address these challenges, this paper proposes a novel training-free construction object detection framework that leverages similarity algorithms to reduce semantic ambiguity between construction images and text. First, Section 3.1 provides a brief overview of the overall methodology. Then, Section 3.2 introduces a local feature localization strategy for construction objects to effectively narrow the search space. Finally, Section 3.3 details a cross-modal alignment method between construction images and construction-related text to mitigate semantic ambiguity and improve detection accuracy.

3.1 | Training-free knowledge-enhanced algorithm overview

The training-free detection framework follows a two-stage object detection paradigm (S. Ren et al., 2017). As illustrated in Figure 2, the first stage, object localization, aims to identify and delineate foreground objects (construction tools and materials) within the image. The second stage, object classification, assigns semantic labels to the localized objects based on textual category information. In the localization stage, the proposed method narrows the search space of potential targets across the image, thereby reducing redundant information and facilitating zero-shot object recognition. In the classification stage, the method computes cross-modal distributional features to

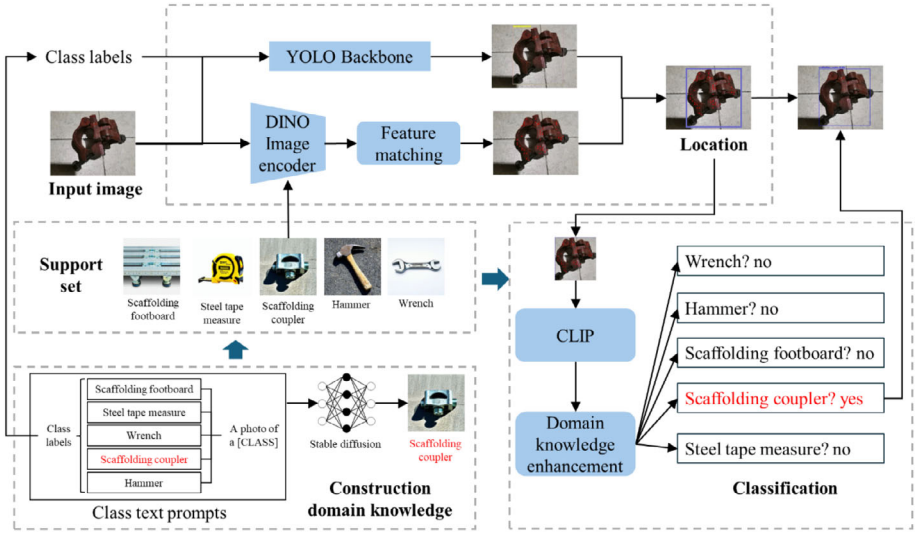


FIGURE 2 Training-free knowledge-enhanced object detection method.

enhance semantic alignment between construction images and textual descriptions.

3.2 | Construction T&M localization method

This method first identifies the object locations of construction objects within the input image. To achieve this, this study employs a feature matching strategy that localizes objects based on correspondences between template images and the input image. The approach consists of two components, namely, feature extraction and feature matching.

3.2.1 | Feature extraction

Self-supervised feature extraction could reduce dependence on labeled data and offer a potential approach for identifying features of similar objects (Huang et al., 2024). To determine the object locations within the input image, this study first provides a template image and a candidate construction T&M image, where the object and background regions in the template image are predefined. In this study, DINOv2 (Oquab et al., 2024) was employed to extract visual features from the template image x_r and the input image x_t. There are two main reasons: (1) DINOv2 is a purely vision-based self-supervised learning model trained on large-scale image datasets; and (2) DINOv2 learns features by decomposing images into patches, enabling precise patch-level matching across different images. The image encoder, parameterized by the self-supervised network, computes the output features, resulting in feature maps f_r and f_t for the template and input images, respectively, where $f_r, f_t \in \mathbb{R}^{H\times W\times C}$. In the approach, foreground and background regions in template images are delineated manually. For each template, the foreground is annotated to include only the target objects, whereas the background comprises all pixels outside the targets. As illustrated in Figure 3a, the method extracts feature maps from the input image at this stage. With DINOv2, the input image size is 518×518 and each patch measures 14×14, resulting in a feature map of 37×37.

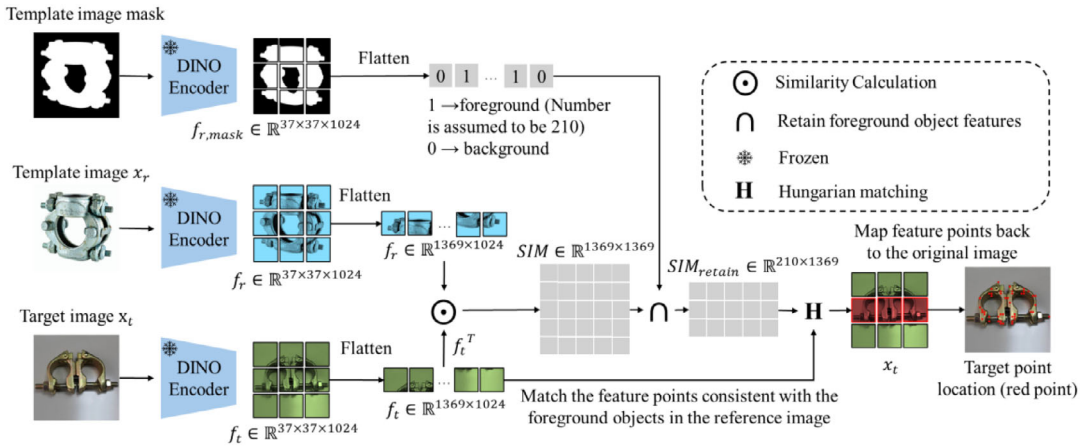
3.2.2 | Feature matching

In this section, the study first locates candidate foreground objects through feature matching. As shown in Figure 3a, the feature map is flattened, and a similarity matrix is computed between the reference image and the candidate image features. The similarity matrix was computed as

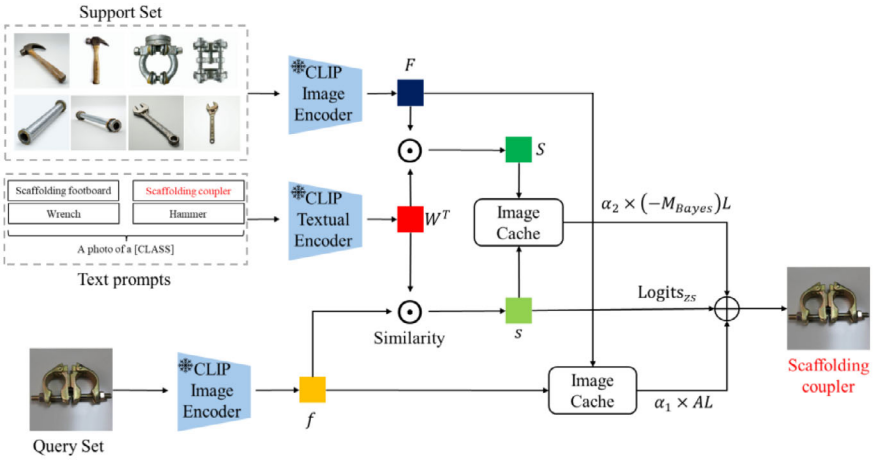
$$SIM = F(x_r) \cdot F(x_t)^T \quad (1)$$

where SIM represents the feature similarity matrix between the template image and the detection image. F denotes the pre-trained image encoder. $F(x_r) = f_r, F(x_t) = f_t$.

As shown in Figure 3a, based on the mask features of the reference image, the foreground similarity matrix SIM_{retain} is preserved. To match the features between the template and candidate images, the Hungarian algorithm (Dong et al., 2024), a mathematical method for solving the linear assignment problem aimed at minimizing the total assignment cost, was employed and has widely been used in feature matching tasks (Atik et al., 2025; Y. Liu et al.,



(a) Domain knowledge fusion-based construction T&M localization method.



(b) Construction T&M category classification method.

FIGURE 3 An illustration of the proposed training-free localization and classification method.

2023). This study formulates the similarity matrix as the cost matrix for the Hungarian algorithm, with the objective of identifying the optimal correspondence between the template and candidate images.

This study defines f_r^i as the ith feature point of the template image and f_t^i as the ith feature point of the candidate image. The output is the set of position coordinates $\mathcal{P} = \{(p_i^x, p_i^y)\}_{i=1}^N$ corresponding to the matched feature points within the object region of the candidate image, where N is the number of feature points associated with the object.

