Hybrid Scene Graph Generation Framework for Structural Construction Activity Understanding Leveraging Human-Object Interaction and Vision-Language Models

Bryan D. Tan[1], Aureley Pradipta[1], and Jacob J. Lin[1]

[1]Department of Civil Engineering, National Taiwan University, Taiwan

r13521725@ntu.edu.tw, r13521730@ntu.edu.tw, jacoblin@ntu.edu.tw

Abstract -

Understanding on-site activities is fundamental for supporting construction planning, coordination, and performance assessment; however, this task is commonly performed through manual interpretation by human supervisors, limiting information integration with autonomous systems. Scene graph generation (SGG) provides a structured representation of visual scenes, but existing image-based methods rely heavily on supervised learning and generic ontologies, limiting their applicability to construction environments where task-specific semantics and annotated data are scarce. This paper proposes a construction-oriented scene graph representation that detects human–object interactions relevant to forming, reinforcing, and concrete pouring activities. We present a hybrid SGG framework that combines a conventional object detection module with zero-shot vision–language–model-based relation inference, enabling relation prediction without task-specific fine-tuning. The proposed framework is evaluated under scene graph detection and scene graph classification settings, achieving Recall@20 scores of 50.30% and 78.70%, respectively, on a construction-specific dataset. These results demonstrate the ability to generate structured activity representations from raw construction images under realistic data constraints, highlighting the effectiveness of vision–language models for zero-shot scene understanding in construction environments.

Keywords -

Scene Understanding; Scene Graph Generation; Human-Object Interaction; Construction Activity Monitoring

1 Introduction

Construction site activity understanding involves identifying who is performing what activity and where, and serves as a foundation for progress monitoring, coordination, and decision making [1]. Such understanding is commonly derived through human supervision based on visual observations from on-site imagery. Its current reliance on human interpretation limits direct integration of activity information with automated decision-making systems. As the construction industry moves towards adopting autonomous systems and robotics for on-site activity assistance, there is a growing need for methods that can transform raw visual data into structured, machine-interpretable representations of construction activities [2].

Activity understanding is closely coupled with human–object interaction (HOI), as construction activities are characterized by how workers interact with tools, materials, and built structures [3]. Effective activity understanding therefore requires reasoning about which subject is interacting with which objects. This perspective formulates activity understanding as a relational inference problem that explicitly captures interactions and dependencies among scene entities. In contrast, current approaches predominantly perform activity understanding as isolated activity classification, limiting its utility in downstream tasks.

Recent works in scene understanding for autonomous systems have highlighted scene graphs as a promising form of structured representation. Scene graphs encode objects and their interactions as structured graphs, providing a compact yet semantically rich abstraction of complex visual scenes. There is a significant potential for scene graphs to support downstream tasks such as visual question answering, activity analysis, and path planning. However, existing scene graph generation (SGG) approaches are largely benchmark-driven, developed on generic datasets with fixed object and predicate vocabularies, and rely heavily on supervised learning [4]. These drawbacks limit direct domain transfer to construction environments where scenes are cluttered, object and interaction ontologies are domain-specific, and annotated data are scarce.

To address these challenges, this work introduces a hybrid SGG framework tailored to complex construction environments. The proposed framework combines a conventional object detection module with a vision-language model (VLM)-based relation module for HOI reasoning. By decoupling object detection from relation detection, the framework enables more flexible integration of domain-

specific semantics without requiring a large volume of labeled data. This seamless integration is credited to relation detection being performed in a zero-shot manner by leveraging the pretrained vision-language priors of a VLM. As a result, the framework is enabled to reason novel interactions without task-specific fine-tuning. The resulting scene graph provides a structured and machine-interpretable representation of construction activities suitable for supporting downstream tasks. The objective of this work is to evaluate the feasibility of the proposed framework in generating activity representations from raw construction images. The key contributions of this work are threefold:

1. We introduce a construction-oriented scene graph formulation that models HOIs in forming, reinforcing, and pouring (FRP) activities, and propose a hybrid SGG framework for producing scene graphs from raw construction images.
2. We develop a zero-shot relation inference approach for construction scene graphs by integrating VLMs with conventional object detectors, enabling relation prediction over domain-specific ontologies without additional fine-tuning.
3. We investigate two complementary prompting strategies for zero-shot human–object relation detection: a multi-selection contrastive reasoning approach and a sequentially gated verification approach. Performance is evaluated under standard scene graph detection (SGDET) and scene graph classification (SGCLS) task settings.