3.2.3 | Object location

After obtaining the foreground feature points from DINOv2 matching, we first compute their spatial distribution and count the number of matched points falling inside each bounding box predicted by YOLO-World. Specifically, the final target location information is determined by a post-processing strategy to filter bounding boxes by incorporating point annotations and overlap constraints. For each bounding box $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ obtained from YOLO-World, the method selects the set of points located inside it:

$$\mathcal{P}_b = \{(p^x, p^y) \in \mathcal{P} | x_{\min} \le p^x \le x_{\max}, y_{\min} \le p^y \le y_{\max}\} \quad (2)$$

where $\mathcal{P}$ denotes the set of all annotated points.

The procedure consists of three steps:

1. If the number of inside points does not satisfy: $|\mathcal{P}_b| \ge 8$. Then the bounding box b is discarded. Here, the method sets the number of inside points to 8. Since the minimum image resolution is preset to 224×224 and the patch size in DINOv2 is 14×14, the maximum number of inside points is 16×16. The method defines the threshold as half of the minimum number of points along one side, that is, half of 16.
2. If no bounding box remains after the previous step but there exist points $(p^x, p^y) \in \mathcal{P}$, the method generates a new bounding box that tightly encloses all points.

3. Finally, if two or more bounding boxes overlap with an $IoU > 0.9$, only the one with the largest area is retained to avoid redundant localization.

3.3 | T&M category classification

After the construction of T&M region localization in Section 3.2, this study aligns the semantic knowledge of images and texts using a multimodal image classification model. This study first provides a brief review of the CLIP multimodal model and its improvements and discusses the limitations of existing methods.

3.3.1 | Training-free zero-shot CLIP

CLIP uses contrastive learning to learn joint visual and textual representations, encouraging high cosine similarity between matched pairs. For C-class image classification, training-free CLIP transforms each class name into a descriptive text prompt and encodes it through its text encoder for comparison with image features. The prompt embeddings form the classifier weight matrix $W \in \mathbb{R}^{C\times d}$, where d represents the dimensionality of the text embeddings. Given a test set $T = \{y_1, y_2, \dots, y_t\}$ with t images, CLIP generates its visual feature representations via its image encoder:

$$f_i = ImageEncoder\left(y_i\right), i \in [1, t], f_i \in \mathbb{R}^d \tag{3}$$

$$f = Concat\left(\left[f_1, f_2, \dots, f_t\right]\right), f \in \mathbb{R}^{t\times d} \tag{4}$$

$$\text{Logits}_{zs} = fW^T \tag{5}$$

3.3.2 | TIP-adapter

Given a dataset $D = \{x_1, x_2, \dots, x_{CK}\}$ of size CK, the TIP-Adapter method utilizes a pre-trained CLIP encoder to encode the given dataset D as follows:

$$F_i = ImageEncoder\left(x_i\right), i \in [1, CK], F_i \in \mathbb{R}^d \tag{6}$$

$$F = Concat\left(\left[F_1, F_2, \dots, F_{CK}\right]\right), f \in \mathbb{R}^{CK\times d} \tag{7}$$

Then, TIP-Adapter constructs an affinity matrix to obtain the similarity matrix between F and f as follows:

$$A = \exp\left(-\beta\left(1 - fF^T\right)\right) \tag{8}$$

where β is a tunable hyperparameter.

Finally, the logits under zero-shot conditions and the computed similarity matrix A are jointly applied to the final mixed logits as follows:

$$L_{TIP-Adapter} = \alpha_1 AL + fW^T \tag{9}$$

where $L \in \mathbb{R}^{CK\times C}$ is the one-hot vector for each few-shot class label.

3.3.3 | TIP-X

TIP-X considers the modality gap between the image and text spaces of CLIP, calculating the relationship between the test image features $f \in \mathbb{R}^{t\times d}$ and the few-shot image features with the text space features as follows:

$$S = \text{softmax}\left(FW^T\right), S \in \mathbb{R}^{CK\times C} \tag{10}$$

$$s = \text{softmax}\left(fW^T\right), S \in \mathbb{R}^{t\times C} \tag{11}$$

Then, TIP-X constructs an affinity matrix $M_{KL} \in \mathbb{R}^{t\times CK}$, and uses the Kullback–Leibler (KL) divergence to compute the affinity matrix, as follows:

$$M_{i,j} = D_{KL}\left(s_i \| S_j\right), i \in [1, t], j \in [1, CK] \tag{12}$$

The final mixed logits calculation method is as follows:

$$L_{TIP-X} = fW^T + \alpha_1 AL + \alpha_2\left(-M_{KL}\right)L \tag{13}$$

3.3.4 | Knowledge-enhanced multimodal knowledge alignment

Both visual and textual inputs are projected by CLIP into a unified embedding space. For this study, the text prompt "A photo of a [object]" was employed, replacing "[object]" with construction-related class labels. CLIP demonstrates excellent zero-shot recognition performance on a variety of public datasets. However, the construction scene is highly specialized and often confidential, which results in poor zero-shot performance for CLIP, as it cannot directly access construction scene images.

In this section, a support set of images for image classification was built, where their image embeddings and class labels are stored as "key-value" pairs, as shown in Figure 3b. CLIP optimizes the cosine similarity between modalities through contrastive training. In this study, the authors draw inspiration from the TIP-X method, which introduces an image–text similarity cache based on object dataset text labels.

TIP-X has limitations when applied to recognize construction domain objects. The TIP-X method calculates

the distance between the image distribution and the text distribution using the KL divergence. However, KL divergence is primarily used to measure the difference between fixed distributions and neglects the characteristic of sparse data samples in complex construction scenes. It does not dynamically update the sample distribution of construction scenes, which can lead to incorrect estimation of the scene's sample distribution. To address the limitations of the TIP-X method, this research puts forward to dynamically update the prior distribution of samples using Bayes' theorem before calculating the distance between the image distribution and the text distribution (Yoon et al., 2018). This study uses a Dirichlet prior to compute the prior distribution, and the calculation process is as follows:

$$M_{i,j} = D_{KL}^{Bayes}\left(s_i \| S_j\right), i \in [1,t], j \in [1,CK] \tag{14}$$

$$D_{KL}^{Bayes}(P\|Q) = \sum_i P_i \log \frac{P_i}{Q_i} \tag{15}$$

$$\hat{P}_i = \frac{P_i + \alpha_3}{\sum_{i=1}^{t}(P_i + \alpha_3)} \tag{16}$$

where s_i represents the ith test signature for the t test samples, and S_j represents the jth few-shot signature. α_3 is the hyperparameter for Bayesian estimation. In the implementation, the value of α_3 was determined from a predefined candidate set ranging from 0 to 1 with a step size of 0.1, that is, $\alpha_3 \in \{0.1, 0.2, 0.3, \ldots, 1.0\}$.

The final mixed logits calculation is as follows:

$$\text{Logits} = \text{Logits}_{zs} + \alpha_1 \times AL + \alpha_2 \times \left(-M_{Bayes}\right) L \tag{17}$$

where M_{Bayes} is the affinity matrix computed using the Bayesian prior in this study. All three hyperparameters α_1, α_2, and α_3 were jointly optimized through a nested loop search procedure, that is, iterating over all possible combinations of α_1, α_2, and α_3 and selecting the combination that achieved the best performance.

4 | EXPERIMENTS AND RESULTS

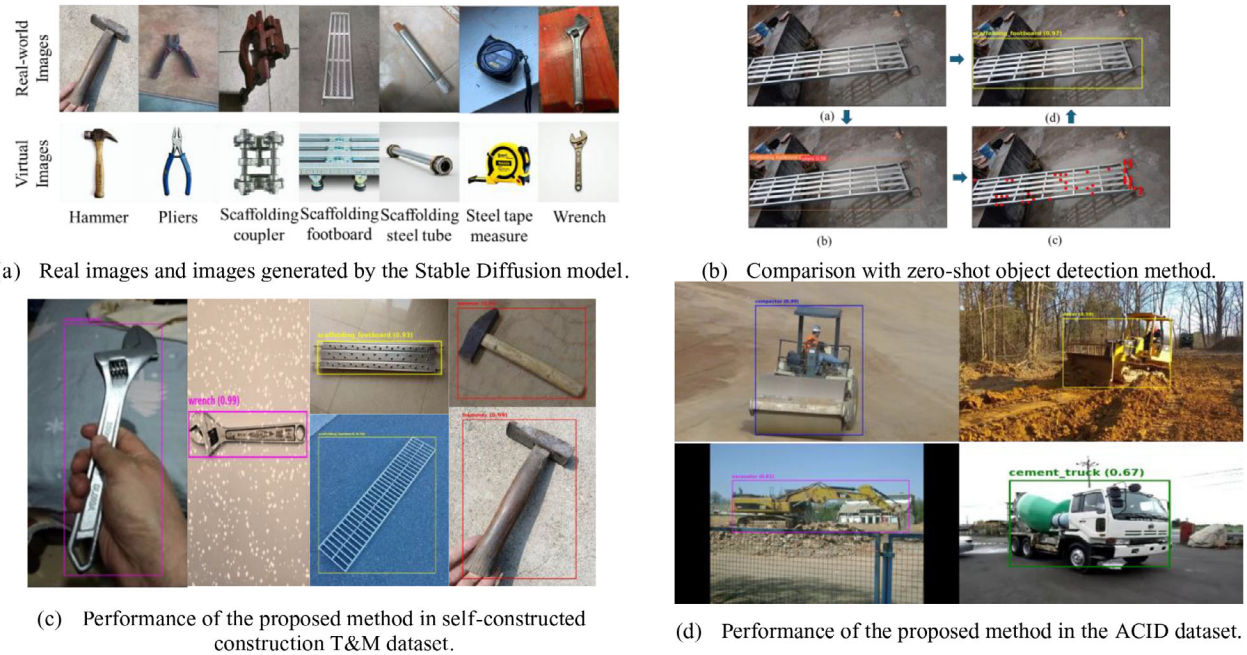
To evaluate the performance of the proposed knowledge-enhanced object detection method for construction T&M detection, experiments on a newly developed construction T&M dataset were conducted, and a public construction machinery object detection dataset (Xiao & Kang, 2021) was used. Section 4.1 details the self-constructed and public datasets. Section 4.2 presents a comparison between our method and existing zero-shot multimodal approaches. To further demonstrate the effectiveness of our method, Section 4.3 provides a comparison with other training-free methods. Section 4.4 reports the results of ablation studies. Section 4.5 reports the results of comparative experiments, and Section 4.6 presents the feature visualizations. All experiments were conducted on a local machine equipped with an AMD EPYC 9654 CPU and an NVIDIA GeForce RTX 4060 Ti GPU with 16 GB of VRAM. Notably, since the proposed method does not require backpropagation or parameter training, it can be effectively deployed on-site, including edge devices or mobile construction robots equipped with similar-level GPUs.

4.1 | Datasets and experiment

4.1.1 | Self-constructed construction T&M dataset

To validate the proposed method, this study develops a custom small-scale dataset tailored for evaluation purposes as shown in Figure 4a. The dataset was collected based on the T&M that workers frequently access during construction activities, aiming to support effective worker–robot collaboration. Specifically, we selected objects used in scaffolding assembly, which is considered one of the most representative temporary structural processes in construction (Younghan Jung et al., 2019). During such temporary structure assembly, training-free vision-language methods can provide rapid and practical solutions for worker–robot collaboration, where fast recognition and manipulation are required on-site. Accordingly, we included scaffolding couplers, scaffolding footboards, steel tape measures, and scaffolding steel tubes as core materials that robots can assist in transporting or handling. In addition, categories such as hammers and wrenches were not chosen for their generality but because they are integral components of scaffolding installation and structural assembly tasks, where workers frequently handle them alongside other construction-specific tools (e.g., couplers and footboards). Although these tools are commonly seen in other domains, recognizing them accurately in complex construction scenes is critical for ensuring safety and coordination. Therefore, they were also incorporated into the dataset to realistically represent few-shot challenges in construction-related perception tasks.

The dataset was collected from real-world images of construction tools and materials sourced from the Internet, and it is representative of visual recognition tasks relevant to robotic grasping in construction scenarios. The method collected 200 original images and expanded the dataset through data augmentation, generating a total of 1000 images that include conditions such as rain, snow, fog, and motion blur. It includes seven different object categories, with a total of 1000 images and 1115 labeled category annotations. Detailed information about the object categories



(a) Real images and images generated by the Stable Diffusion model.

(b) Comparison with zero-shot object detection method.

(c) Performance of the proposed method in self-constructed construction T&M dataset.

(d) Performance of the proposed method in the ACID dataset.

FIGURE 4 An illustration of the self-constructed construction T&M dataset and the recognition results.

TABLE 3 Number of labels for each category sample in self-constructed construction tools and materials (T&M) dataset.

Hammer	Pliers	Scaffolding coupler	Scaffolding footboard	Scaffolding steel tube	Steel tape measure	Wrench
100	155	140	155	130	125	310

can be found in Table 3. Compared to other construction object detection datasets, this dataset provides fewer images; however, it represents categories that are rarely collected and annotated in construction scenes and are essential auxiliary T&M for worker–robot collaboration during scaffold installation. This dataset is primarily used for validating the construction T&M recognition method.

In construction scenarios, image collection and annotation require professional expertise, whereas textual knowledge is comparatively easier to obtain. In this study, the only knowledge available was the categories of textual information. Construction domain images are required as prior knowledge for construction T&M recognition. Therefore, this study utilizes images generated by a text-to-image model (Rombach et al., 2022) as the domain knowledge for construction images. The intention was not to rely on large-scale synthetic data to replace the training-free nature of the framework but rather to use synthetic images as auxiliary references in scenarios where real template images are insufficient or unavailable. For this reason, the collected and labeled dataset was exclusively used for model validation and was not involved in the construction of domain image knowledge. To better utilize Stable Diffusion for image generation, Chat Generative Pre-trained Transformer (ChatGPT) (OpenAI, 2023) was employed to generate prompts for the text-to-image model. This study inputs the object category textual labels required for image generation into ChatGPT with the following prompt: "Help me generate a prompt to generate a [object]," where [object] represents the object category textual label. In the localization stage, this study employs text-generated synthetic images as visual knowledge. In the classification stage, text–image pairs composed of textual descriptions and generated images are used as image–text knowledge. Table 4 presents the generated category prompts. The image generation results from the Stable Diffusion model are shown in Figure 4a. In the object localization stage, the method selected images with minimal occlusion and clear views of the target objects to ensure reliable feature extraction. For each tool or material category, one representative template image was chosen as the reference for matching.

4.1.2 | Alberta Construction Image Data (ACID) dataset

The ACID construction machinery detection dataset (Xiao & Kang, 2021) was employed in this study to rigorously evaluate the effectiveness of the proposed method. The

TABLE 4 Examples of generated prompts.

Class	Hammer	Pliers	Scaffolding coupler	Scaffolding footboard	Scaffolding steel tube	Steel tape measure	Wrench
The prompts for stable diffusion model	High-resolution realistic photo of [Class], placed on a clean background, sharp details, professional product photo, natural lighting, and minimal shadows						

TABLE 5 Label count for each category in the Alberta Construction Image Data (ACID) dataset.

Cement truck	Compactor	Dozer	Excavator	Wheel loader
606	735	751	1844	1336

TABLE 6 Computation time for different methods.

	T&M dataset		ACID dataset	
	YOLO-World	Ours	YOLO-World	Ours
Computation efficiency	17.86 fps	2.60 fps	18.65 fps	2.58 fps

original ACID dataset contains 10,000 images of construction machinery, covering 10 different types of machinery categories. Considering the input requirement of the CLIP image encoder, which is 224 × 224 pixels, images smaller than this threshold were excluded during the preprocessing of the dataset to ensure proper input to the model. Additionally, due to computational resource limitations, only five categories from the dataset were selected for model validation. These preprocessing steps resulted in a customized version of the ACID dataset for object detection. As shown in Table 5, the dataset consists of 3981 images and 5272 category labels, with the following distribution: 606 cement trucks, 735 compactors, 751 dozers, 1844 excavators, and 1336 wheel loaders.

4.2 | Comparison with multimodal object detection methods

YOLO-World (Cheng et al., 2024) vision-language multimodal object detection is a method that heavily relies on pre-trained weights for object recognition. In contrast to the zero-shot knowledge-enhanced approach, existing multimodal detection methods lack any prior knowledge specific to the construction domain (Romera-Paredes & Torr, 2015), making them zero-shot object detection methods.

4.2.1 | Performance on T&M dataset

Using the dataset provided in this study, this study compared the performance of the proposed method in a 4-shot scenario with YOLO-World. YOLO-World identifies objects based on text prompts, where the text prompts are the category labels shown in Table 3. The model's performance was measured using the mean average precision (mAP) metric (Everingham et al., 2010), which is among the most commonly utilized metrics in vision-based object detection tasks.