2 Related Work

This section reviews prior research on vision-based construction activity understanding, focusing on how visual information is modeled and represented. First, classification based approaches to activity understanding are summarized. Then methods that explicitly model HOI using structured representations such as SGGs are reviewed. Finally, recent VLM based approaches for relation reasoning and its implications to our scope are discussed.

2.1 Activity Understanding as a Classification Problem

Vision-based construction activity understanding aims to capture worker behaviors and site operations from visual data. This task has predominantly been addressed through activity recognition, with early studies focusing on single-worker actions for monitoring productivity, safety, and task intent. Initial approaches relied on engineered visual features and conventional machine learning models [5]. Subsequent works adopted deep learning-based approaches to improve recognition accuracy in cluttered construction site conditions, for example [6] utilized a You Only Watch Once model to recognize discrete forming activities of workers in a 2D scene. More recently, unified frameworks have been proposed to jointly recognize construction activities across multiple levels, including individual, crew, and global activities, within a single learning architecture [3]. Despite the progression towards more structured recognition, existing approaches treat activities as categorical labels, and do not represent worker-object interactions (i.e., HOIs) as reusable, structured scene-level representations.

2.2 HOI and SGG-Based Representations

Effective construction activity understanding requires explicit reasoning over interactions between workers, tools, materials, and surrounding entities, rather than relying solely on activity labels. Recent domain-driven studies have highlighted HOI detection as a critical stage for enabling worker accountability, progress attribution, and productivity assessment in construction environments [7]. An early attempt to represent construction activity HOI in a structured manner was proposed in [8], which modeled construction activities as graphs over detected objects by encoding semantic relevance and spatial relationships between object pairs to support recognition of diverse and concurrent site activities. A more recent study has introduced deep learning–based interaction graphs, integrating convolutional neural networks with graph neural networks to infer worker–object interactions across multiple object categories, spatial scales, and contextual levels for construction accountability monitoring [9].

HOI detection can be viewed as a human-constrained instantiation of SGG, where interactions involving humans are embedded within a more general scene-level representation [10]. This formulation allows advances in SGG to extend to HOI by enabling structured, reusable representations of worker-object interactions.

Within construction research, scene graph-based scene understanding has been explored in prior works, particularly for safety and hazard analysis. For example, [11] leveraged a Pixel2Graph deep neural network–based SGG model to identify contact-driven safety hazards by detecting relational patterns between workers and heavy equipment in site images. Similarly, [12] have automated construction hazard identification by combining instance-level object detection with a transformer-based SGG model for relation inference, where the resulting triplets are further fused with domain knowledge in safety regulations for downstream reasoning and classification. These studies highlight the effectiveness of scene graphs as a semantically rich interface between visual perception and high-level reasoning in construction environments.

2.3 VLMs for Zero-Shot Relation Reasoning

Most existing SGG approaches are developed and evaluated on generic benchmarks with fixed object and predicate vocabularies and rely heavily on supervised learning [4]. These assumptions limit their applicability to construction environments, where scenes are highly cluttered, interaction semantics are domain-specific, and large-scale annotated relation data are rarely available. To address this domain dependence, several recent SGG approaches have explored VLMs for relation reasoning, leveraging their pretrained semantic priors to enable zero-shot or weakly supervised scene understanding without task-specific relation annotations.

For example, [13] introduced a novel open-vocabulary SGG framework that utilizes a Bootstrapping Language-Image Pre-training VLM to generate scene captions from 2D images and subsequently mines linguistic cues for object grounding and predicate alignment. While effective, this approach remains weakly supervised and relies on domain-specific datasets for grounding and alignment. In contrast, IndVisSGG [14] adopts a multi-agent GPT-4 with vision VLM-based pipeline to directly infer relation triplets from images without instance localization, enabling fully training-free relation prediction.

Despite promising results on generic benchmarks, transferring such VLM-based SGG approaches to construction domains remains non-trivial. Construction activity understanding requires instance-level grounding of workers, tools, and materials, as well as task-oriented ontologies aligned with construction processes. Practical deployment in construction settings necessitates integrating object detection and adapting relation vocabularies to domain-specific interaction semantics.

3 Methodology

3.1 Method Overview

The objective of the proposed framework is to decompose a construction site image into a structured graph representation $G = \{N, E\}$, where nodes N correspond to object instances and edges E represent interactions between object pairs. Each object instance $o_i \in N$ is localized to a bounding box o_i^{bbox} and classified into an object category o_i^{cat} using an off-the-shelf transformer-based object detection model.

Given the detected object instances, the relation detection module constructs candidate subject-object pairs and infers interaction predicates using a zero-shot VLM. For each candidate pair, a pairwise crop is generated from the union of the corresponding bounding boxes, where the subject and object regions are emphasized using translucent masks while background pixels are suppressed. The relation predicate r_j of each interaction is inferred directly by the VLM without fine-tuning.

To balance interaction coverage and inference efficiency, this work evaluates two complementary strategies for generating candidate subject-object pairs and VLM inference prompting. The first strategy selects a fixed number of spatially nearest object instances for each worker instance and composes the pairwise crops into a single collage image to be jointly processed by the VLM in a single inference pass. The second strategy considers all object instances whose bounding boxes intersect with a worker instance and processes each resulting pairwise crop independently through a sequential VLM-based reasoning pipeline.

Each inferred interaction is represented as a directed triplet $\langle s_j, r_j, o_j \rangle$ where s_j denotes the worker subject, o_j denotes the object node, and r_j represents the interaction (relation predicate). Since our scope focuses on interactions between human-object pairs, the subject node s_j is constrained to only correspond to $worker$ class instances, while other object nodes may represent tools, materials, or constructed elements detected in the construction scene. The set of triplets within the resulting scene graph encodes both spatially grounded object instances and their task-relevant interactions. This set is further enriched using domain knowledge to semantically enhance the final scene graph. The hybrid SGG framework pipeline is illustrated in Figure 1.

3.2 Ontology Selection and Object Detection

The complexity of SGG models increases exponentially with the number of object categories and interaction predicates. Given the limited availability of domain-specific annotated data in dynamic construction environments, it is necessary to balance the level of granularity captured in the scene graph and the complexity of the model. In this work, the scope is limited to FRP activities; therefore, a task-oriented set of object categories and interaction predicates is defined in Table 1 to preserve sufficient semantic detail while limiting complexity.

Table 1. Ontology of object and predicate categories considered in the proposed framework. Items marked with * denote semantically enriched entities or relations inferred via domain knowledge, rather than directly detected by the object detector or VLM.

Objects		Predicates
concrete_broom	rebar_material	assemble
concrete_hose	rebar_structure	erect
concrete_leveler	vibrator_pack	transport
concrete_separator	vibrator_rod	operate
concrete_trowel	worker	pour*
formwork_material	concrete*	spread*
formwork_structure		compact*