Although Table 6 shows that YOLO-World achieves faster zero-shot detection than our method, there remains an urgent need to improve detection accuracy in construction site scenarios. As shown in Table 7, our method shows a significant improvement in the accuracy of object detection, compared to zero-shot detection methods. YOLO-World's zero-shot detection performance was limited, with an overall mAP of only 20.38%. The proposed method achieved an overall mAP of 59.45%. For example, the Scaffolding Coupler, an essential T&M for workers during scaffolding installation, is effectively recognized by the proposed method. In contrast, existing zero-shot methods struggled to recognize this tool, as current vision-language multimodal methods are unable to handle domains with high domain-specific knowledge. As illustrated in Figure 5d,g, the positional correspondence between the reference and test images in Figure 5d determines the object's location within the field of view, while the image–text similarity comparison shown in Figure 5g enables semantic classification of the detected object. This highlights the serious challenges existing multimodal methods face when processing T&M images from construction scenarios.

Directly applying zero-shot multimodal models to object detection in construction scenarios often fails to recognize uncommon components such as scaffold couplers as illustrated in Figure 5a. This is primarily due to the semantic misalignment between visual content and textual knowledge (Jeoung et al., 2025). In contrast, the method proposed in this paper successfully identifies such components, demonstrating their ability to enhance cross-modal alignment under a training-free paradigm. This capability can significantly improve the efficiency of worker–robot collaboration in construction environments.

As illustrated in Figure 4b, subfigure (a) shows the original input image, while subfigure (b) presents the output of

TABLE 7 Mean average precision (mAP) in self-constructed construction T&M dataset.

	Hammer	Pliers	Scaffolding coupler	Scaffolding footboard	Steel tape measure	Scaffolding steel tube	Wrench	All mAP
YOLO-World	10.09	15.86	1.00	65.44	25.40	6.20	18.67	20.38
3.2+CLIP (contrastive language-image pre-training)	41.56	30.38	34.04	71.55	68.15	74.55	31.96	50.31
3.2+TIP-Adapter	42.56	47.25	50.24	63.60	71.77	79.30	55.83	58.65
3.2+TIP-X	37.58	48.74	52.12	66.92	65.54	80.89	42.60	56.34
Ours	44.56	43.58	54.35	72.14	67.34	80.92	53.25	59.45

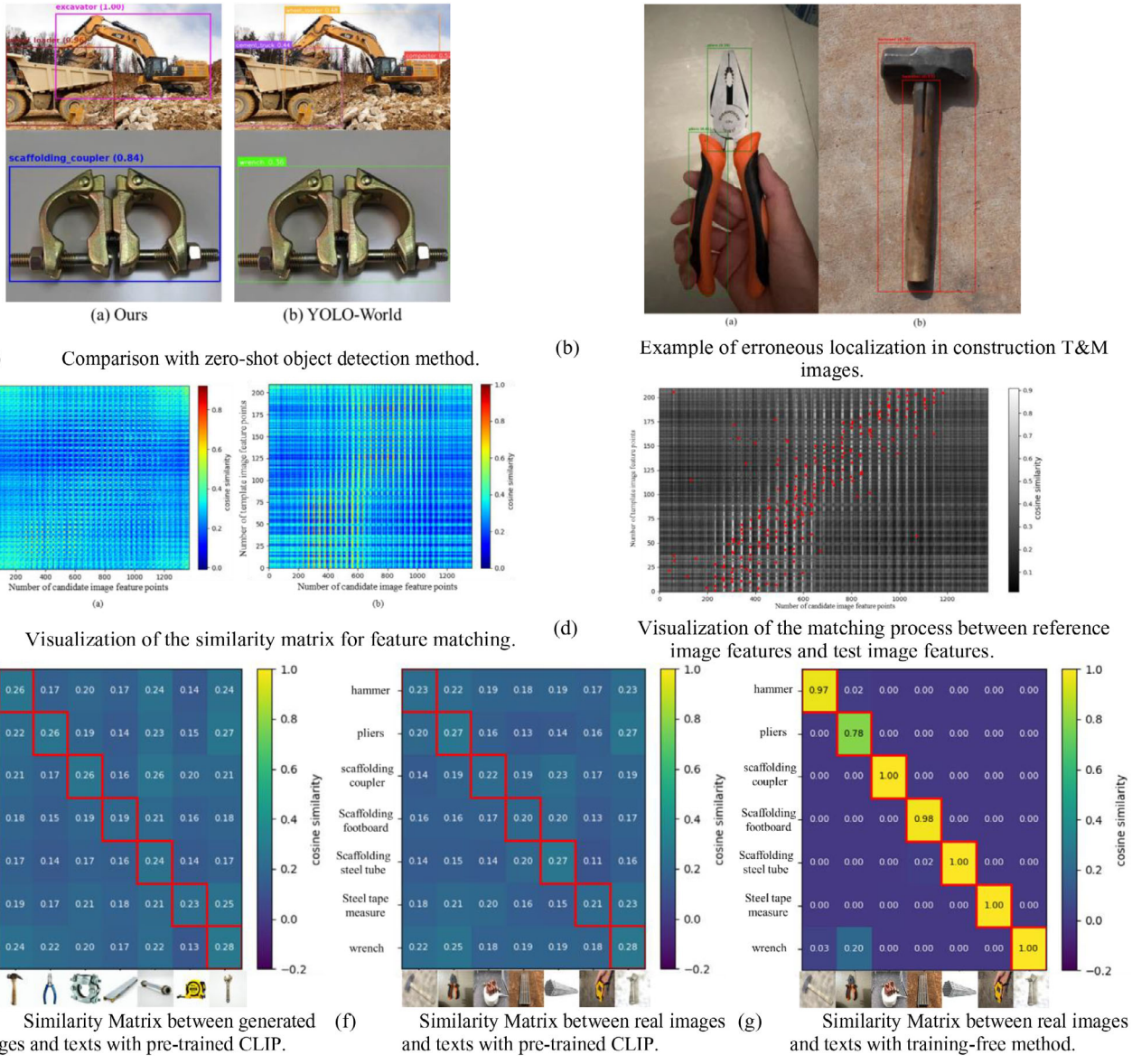


FIGURE 5 An illustration of the recognition results.

a zero-shot object detection model. The pre-trained multimodal model detects two categories, namely, scaffold planks (correct) and pliers (incorrect). If using localization with only YOLO-World, objects that should not be present (such as the pliers in the figure) may appear erroneously. This false positive is primarily attributed to the lack of domain-specific knowledge related to construction sites in the pre-trained model. After applying the

TABLE 8 mAP in the ACID dataset.

	Cement truck	Compactor	Dozer	Excavator	Wheel loader	All mAP
YOLO-World	66.29	15.51	0.26	47.59	25.87	31.11
3.2+CLIP	82.29	0.97	35.23	69.07	55.20	48.55
3.2+TIP-Adapter	81.86	3.31	56.95	69.14	53.41	52.93
3.2+TIP-X	74.16	16.99	61.95	64.25	59.72	55.41
Ours	78.19	59.29	45.08	67.59	34.44	56.92

proposed training-free method, as shown in subfigure (c), the model effectively eliminates the incorrect detection of pliers via an image matching mechanism. This not only reduces computational cost and prevents overfitting of network parameters but also removes the need for manual data annotation. These results demonstrate that the proposed method significantly improves semantic alignment between visual and textual modalities and offers a practical solution for deploying multimodal models in construction safety management and worker–robot collaborative grasping tasks.

Experimental results suggested that general-purpose multimodal foundation models with zero-shot capabilities struggle in domains requiring specialized knowledge, such as the construction industry. The proposed training-free object detection framework requires neither labeled training samples nor network parameter updates. It thus avoids overfitting risks and high computational demands commonly associated with deep learning under limited data. With these advantages, the method is well-suited for real-world construction environments where both familiar and novel objects may coexist.

4.2.2 | Performance on ACID dataset

Based on the ACID dataset, this study compared the detection results of our method in a 16-shot scenario with those of YOLO-World. YOLO-World performs object recognition based on textual prompts, which are the category labels shown in Table 5. From Table 8, it was evident that YOLO-World's zero-shot object detection performance was limited, with the mAP of only 31.11% across all categories. In contrast, the mAP of the proposed method across all categories was 56.92%. For example, the detection performance for the "Dozer" category was limited with existing multimodal methods, possibly due to feature discrepancies between the images and textual information of the "Dozer" category. Overall, the detection results objectively highlighted the imbalance in inter-class modal distributions within the YOLO-World multimodal method, leading to significant variations in detection performance across categories. This imbalance is unfavorable for the widespread application of such methods in future construction scenarios.

The visualization results are shown in Figure 5a. Due to YOLO-World's biased understanding of the image and text features of construction machinery, significant misidentifications and redundant object detections were observed (e.g., the "Compactor" category in the upper-right image). By applying the domain knowledge enhancement and image–text knowledge alignment methods proposed in this paper, the model can confidently and accurately recognize construction machinery objects.

4.3 | Comparison with other training-free methods

The method proposed in this paper consists of two stages. The first stage (Section 3.2) is based on image matching for object localization, while the second stage (Section 3.3) is based on knowledge alignment for object semantic recognition. To fully evaluate the effectiveness of the proposed method, this section uses three training-free methods (Radford et al., 2021; Udandarao et al., 2023; R. Zhang et al., 2022) as the baseline models for the second stage, which facilitates a comprehensive assessment of the proposed approach.

The experimental results, as shown in Tables 7 and 8, reveal that CLIP's zero-shot classification performance is limited. The mAP across all categories in the construction T&M detection dataset is 50.31%, and the mAP across all categories in the ACID dataset is 48.55%. Although the overall recognition accuracy of the CLIP method is relatively low, it maintains higher accuracy for certain categories. For instance, in the construction T&M detection dataset, the categories "Steel tape measure" and "Scaffolding steel tube" exhibit relatively high accuracy. In the ACID dataset, the "Cement truck" category also has higher recognition accuracy. This phenomenon is likely due to the higher similarity in image–text distribution for these categories in the pre-trained CLIP model.

It is worth noting that the results of Section 3.2 + CLIP zero-shot classification show a significant improvement over the YOLO-World zero-shot detection method in terms

TABLE 9 Optimal hyperparameter settings for different shots in self-constructed construction T&M dataset.

	β	α_1	α_2	α_3
Search scale	[0, 50]	[0, 50]	[0, 30]	[0, 1]
1-shot	3.695	15.07	12.06	0.1
2-shot	1.0	22.555	12.06	0.2
4-shot	3.45	7.585	7.874	0.9

of detection accuracy. This indicates that multimodal object detection presents greater research challenges compared to multimodal image classification in the context of foundational model research, highlighting the necessity and urgency of the approach proposed in Sec. 3.2. Last, while Tip-Adapter and TIP-X show significant improvements over CLIP zero-shot, the algorithm presented in this paper outperforms all the comparative methods.

4.4 | Ablation study

The proposed construction T&M classification method involves three hyperparameters, namely, β in Equation (8), α_1 and α_2 in Equation (13), and α_3 in Equation (16). Table 9 summarizes the hyperparameter values that yielded the highest accuracy during grid search on the validation dataset.

Table 10 summarizes the experimental results across various few-shot detection settings ($k = 1$, 2, and 4). Notably, the model attained a mAP of 59.45% and successfully identified all construction T&M in the 4-shot scenario. Additionally, as the number of images increased, the mAP also improved. Although the results are promising, some counterintuitive outcomes were also observed. For example, the recognition accuracy for the "Steel tape measure" category in the 4-shot scenario was lower than that in the 1-shot scenario. In the few-shot setting, sample templates contribute unevenly. Multiple samples from the same category may vary in representativeness due to occlusion or background clutter, which can amplify noise and reduce class discriminability, ultimately leading to performance degradation, where increasing the number of samples does not yield a positive impact on model performance.

Through grid search on the validation set, the hyperparameter configuration yielding maximum accuracy is listed in Table 11. Table 12 details the experimental outcomes for few-shot detection with k values of 1, 2, 4, and 16. Under the 16-shot condition, the model detected all construction machinery, reaching a mAP of 56.92%. Additionally, as the number of images increased, the mAP also increased. Although the results are promising, some counterintuitive outcomes were observed. For example, the recognition accuracy for the "Cement truck" category in the 1-shot scenario was higher than in other shot scenarios. This phenomenon could be attributed to the marginal effects caused by the modality gap between images and text, where the increase in samples did not have a positive impact on the images of this category.

To highlight the effectiveness of the second-stage classification, the method conducted experiments on two datasets, comparing the method with a classification-only baseline (with labeled object locations) using the mAP metric. Specifically, we employed different localization methods to analyze their impact on the final detection accuracy, including the proposed localization using DINOv2 and feature matching (DINOv2+cls), localization using only YOLO-World (Y-W-loc+cls), and localization using ground-truth bounding boxes (Gt-loc+cls). As shown in Tables 10 and 12, different localization strategies lead to varying detection performances, and the experimental results further confirm the improvement effect of fusing YOLO-World on localization accuracy. The results confirm that the proposed training-free two-stage framework is influenced by both localization and classification, suggesting that future studies should investigate strategies to improve each stage independently.

4.5 | Comparative evaluation with few-shot fine-tuning method

To further highlight the advantages of the proposed method, the method additionally benchmarked our training-free framework against the few-shot fine-tuning based object detection method Few-Shot object detection via Contrastive proposals Encoding (FSCE) (B. Sun et al., 2021). The implementation of FSCE was reproduced using the mmfewshot module from OpenMMLab under the same experimental settings.

The results, summarized in Tables 13 and 14, show that the method consistently outperforms the fine-tuning-based FSCE across different shot settings. Notably, while fine-tuning approaches require additional training time and are prone to overfitting due to the limited availability of construction-domain data, our training-free approach leverages domain-specific visual-textual alignment without retraining, thereby achieving higher accuracy with improved efficiency.

4.6 | Visualization of features

To make the intermediate steps of our method more transparent, the method provides visualizations for both localization and classification. Figure 5c,d illustrates the feature matching process: after encoding with DINOv2,

TABLE 10 Ablation experimental results of the proposed method in self-constructed construction T&M dataset.

		Hammer	Pliers	Scaffolding coupler	Scaffolding footboard	Steel tape measure	Scaffolding steel tube	Wrench	All mAP
1-shot	Y-W-loc+cls	21.38	18.06	43.02	39.13	25.98	27.65	16.90	27.45
	DINO v2+cls	46.39	37.19	51.98	61.33	69.36	80.08	44.79	55.87
	Gt-loc+cls	54.56	47.25	70.28	70.61	91.94	91.86	40.60	66.73
2-shot	Y-W-loc+cls	22.46	21.64	33.48	42.92	24.40	27.68	17.97	27.22
	DINO v2+cls	51.91	39.66	41.60	77.30	67.81	79.92	45.88	57.73
	Gt-loc+cls	52.39	48.33	54.43	81.03	89.66	94.52	53.19	67.65
4-shot	Y-W-loc+cls	19.40	27.10	39.34	40.27	22.41	28.34	14.61	27.35
	DINO v2+cls	44.56	43.58	54.35	72.14	67.34	80.92	53.25	59.45
	Gt-loc+cls	41.21	66.05	69.81	70.60	88.67	94.30	47.63	68.33

Note: "Y-W-loc+cls" denotes the detection performance based on YOLO-World localization method; "DINOv2+cls" denotes the detection performance based on the proposed localization method; and "Gt-loc+cls" denotes the detection performance under ground-truth localization, where only bounding boxes are provided without class labels.

TABLE 11 Optimal hyperparameter settings for different shots in the ACID dataset.

	β	α_1	α_2	α_3
Search scale	[0, 50]	[0, 50]	[0, 30]	[0, 1]
1-shot	0.1	5.41	5.48	0.9
2-shot	16.68	2.595	27.6	0.4
4-shot	12.27	10.08	29.402	0.4
16-shot	6.145	7.58	17.442	0.4

similarity matrices are computed and refined via Hungarian matching. To clearly highlight correspondences, the matrix is displayed in grayscale with red dots marking the matched feature pairs, which correspond to high-similarity regions and can be projected back to the original image to locate features.

Figure 5e–g compares image–text similarity matrices in the classification stage. With CLIP encoding, generated images and text (Figure 5e) as well as real images and text (Figure 5f) show nearly indistinguishable category scores, often leading to errors. In contrast, our training-free method (Figure 5g) produces sharper alignment between images and corresponding text, resulting in higher classification accuracy.