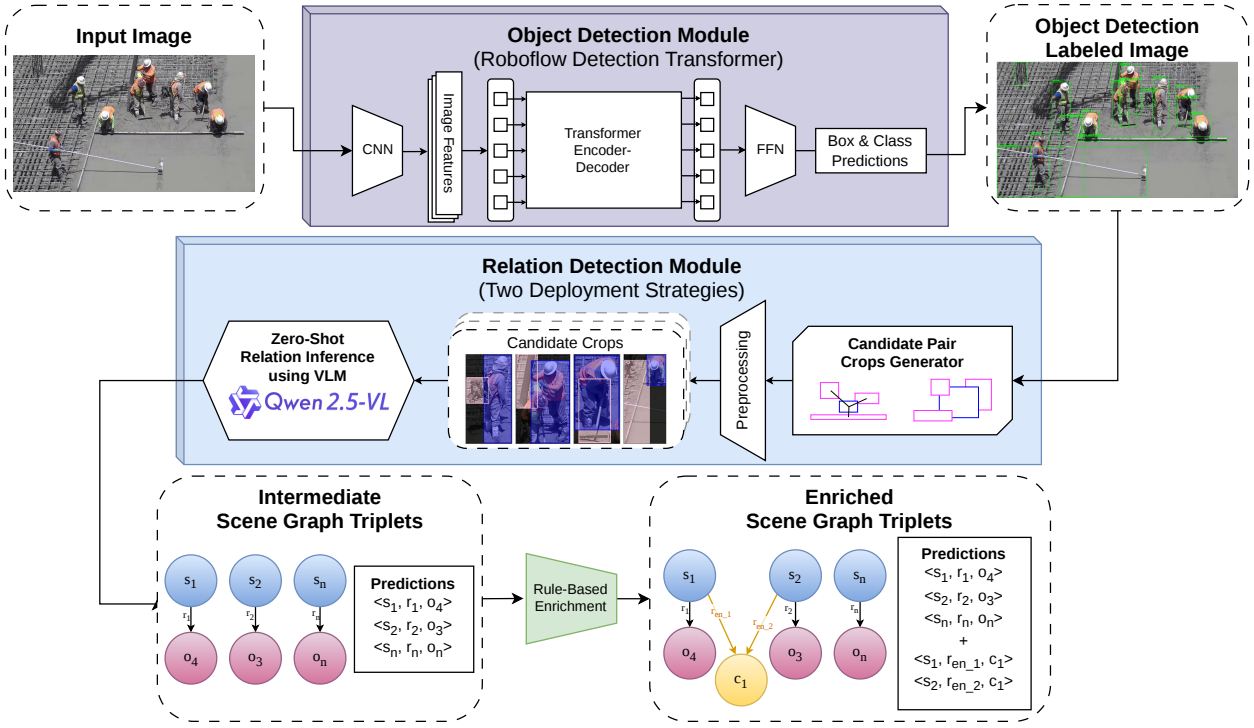

Figure 1. Overview of the Hybrid SGG framework for Construction Activity Understanding.

For object instance detection, a fine-tuned Roboflow Detection Transformer (RF-DETR) object detection model [15] is adopted for its transformer-based architecture, enabling stable end-to-end detection without hand-crafted anchor designs and competitive performance in cluttered scenes. Since the primary contribution of this work lies in HOI reasoning and SGG, the proposed framework is detector-agnostic, hence alternative off-the-shelf object detectors are expected to produce comparable object detection performance without affecting overall formulation.

3.3 VLM-based Relation Detection

For relation detection, VLMs are leveraged for their strong prior knowledge and reasoning capabilities. This approach is advantageous in our scope as the interaction ontology is domain-specific and does not align with existing large-scale benchmarks. As no relation annotations are used to train or fine-tune the model, relation inference is performed in a zero-shot manner. We formulate relation inference as a prompting-based reasoning task. As different prompting strategies exhibit trade-offs in inference performance, we investigate two complementary prompting strategies for human-object relation detection.

Bridging bounding boxes to VLMs. Simply presenting the full image annotated with bounding boxes and object labels to the VLM is insufficient for reliable relation inference as it lacks spatial and contextual cues. To bridge this gap, the proposed method generates pairwise crop candidates of the union region of the subject and object bounding boxes for each potential interaction. To further guide VLM attention, a translucent background mask is applied to suppress background information while retaining visual structure of the region of interest. Additionally, a light translucent color mask is overlaid on the subject and object regions to emphasize roles. Each masked crop candidate is provided to the VLM together with a textual prompt that explicitly outlines the object class categories. The image preprocessing stages are illustrated in Figure 2.

Strategy 1: Multi-selection contrastive reasoning; This strategy infers interaction at the worker level. For each detected worker instance $s_i \in N$ in a labeled image, we select the top K spatially nearest non-subject instances based on the Euclidean distance of the bounding box centroids. For each resulting worker-object pair, a pairwise masked crop candidate is generated as described in the preceding section. The K crop candidates containing the same subject s_i are then composed into a single collage image and provided to the VLM for reasoning. The model is prompted to identify whether any collage element exhibits a clear directional interaction between the worker and the corresponding object included in the crop. In the event of a clear interaction, the VLM is also prompted to identify the most appropriate predicate. This procedure is repeated once per detected worker instance.

Strategy 2: Sequentially gated verification. In con-

Figure 2. The preprocessing stages applied to each candidate pair in both deployment strategies. The red dotted outline is shown for illustrative purposes only and is not part of the candidate crop.