5 | DISCUSSION

The study found that the training-free knowledge-enhanced method for construction T&M detection can leverage multimodal foundation models for few-shot learning in the construction domain. First, the method allows learning the feature knowledge of construction T&M images using template images from specific classes, which aids in accurately localizing construction T&M within images. Second, the method can effectively distinguish the categorical attributes of construction T&M after being provided with some image and textual knowledge. The detection outcomes of the developed model are presented in Figure 4c,d. These results demonstrate the model's strong performance even when construction T&M vary in shape and color. Figure 4c shows that the method maintains robust tool detection across varying viewpoints and rotations. These findings suggest that the proposed method effectively utilizes pre-trained foundation models and domain-specific knowledge to represent construction scenes, without requiring the training or optimization of learnable parameters. To further enhance localization precision, particularly in complex construction environments, the paper plans to integrate SAM (segment anything model) for subsequent processing. SAM's ability to generate precise object masks based on prompt points will enable us to easily extract foreground object masks for the regions around the prompt points, offering a more robust approach for improving localization accuracy. This method will significantly reduce partial localization errors and increase the overall robustness of the detection model.

These findings have significant implications for both academia and industry. The framework offers a solution for adapting VLM-based systems to specialized domains with limited data availability, while providing a practical pathway for detecting construction tools and materials in real-world deployments. Compared with conventional deep learning approaches, the method facilitates easier application in construction scenarios by leveraging natural language rather than machine language, enabling more effective on-site deployment. Unlike existing VLM fine-tuning methods, it does not require extensive data, labor, or complex system design. Moreover, the framework can be readily used by both site managers and field operators for construction site monitoring without requiring in-depth knowledge of data or models.

TABLE 12 Ablation experimental results of the proposed method in the ACID dataset.

		Cement truck	Compactor	Dozer	Excavator	Wheel loader	All mAP
1-shot	Y-W-loc+cls	28.99	0.49	13.42	29.74	15.15	17.56
	DINO v2+cls	82.25	0.80	30.36	71.66	54.08	47.83
	Gt-loc+cls	93.59	1.14	37.49	89.03	60.70	56.39
2-shot	Y-W-loc+cls	25.94	0.40	8.92	18.64	21.99	15.18
	DINO v2+cls	84.51	10.87	36.76	68.41	57.42	51.59
	Gt-loc+cls	97.28	14.34	41.42	83.60	63.00	59.93
4-shot	Y-W-loc+cls	24.89	1.92	23.43	25.06	12.62	17.58
	DINO v2+cls	80.70	17.27	51.46	68.49	58.00	55.18
	Gt-loc+cls	91.77	21.82	52.41	82.65	62.12	62.15
16-shot	Y-W-loc+cls	21.84	2.98	13.03	24.61	18.23	16.14
	DINO v2+cls	78.19	59.29	45.08	67.59	34.44	56.92
	Gt-loc+cls	94.62	63.51	41.96	79.96	35.69	63.15

Note: "Y-W-loc+cls" denotes the detection performance based on YOLO-World localization method; "DINOv2+cls" denotes the detection performance based on the proposed localization method; and "Gt-loc+cls" denotes the detection performance under ground-truth localization, where only bounding boxes are provided without class labels.

TABLE 13 Comparative evaluation with few-shot fine-tuning method in self-constructed construction T&M dataset.

		Hammer	Pliers	Scaffolding coupler	Scaffolding footboard	Steel tape measure	Scaffolding steel tube	Wrench	All mAP
1-shot	Ours	46.39	37.19	51.98	61.33	69.36	80.08	44.79	55.87
	FSCE	2.10	14.20	12.90	14.40	24.40	2.50	12.50	11.86
2-shot	Ours	51.91	39.66	41.60	77.30	67.81	79.92	45.88	57.73
	FSCE	15.60	12.70	29.70	14.50	34.10	10.40	14.00	18.71
4-shot	Ours	44.56	43.58	54.35	72.14	67.34	80.92	53.25	59.45
	FSCE	33.50	27.60	60.40	42.10	51.90	18.60	33.80	38.27

TABLE 14 Comparative evaluation with few-shot fine-tuning method in the ACID dataset.

		Cement truck	Compactor	Dozer	Excavator	Wheel loader	All mAP
1-shot	Ours	82.25	0.80	30.36	71.66	54.08	47.83
	FSCE	12.20	17.20	15.00	21.10	22.10	17.52
2-shot	Ours	84.51	10.87	36.76	68.41	57.42	51.59
	FSCE	21.90	26.60	28.20	26.40	16.60	23.94
4-shot	Ours	80.70	17.27	51.46	68.49	58.00	55.18
	FSCE	51.90	49.00	37.90	43.20	38.30	44.06

6 | LIMITATIONS

This study has some limitations. First, while it can predict the types of construction objects present in an image (e.g., the pliers category), the proposed image encoding feature extraction is relatively discrete and not fully pixel-based, which often results in only partial localization of the object, as illustrated in Figure 5b(a). As a training-free approach, the method lacks a learning mechanism, which may lead to the misidentification of local features as complete objects, as shown in Figure 5b(b). To address this issue, an effective solution would be to integrate the (Kirillov et al., 2023) model to enable full-pixel segmentation of the detected objects. Moreover, the performance of the model could be further impacted by the diversity and quality of the available domain knowledge. Despite showing promising results in the construction domain, the generalization of the model to other

domains with limited prior knowledge still needs to be explored.

Second, although the proposed method achieves promising results in object detection tasks within construction scenarios, it has not yet been evaluated on large-scale multi-class detection tasks (e.g., detecting 100 distinct object categories simultaneously). Future work should explore the impact of category diversity on VLMs to improve their generalizability and scalability. The method requires workers or professionals to perform a small amount of foreground object annotation for synthetic image generation. While synthetic images help enrich domain knowledge and improve visual-text alignment, they may introduce minor biases that slightly affect localization precision in cluttered backgrounds. For example, different workers may have varying definitions of construction tools, and fine-grained distinctions among tools and materials could lead to errors in the image generation process by the generative model, which in turn affects subsequent image feature matching. In future research, exploring domain adaptation and fine-grained feature alignment techniques will be a key direction to further mitigate the impact of modality gaps.

Third, while the training-free detection framework demonstrates reliable performance on two different datasets, it also has certain limitations. For instance, the robustness of the system may be constrained in small-object detection. More fine-grained feature extraction, such as increasing the input image resolution of the DINO model and enlarging the number of patches, can improve detection accuracy in visual information processing. Moreover, the current approach has not been tested on real construction robots. Future research should integrate the proposed method with robotic grasping strategies to enable real-time deployment in actual construction environments. In addition, while this is slower than fully optimized detection networks such as YOLO-World (17.86 FPS), the proposed training-free framework emphasizes flexibility and data efficiency rather than real-time performance. Future work will focus on CUDA acceleration and lightweight optimization to improve deployment efficiency.

7 | CONCLUSION

This paper presents a training-free, knowledge-enhanced T&M recognition method using VLM in construction. Experimental results demonstrated the effectiveness of the approach. In the 4-shot scenario, the method achieved a recognition accuracy of 59.45%, compared to 20.38% in the zero-shot setting, highlighting the benefits of integrating domain-specific visual and textual knowledge. These findings confirm that the proposed method can provide reliable recognition performance while avoiding costly data annotation and retraining. The proposed framework contributes to knowledge-enhanced deployment of intelligent visual systems in construction, enhancing their adaptability to dynamic environments and supporting worker–robot collaboration. Beyond tool and material recognition, this approach also holds potential for broader applications such as construction safety monitoring and automated resource management. Future work will focus on improving robustness under complex field conditions and exploring complementary supervised learning approaches, such as neural dynamic classification (Rafiei & Adeli, 2017), dynamic ensemble learning (Alam et al., 2020), and the finite element machine (Pereira et al., 2020), to further enhance recognition accuracy and generalization.

NOTATION

VLM	Vision-language model
T&M	Tools and materials data
ACID	Alberta Construction Image Data
DINOv2	A self-supervised image pre-training model
CLIP	Contrastive language-image pre-training model
TIP-Adapter	Training-free vision-language image classification model
TIP-X	Training-free vision-language image classification model
KL	Kullback–Leibler divergence
mAP	Mean average precision, evaluation metric for detection accuracy
YOLO-World	Zero-shot detection model used for baseline comparison
SIM	Represents the feature similarity matrix between the template image and the detection image.
Logits_{zs}	The logits under zero-shot CLIP conditions
$L_{TIP-Adapter}$	The final mixed logits of TIP-Adapter method
L_{TIP-X}	The final mixed logits of TIP-X method

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [grant number 72201226]; the Research Grants Council (Hong Kong) [grant number 26208323, C6044-23GF].

REFERENCES

Alam, K. M. D. R., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*(12), 8675–8690. https://doi.org/10.1007/s00521-019-04359-7

Areerob, K., Nguyen, V., Li, X., Inadomi, S., Shimada, T., Kanasaki, H., Wang, Z., Suganuma, M., Nagatani, K., Chun, P., & Okatani, T. (2025). Multimodal artificial intelligence approaches using large language models for expert-level landslide image analysis. *Computer-Aided Civil and Infrastructure Engineering*, *40*(19), 2900–2921. https://doi.org/10.1111/mice.13482

Atik, A., Li, K., Hajibabai, L., & Hajbabaie, A. (2025). Integrated column generation for volunteer-based delivery assignment and route optimization. *Computer-Aided Civil and Infrastructure Engineering*, *40*(15), 2227–2243. https://doi.org/10.1111/mice.13439

Bang, S., Hong, Y., & Kim, H. (2021). Proactive proximity monitoring with instance segmentation and unmanned aerial vehicle-acquired video-frame prediction. *Computer-Aided Civil and Infrastructure Engineering*, *36*(6), 800–816. https://doi.org/10.1111/mice.12672

Bavelos, A. C., Anastasiou, E., Dimitropoulos, N., Michalos, G., & Makris, S. (2025). Virtual reality-based dynamic scene recreation and robot teleoperation for hazardous environments. *Computer-Aided Civil and Infrastructure Engineering*, *40*(3), 392–408. https://doi.org/10.1111/mice.13337

Cai, J., Gao, Z., Guo, Y., Wibranek, B., & Li, S. (2024). FedHIP: Federated learning for privacy-preserving human intention prediction in human-robot collaborative assembly tasks. *Advanced Engineering Informatics*, *60*, 102411. https://doi.org/10.1016/j.aei.2024.102411

Cai, J., Zhang, Y., Yang, L., Cai, H., & Li, S. (2020). A context-augmented deep learning approach for worker trajectory prediction on unstructured and dynamic construction sites. *Advanced Engineering Informatics*, *46*, 102411. https://doi.org/10.1016/j.aei.2020.101173

Chan, C.-F., Wong, P. K.-Y., Guo, X., Cheng, J. C. P., Chan, J. P.-C., Leung, P.-H., & Tao, X. (2025). Context-aware vision-language model agent enriched with domain-specific ontology for construction site safety monitoring. *Automation in Construction*, *177*, 106305. https://doi.org/10.1016/j.autcon.2025.106305

Chen, C., Xiao, B., Zhang, Y., & Zhu, Z. (2023). Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition. *Automation in Construction*, *146*, 104702. https://doi.org/10.1016/j.autcon.2022.104702

Chen, J., Huang, Q., Chen, W., Li, Y., & Chen, Y. (2024). Automated counting of steel construction materials: Model, methodology, and online deployment. *Buildings*, *14*(6), 1661. https://doi.org/10.3390/buildings14061661

Chen, K., Zhou, X., Bao, Z., Skibniewski, M. J., & Fang, W. (2025). Artificial intelligence in infrastructure construction: A critical review. *Frontiers of Engineering Management*, *12*(1), 24–38. https://doi.org/10.1007/s42524-024-3128-5

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., & Shan, Y. (2024). YOLO-World: Real-time open-vocabulary object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI (pp. 16901–16911).

Chun, P., Yamane, T., & Maemura, Y. (2022). A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. *Computer-Aided Civil and Infrastructure Engineering*, *37*(11), 1387–1401. https://doi.org/10.1111/mice.12793

Deng, H., Hong, H., Luo, D., Deng, Y., & Su, C. (2020). Automatic indoor construction process monitoring for tiles based on BIM and computer vision. *Journal of Construction Engineering and Management*, *146*(1), 0001744. https://doi.org/10.1061/(asce)co.1943-7862.0001744

Dong, W., Wu, X. J., Xu, T., Feng, Z., Ahmed, S. A. A., Awais, M., & Kittler, J. (2024). View-shuffled clustering via the modified Hungarian algorithm. *Neural Networks*, *179*, 106602. https://doi.org/10.1016/J.NEUNET.2024.106602

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*, 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fan, J., Yin, Y., Wang, T., Dong, W., Zheng, P., & Wang, L. (2025). Vision-language model-based human-robot collaboration for smart manufacturing: A state-of-the-art survey. *Frontiers of Engineering Management*, *12*(1), 177–200. https://doi.org/10.1007/s42524-025-4136-9

Fei, Y., Liao, W., Lu, X., & Guan, H. (2024). Knowledge-enhanced graph neural networks for construction material quantity estimation of reinforced concrete buildings. *Computer-Aided Civil and Infrastructure Engineering*, *39*(4), 518–538. https://doi.org/10.1111/mice.13094

Gil, D., & Lee, G. (2024). Zero-shot monitoring of construction workers' personal protective equipment based on image captioning. *Automation in Construction*, *164*, 105470. https://doi.org/10.1016/J.AUTCON.2024.105470

Huang, J., Yang, X., Zhou, F., Li, X., Zhou, B., Lu, S., Ivashov, S., Giannakis, I., Kong, F., & Slob, E. (2024). A deep learning framework based on improved self-supervised learning for ground-penetrating radar tunnel lining inspection. *Computer-Aided Civil and Infrastructure Engineering*, *39*(6), 814–833. https://doi.org/10.1111/mice.13042

Jeoung, J., Jung, S., & Hong, T. (2025). Zero-shot framework for construction equipment task monitoring. *Computer-Aided Civil and Infrastructure Engineering*, *40*(21), 3218–3238. https://doi.org/10.1111/mice.13506

Jiang, T., Frøseth, G. T., Rønnquist, A., Kong, X., & Deng, L. (2024). A visual inspection and diagnosis system for bridge rivets based on a convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, *39*(24), 3786–3804. https://doi.org/10.1111/mice.13274

Jung, Y., Cho, I., Hsu, S. H., & Golparvar-Fard, M. (2024). VisualSiteDiary: A detector-free Vision-Language Transformer model for captioning photologs for daily construction reporting and image retrievals. *Automation in Construction*, *165*, 105483. https://doi.org/10.1016/J.AUTCON.2024.105483

Jung, Y., Oh, H., & Jeong, M. M. (2019). An approach to automated detection of structural failure using chronological image analysis in temporary structures. *International Journal of Construction Management*, *19*(2), 178–185. https://doi.org/10.1080/15623599.2017.1411457

Karim, A., & Adeli, H. (1999a). CONSCOM: An OO construction scheduling and change management system. *Journal of Construction Engineering and Management*, *125*(5), 368–376. https://doi.org/10.1061/(ASCE)0733-9364(1999)125:5(368)

Karim, A., & Adeli, H. (1999b). OO information model for construction project management. *Journal of Construction Engineering and Management*, *125*(5), 361–367. https://doi.org/10.1061/(ASCE)0733-9364(1999)125:5(361)

Kim, J. (2020). Visual analytics for operation-level construction monitoring and documentation: State-of-the-art technologies, research challenges, and future directions. *Frontiers in Built Environment*, *6*, 575738. https://doi.org/10.3389/fbuil.2020.575738

Kim, J., & Chi, S. (2017). Adaptive detector and tracker on construction sites using functional integration and online learning. *Journal of Computing in Civil Engineering*, *31*(5), 0000677. https://doi.org/10.1061/(asce)cp.1943-5487.0000677

Kim, J., & Chi, S. (2021). A few-shot learning approach for database-free vision-based monitoring on construction sites. *Automation in Construction*, *124*, 103566. https://doi.org/10.1016/j.autcon.2021.103566

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., Dollár, P., & Girshick, R. (2023). Segment anything. *Proceedings of the IEEE International Conference on Computer Vision*, Paris, France. https://doi.org/10.1109/ICCV51070.2023.00371

Kumar, S., Gupta, H., Yadav, D., Ansari, I. A., & Verma, O. P. (2022). YOLOv4 algorithm for the real-time detection of fire and personal protective equipments at construction sites. *Multimedia Tools and Applications*, *81*(16), 22163–22183. https://doi.org/10.1007/s11042-021-11280-6

Kunlamai, T., Yamane, T., Suganuma, M., Chun, P., & Okatani, T. (2024). Improving visual question answering for bridge inspection by pre-training with external data of image–text pairs. *Computer-Aided Civil and Infrastructure Engineering*, *39*(3), 345–361. https://doi.org/10.1111/mice.13086

Li, Y., & Chen, J. (2022). Computer vision–based counting model for dense steel pipe on construction sites. *Journal of Construction Engineering and Management*, *148*(1), 0002217. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002217

Liang, Y., Vadakkepat, P., Chua, D. K. H., Wang, S., Li, Z., & Zhang, S. (2024). Recognizing temporary construction site objects using CLIP-based few-shot learning and multi-modal prototypes. *Automation in Construction*, *165*, 105542. https://doi.org/10.1016/j.autcon.2024.105542

Liu, K., & Golparvar-Fard, M. (2015). Crowdsourcing construction activity analysis from jobsite video streams. *Journal of Construction Engineering and Management*, *141*(11), 0001010. https://doi.org/10.1061/(asce)co.1943-7862.0001010

Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., & Shen, C. (2023). Matcher: Segment anything with one shot using all-purpose feature matching. *The Twelfth International Conference on Learning Representations*, Vienna, Austria.