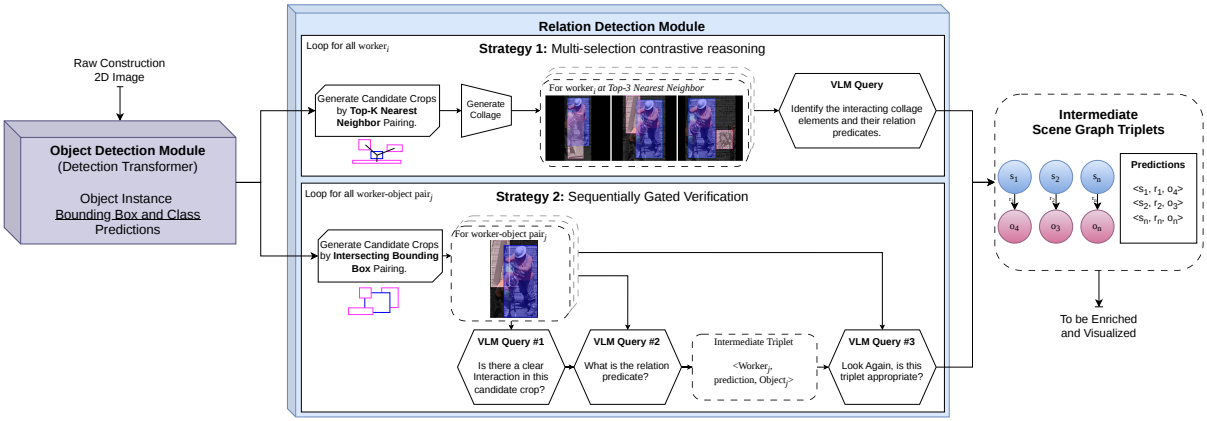

Figure 3. Illustration of the two deployment strategies used in the Relation Detection Module.

trast, this strategy infers interaction at the pairwise level. For each detected worker instance $s_i \in N$, pairwise crop candidates are generated for all object instances whose bounding box intersects with a worker's bounding box. Each crop candidate is independently processed through a three-stage gated reasoning pipeline to progressively verify the validity and type of interaction. The first stage is a binary relation gate which determines if there is any meaningful interaction between the worker-object pair in the given crop candidate; non-meaningful crops are discarded immediately. The second stage performs predicate classification on the surviving candidate pairs from the first stage, identifying the most appropriate interaction label from a closed predicate set.

In the third and final stage, a verifying "look-again" gate re-evaluates the predicted triplet by jointly prompting the inferred predicate from the second stage to repair hallucinated or inconsistent outputs. In contrast to the first strategy, this method decomposes the relation detection inference into simpler decision steps. This procedure is repeated for each worker–object pair with intersecting bounding boxes.

Both pairwise crop candidate generation methods relies on spatial heuristics such as bounding box proximity and overlap. Although these assumptions align well with the defined ontology, which comprises of short-range interactions, they may not capture all possible interaction cases (e.g., non-contact interactions). However, such cases are less prevalent within the scoped activity definitions. Figure 3 illustrates both strategies employed in the Relation Detection Module.

3.4 Rule-based Scene Graph Enrichment

Certain construction objects, such as wet concrete, exhibit amorphous visual characteristics that pose a great challenge for instance-level object detection. As such, these objects are not explicitly detected by the object detection module. To address this, we introduce a rule-based scene graph enrichment strategy that leverages domain knowledge of tool usage in construction activities. Specifically, using the assumption that activities such as concrete pouring, spreading, and compacting require distinct and dedicated tools, additional interaction triplets can be directly inferred from detected human-tool relations.

For example, $\langle worker, operate, vibrator\_rod \rangle$ im-

plies a concrete compaction activity. This implication allows the inferred triplet $\langle worker, compact, concrete \rangle$ to be appended to the final scene graph. This enrichment process augments the scene graph with semantically meaningful but visually ambiguous entities, resulting in a more complete graph representation of construction activities. These enriched triplets are considered *inferred relations*, derived from rule-based assumptions rather than direct visual grounding. As such, they are excluded from quantitative evaluation to avoid inflating performance metrics.

4 Experiments

4.1 Implementation

Object Detection Dataset. The object detection dataset consists of 1,107 manually annotated images spanning 12 object classes. Images are sourced from publicly available FRP construction videos collected from online platforms. To mitigate overfitting, frames are sampled at a strict interval from each video sequence. The dataset is randomly split into 70% training, 20% validation, and 10% testing. To improve diversity and generalization, the training subset is augmented with horizontal flips and random spatial crops in ranges of 10-30% of the original image area.

Relation Evaluation Set. A smaller set of 80 images, separate from the object detection dataset, is labeled with bounding boxes, classes, and subject-object pair interaction predicates within the defined ontology. The images in this set were selected to ensure that at least one valid worker–object interaction is present and clearly visible, such that it can be used to evaluate the hybrid SGG framework despite its limited size. However, the selection prioritizes interaction clarity over scene complexity, which may reduce variation in factors such as occlusion and clutter. Table 2 summarizes the distribution of workers, objects, and predicate support.

Object Detection Training. Object detection training is conducted on the Roboflow online training platform. An RF-DETR Medium model [15] pretrained on the Objects365 dataset [16] is fine-tuned on the custom FRP dataset for 26 epochs using default optimization settings. As training is performed on Roboflow's managed training pipeline, the detailed hyperparameter configurations are not explicitly controlled in this work. The fine-tuned detector achieves a mean average precision (mAP) of 73% on the held-out test set.

Relation Model Inference. Relation inference leverages the Qwen2.5-VL Instruct (7B) VLM sourced from the open-source repository hosted on Hugging Face [17]. The model is deployed locally on a workstation with an NVIDIA RTX 5090 GPU. Evaluation is performed in inference-only mode without fine-tuning.

4.2 Evaluation Details

Scene Graph Evaluation. SGG is typically evaluated under SGDET, SGCLS, and Predicate Classification (PredCLS) [4]. As this work targets end-to-end SGG from raw construction images, we focus on SGDET and SGCLS. PredCLS is excluded as it reduces relation prediction to a Visual Question Answering-style task. Following scene graph evaluation convention, we adopt Recall@K (R@K) and mean Recall@K (mR@K) as our evaluation metrics [4]. We set $K = 20$ to exceed valid per-image candidates while balancing inference cost.

4.3 Results

Table 3 reports the performance of the proposed framework evaluated on the 80-image evaluation set. Under the SGDET setting, the framework achieves a $R@20$ of 50.30% and a $mR@20$ of 34.39%. Higher performances are observed under the SGCLS setting, with the best configuration seeing 78.70% $R@20$.

5 Discussion

Performance. Table 3 shows that the sequentially gated verification prompting approach consistently achieves higher overall $R@20$ and $mR@20$ across both SGDET and SGCLS settings, indicating that decomposing relation inference into successive stages is effective in detecting relations in visually ambiguous scenes. A performance gap exceeding 20% between SGDET and SGCLS highlights the impact of object localization errors on relation prediction. While the gated approach yields the strongest overall performance, the multi-selection contrastive reasoning strategy attains the highest $mR@20$ under SGCLS, indicating reduced confusion. Variation in per-predicate $R@20$ reflects differences in interaction difficulty, with $erect$ and $operate$ exhibiting lower recall. The multi-selection strategy shows more balanced performance across predicates, whereas the sequentially gated approach achieves strong performance on $assemble$ but fails to detect $erect$, indicating a bias toward more visually distinct interactions.

Failure Modes. Figure 4 shows a case where the interaction between $worker_5$ and $concrete\_seperator_2$ is misclassified as $transport$ instead of $operate$. This error stems from visual pose ambiguity and single-frame inference limitations, as both interactions are plausible based on appearance cues alone. Confusion is observed most often between $operate$ and $transport$, where tool presence does not imply active usage, and between $assemble$ and $erect$ due to comparable spatial configurations.

Effects of Preprocessing. The ablation study (Table 3) shows that applying translucent color masks to subjects reduces performance, indicating that color cues are important for VLM-based interaction reasoning.

Table 2. Breakdown of per-object and per-predicate support counts in the 80-image Relation Evaluation Set

object	count	object	count	object	count	predicate	count
concrete_broom	33	concrete_trowel	11	rebar_structure	6	assemble	103
concrete_hose	18	formwork_material	16	vibrator_pack	5	erect	40
concrete_leveler	4	formwork_structure	9	vibrator_rod	19	transport	18
concrete_separator	17	rebar_material	10	worker	161	operate	8

Table 3. Hybrid SGG evaluation results. k denotes the collage size for multi-selection contrastive reasoning.