Mitterberger, D., Atanasova, L., Dörfler, K., Gramazio, F., & Kohler, M. (2022). Tie a knot: Human–robot cooperative workflow for assembling wooden structures using rope joints. *Construction Robotics*, *6*(3-4), 277–292. https://doi.org/10.1007/s41693-022-00083-2

OpenAI. (2023). *Introducing ChatGPT*. https://openai.com/blog/chatgpt/

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2024). DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=a68SUt6zFt

Pan, Z., & Yu, Y. (2024). Learning multi-granular worker intentions from incomplete visual observations for worker-robot collaboration in construction. *Automation in Construction*, *158*, 105184. https://doi.org/10.1016/j.autcon.2023.105184

Pereira, D. R., Piteri, M. A., Souza, A. N., Papa, J. P., & Adeli, H. (2020). FEMa: A finite element machine for fast learning. *Neural Computing and Applications*, *32*(10), 6393–6404. https://doi.org/10.1007/s00521-019-04146-4

Pour Rahimian, F., Seyedzadeh, S., Oliver, S., Rodriguez, S., & Dawood, N. (2020). On-demand monitoring of construction projects through a game-like hybrid application of BIM and machine learning. *Automation in Construction*, *110*, 103012. https://doi.org/10.1016/j.autcon.2019.103012

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of Machine Learning Research*, Virtual.

Rafiei, M. H., & Adeli, H. (2017). A new neural dynamic classification algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 3074–3083. https://doi.org/10.1109/TNNLS.2017.2682102

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Ren, T., & Jebelli, H. (2025). Efficient 3D robotic mapping and navigation method in complex construction environments. *Computer-Aided Civil and Infrastructure Engineering*, *40*(12), 1580–1605. https://doi.org/10.1111/mice.13353

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022*, New Orleans, LA. https://doi.org/10.1109/CVPR52688.2022.01042

Romera-Paredes, B., & Torr, P. H. S. (2015). An embarrassingly simple approach to zero-shot learning. In R. Feris, C. Lampert, & D. Parikh (Eds.), *32nd International Conference on Machine Learning, ICML 2015* (Vol. *3*, pp. 11–30). Springer. https://doi.org/10.1007/978-3-319-50077-5_2

Shen, J., Xiong, X., Xue, Z., & Bian, Y. (2019). A convolutional neural-network-based pedestrian counting model for various crowded scenes. *Computer-Aided Civil and Infrastructure Engineering*, *34*(10), 12454. https://doi.org/10.1111/mice.12454

Sun, B., Li, B., Cai, S., Yuan, Y., & Zhang, C. (2021). FSCE: Few-shot object detection via contrastive proposal encoding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Virtual (pp. 7348–7358). https://doi.org/10.1109/CVPR46437.2021.00727

Sun, H., Song, L., & Yu, Z. (2025). Bridge damage identification using a small amount of damage labeling data. *Computer-Aided Civil and Infrastructure Engineering*, *40*(18), 2729–2750. https://doi.org/10.1111/mice.13470

Sun, M., & Gao, Y. (2022). GATER: Learning grasp-action-target embeddings and relations for task-specific grasping. *IEEE Robotics*

and Automation Letters, *7*(1), 3131378. https://doi.org/10.1109/LRA.2021.3131378

Udandarao, V., Gupta, A., & Albanie, S. (2023). SuS-X: Training-free name-only transfer of vision-language models. *Proceedings of the IEEE International Conference on Computer Vision*, Paris, France. https://doi.org/10.1109/ICCV51070.2023.00257

Wang, X., & El-Gohary, N. (2024). Few-shot object detection and attribute recognition from construction site images for improved field compliance. *Automation in Construction*, *167*, 105539. https://doi.org/10.1016/j.autcon.2024.105539

Wang, X., Wang, S., Menassa, C. C., Kamat, V. R., & McGee, W. (2023). Automatic high-level motion sequencing methods for enabling multi-tasking construction robots. *Automation in Construction*, *155*, 105071. https://doi.org/10.1016/j.autcon.2023.105071

Wang, Y., Liao, P. C., Zhang, C., Ren, Y., Sun, X., & Tang, P. (2019). Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety. *Advanced Engineering Informatics*, *42*, 101001. https://doi.org/10.1016/j.aei.2019.101001

Wang, Y., Xiao, B., Bouferguene, A., & Al-Hussein, M. (2024). Proactive safety hazard identification using visual–text semantic similarity for construction safety management. *Automation in Construction*, *166*, 105602. https://doi.org/10.1016/J.AUTCON.2024.105602

Wang, Z., Zhang, Q., Yang, B., Wu, T., Lei, K., Zhang, B., & Fang, T. (2021). Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. *Journal of Computing in Civil Engineering*, *35*(1). https://doi.org/10.1061/(asce)cp.1943-5487.0000933

Wu, Z., Pei, T., Bao, Z., Ng, S. T., Lu, G., & Chen, K. (2025). Utilizing intelligent technologies in construction and demolition waste management: From a systematic review to an implementation framework. *Frontiers of Engineering Management*, *12*(1), 1–23. https://doi.org/10.1007/s42524-024-0144-4

Xiao, B., & Kang, S.-C. (2021). Development of an image data set of construction machines for deep learning object detection. *Journal of Computing in Civil Engineering*, *35*(2), 05020005. https://doi.org/10.1061/(asce)cp.1943-5487.0000945

Yong, G., Jeon, K., Gil, D., & Lee, G. (2023). Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Computer-Aided Civil and Infrastructure Engineering*, *38*(11), 1536–1554. https://doi.org/10.1111/mice.12954

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., & Ahn, S. (2018). Bayesian model-agnostic meta-learning. *Advances in Neural Information Processing Systems*, *2018*, Montréal, Canada.

Yu, H., Kamat, V. R., Menassa, C. C., McGee, W., Guo, Y., & Lee, H. (2023). Mutual physical state-aware object handover in full-contact collaborative human-robot construction work. *Automation in Construction*, *150*, 104829. https://doi.org/10.1016/j.autcon.2023.104829

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., & Li, H. (2022). Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 13695 LNCS* (pp. 493–510). Springer. https://doi.org/10.1007/978-3-031-19833-5_29

Zhang, Y., Huang, S., Qin, J., Li, X., Zhang, Z., Fan, Q., & Tan, Q. (2025). Detection of helmet use among construction workers via helmet-head region matching and state tracking. *Automation in Construction*, *171*, 105987. https://doi.org/10.1016/J.AUTCON.2025.105987

Zhao, X., & Jebelli, H. (2025). A computational method for real-time roof defect segmentation in robotic inspection. *Computer-Aided Civil and Infrastructure Engineering*, *40*(23), 3596–3623. https://doi.org/10.1111/mice.13471

Zhu, X., Zhang, R., He, B., Zhou, A., Wang, D., Zhao, B., & Gao, P. (2023). Not all features matter: Enhancing few-shot CLIP with adaptive prior refinement. *Proceedings of the IEEE International Conference on Computer Vision*, Paris, France. https://doi.org/10.1109/ICCV51070.2023.00246

How to cite this article: Zhang, Z., Yu, Y., Pan, Z., & Antwi-Afari, M. F. (2025). Training-free few-shot construction tool and material detection using pre-trained vision-language model. *Computer-Aided Civil and Infrastructure Engineering*, 1–20. https://doi.org/10.1111/mice.70129