SGDET Task Evaluation @$IOU \geq 0.5$				**Per-Predicate** R@20			
Strategy	**Preprocess**	R@20	mR@20	assemble	erect	transport	operate
Multi-Select ($k = 3$)	Full	28.40	20.82	37.50	0	29.13	16.67
Multi-Select ($k = 5$)	No Mask	29.58	21.69	32.50	0	32.04	22.22
Multi-Select ($k = 3$)	No Mask	31.36	23.95	35.00	0	33.01	27.78
Sequentially Gated	Full	42.26	25.00	42.50	0	53.40	5.55
Sequentially Gated	**No Mask**	**50.30**	**34.39**	**52.50**	0	**57.28**	**27.78**
SGCLS Task Evaluation				**Per-Predicate** R@20			
Strategy	**Preprocess**	R@20	mR@20	assemble	erect	transport	operate
Multi-Select ($k = 3$)	Full	49.11	43.87	87.50	25.00	40.78	22.22
Multi-Select ($k = 3$)	No Mask	60.35	57.86	92.50	37.50	51.46	50.00
Multi-Select ($k = 5$)	**No Mask**	63.31	**59.46**	**95.00**	**38**	55.34	**50.00**
Sequentially Gated	Full	73.96	47.54	90.00	0	83.50	16.67
Sequentially Gated	**No Mask**	**78.70**	54.83	**95.00**	0	**85.44**	38.89

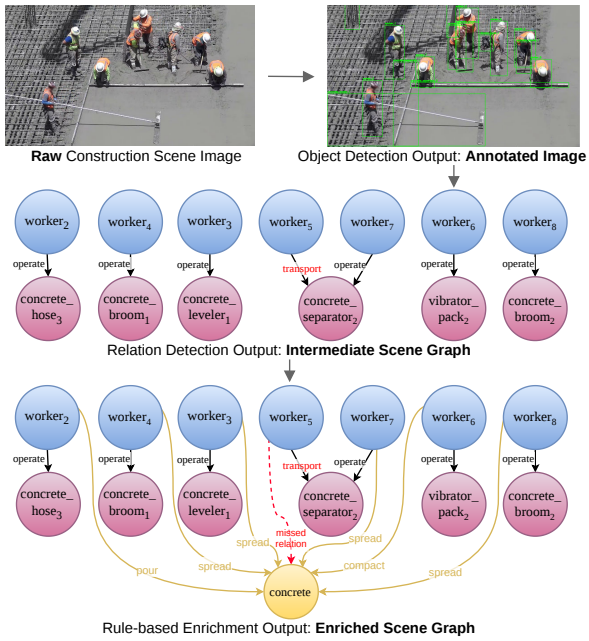

Figure 4. Sample output of the proposed hybrid SGG framework. Red labels indicate missed detections.

Limitations. The evaluation dataset is limited in size due to the cost of manually annotating object instances and interaction predicates for a construction-specific ontology. While the 80-image dataset can validate feasibility, larger datasets are required to comprehensively validate performance under diverse site conditions. In addition, the ontology set is intentionally constrained to a coarse-grained set aligned with FRP activities, which simplifies zero-shot inference but limits fine-grained action representation. This also simplifies the candidate generation stage, which relies on spatial heuristics but may not capture all valid interactions under more complex scenes. Finally, the rule-based scene graph enrichment module augments scene graphs only at the relational level and does not provide spatial grounding for inferred amorphous entities, limiting their direct use in grounded reasoning tasks.

6 Conclusion

This paper investigated the use of scene graphs to generate structured, machine-interpretable representations of construction activities from raw site images under limited labeled data constraints. A worker-centric formulation is introduced to model HOIs in FRP tasks. The proposed Hybrid SGG framework combines conventional object detection with zero-shot VLM-based relation inference, enabling interaction reasoning without task-specific training. Results demonstrate strong performance under standard SGG evaluation settings, highlighting the potential of pretrained VLMs for domain-specific reasoning with minimal annotation. Future work will focus on scaling evaluation, improving candidate generation, extending towards video reasoning, and integration with autonomous systems.

References

[1] Huimin Li, Hui Deng, and Yichuan Deng. Towards worker-centric construction scene understanding: Status quo and future directions. *Automation in*

Construction, 171:106005, 2025. ISSN 0926-5805. doi:10.1016/j.autcon.2025.106005.

[2] Jiajing Liu, Hanbin Luo, and Dongrui Wu. Human–robot collaboration in construction: Robot design, perception and interaction, and task allocation and execution. *Advanced Engineering Informatics*, 65:103109, 2025. ISSN 1474-0346. doi:10.1016/j.aei.2025.103109.

[3] Cheng Yun Tsai, Mik Wanul Khosiin, Jacob J. Lin, and Chuin-Shan Chen. Multi-granular crew activity recognition for construction monitoring. *Automation in Construction*, 179:106428, 2025. ISSN 0926-5805. doi:10.1016/j.autcon.2025.106428.

[4] Hongsheng Li, Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Xia Zhao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. *Neurocomputing*, 566:127052, 2024. ISSN 0925-2312. doi:10.1016/j.neucom.2023.127052.

[5] Jun Yang, Zhongke Shi, and Ziyan Wu. Vision-based action recognition of construction workers using dense trajectories. *Advanced Engineering Informatics*, 30(3):327–336, 2016. ISSN 1474-0346. doi:10.1016/j.aei.2016.04.009.

[6] Min-Yuan Cheng, Akhmad F.K. Khitam, and Hongky Haodiwidjaya Tanto. Construction worker productivity evaluation using action recognition for foreign labor training and education: A case study of taiwan. *Automation in Construction*, 150:104809, 2023. ISSN 0926-5805. doi:10.1016/j.autcon.2023.104809.

[7] Mik Wanul Khosiin, Jacob J. Lin, and Chuin-Shan Chen. Worker accountability in computer vision for construction productivity measurement: A systematic review. In *Proceedings of the International Conference on Construction Engineering and Project Management*, pages 775–782, 2024. doi:10.6106/ICCEPM.2024.0775.

[8] Xiaochun Luo, Heng Li, Dongping Cao, Fei Dai, JoonOh Seo, and SangHyun Lee. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *Journal of Computing in Civil Engineering*, 32(3):04018012, 2018. doi:10.1061/(ASCE)CP.1943-5487.0000756.

[9] Mik Wanul Khosiin, Jacob J. Lin, Eko Andi Suryo, Kartika Puspa Negara, Ismiarta Aknuranda, and Chuin-Shan Chen. Video-based productivity monitoring of worker and large-scale object interactions in construction sites. In *Proceedings of the 42nd International Symposium on Automation and Robotics in Construction*, pages 580–587, Montreal, Canada, July 2025. International Association for Automation and Robotics in Construction (IAARC). ISBN 978-0-6458322-2-8. doi:10.22260/ISARC2025/0076.

[10] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Exploiting scene graphs for human-object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15984–15993, October 2021.

[11] Daeho Kim, Ankit Goyal, Alejandro Newell, SangHyun Lee, Jia Deng, and Vineet R. Kamat. Semantic relation detection between construction entities to support safe human-robot collaboration in construction. In *Computing in Civil Engineering 2019*, pages 265–272, 2019. doi:10.1061/9780784482438.034.

[12] Lite Zhang, Junjie Wang, Yanbo Wang, Hai Sun, and Xuebing Zhao. Automatic construction site hazard identification integrating construction scene graphs with bert based domain knowledge. *Automation in Construction*, 142:104535, 2022. ISSN 0926-5805. doi:10.1016/j.autcon.2022.104535.

[13] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28076–28086, 2024.

[14] Zuoxu Wang, Zhijie Yan, Shufei Li, and Jihong Liu. Indvissgg: Vlm-based scene graph generation for industrial spatial intelligence. *Advanced Engineering Informatics*, 65:103107, 2025. ISSN 1474-0346. doi:10.1016/j.aei.2024.103107.

[15] Isaac Robinson, Peter Robicheaux, Matvei Popov, Deva Ramanan, and Neehar Peri. Rf-detr: neural architecture search for real-time detection transformers. *arXiv preprint arXiv:2511.09554*, 2025. doi:10.48550/arXiv.2511.0955.

[16] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019. doi:10.1109/ICCV.2019.00852.

[17] Qwen Team. Qwen2.5-vl. `https://qwenlm.github.io/blog/qwen2.5-vl/`, January 2025. Accessed via Hugging Face.