CCC 2025 **Proceedings of the Creative Construction Conference** (2025)
Edited by: Miroslaw J. Skibniewski & Miklós Hajdu & Žiga Turk
https://doi.org/10.22260/CCC2025/0023

ZERO-SHOT OBJECT DETECTION AND SEGMENTATION FOR CONSTRUCTION SITES THROUGH MULTI-MODEL INTEGRATION

Aoi Tarutani, Fuku Himuro
Shimizu Corporation, Tokyo, Japan

Abstract

Object detection and segmentation are crucial for managing construction sites, aiding in tasks such as progress tracking, material management, and safety assurance. However, conventional methods encounter persistent challenges, including occlusion, variable lighting conditions, and the labor-intensive nature of dataset creation, which limit their adaptability to dynamic construction environments. This study introduces a novel zero-shot object detection and segmentation framework designed specifically for construction-related objects, including machinery, workers, and materials. The proposed framework integrates three state-of-the-art models: Florence-2, Llama3.2-Vision, and the Segment Anything Model 2 (SAM2). Florence-2 generates region proposals for previously unseen objects using textual descriptions; Llama3.2-Vision predicts and refines accurate labels for detected regions based on textual queries; and SAM2 produces high-precision segmentation masks. The effectiveness of this approach was validated through both qualitative and quantitative experiments. While parts of this framework and qualitative experiments were previously presented, this paper extends our previous work by providing a more detailed methodology and including additional quantitative experiments. Qualitative experiments using images from a specific tunnel excavation site, demonstrating robust detection and segmentation performance under challenging conditions such as occlusion and variable lighting. Quantitative experiments using the Alberta Construction Image Dataset (ACID) showed that the proposed multi-model method significantly outperformed Florence-2 alone, particularly for large objects, despite not achieving the accuracy of the fine-tuned YOLOv11 model. The proposed framework eliminates the need for extensive retraining and manual dataset creation by leveraging the complementary strengths of these models. This scalable and flexible solution offers practical applications in progress tracking, material management, and safety monitoring and thereby addresses the unique complexities of dynamic construction environments.

Keywords: zero-shot detection, semantic segmentation, vision-language model, construction sites, monitoring.

Peer-review under responsibility of the scientific committee of the Creative Construction Conference 2025.

1. Introduction

1.1. Background

The increasing scale and complexity of buildings, coupled with a growing labor shortage at construction sites, have posed significant challenges to efficient management in areas such as construction progress, materials, and safety. To address these issues, object-detection technologies that utilize affordable and easy-to-handle cameras have gained considerable attention. However, traditional object detection models often struggle in construction-site environments because of occlusions, dynamic lighting conditions, and the presence of uncommon objects.

1.2. Related Work

Among the numerous object detection models, You Only Look Once (YOLO) is widely recognized for its high real-time detection performance. However, YOLO relies on pre-trained classes and conditions, making it challenging to adapt to environments like construction sites, where uncommon objects are prevalent, and conditions vary significantly over time and location. Additionally, the detection accuracy

Corresponding author email address: aoi.tarutani@shimz.co.jp

of YOLO often degrades under challenging conditions, such as occlusions or changes in the lighting environment.

To overcome these limitations, existing studies have focused on enhancing the applicability of YOLO through dataset creation and model fine-tuning [1-6]. However, these processes incur significant costs and labor, undermining their practicality in real-world construction sites. Therefore, there is an urgent need to develop more flexible and efficient object detection methods that reduce the effort required to create training datasets.

Recently, zero-shot learning (ZSL) has emerged as a promising solution to these challenges. ZSL enables the detection of objects from previously unseen classes without extensive retraining, leveraging auxiliary information, such as textual descriptions, semantic embeddings, or pre-trained language models. For instance, contrastive language-image pretraining (CLIP) aligns visual and textual inputs within a shared embedding space, allowing novel object detection through textual prompts [7].

Building on these advancements, advanced ZSL models, including Florence-2 [8], Llama3.2-Vision [9], and Segment Anything Model 2 (SAM2) [10], have been developed, each exhibiting distinct functionalities. Florence-2 and Llama3.2-Vision are vision-language models that integrate image and text inputs for zero-shot understanding and reasoning. Florence-2 was particularly effective at estimating the regions associated with noun phrases and output bounding boxes or segmentation masks. Llama3.2-Vision excels at generating detailed descriptions and supports complex visual reasoning. Conversely, SAM2 specializes in zero-shot segmentation, enabling the extraction of object masks for arbitrary prompts without additional training. The complementary strengths of the models are listed in Table 1.

Table 1. Comparison of Zero-Shot Models and YOLO.

Model	Zero-Shot Support	Inputs	Outputs	Key Strengths
YOLO	Not Supported	Image	Bounding box-label, Segmentation mask-label	Real-time detection of pre-defined classes
Florence-2	Supported	Image, Text	Bounding box-label, Segmentation mask-label	Region estimation, Region proposal
Llama3.2-Vision	Supported	Image, Text	Text, Image (labels)	Visual reasoning, Description generation
SAM2	Supported	Image, Bounding Boxes/Points	Segmentation mask	High-Precision Segmentation

1.3. Research Objective

Parts of the proposed method and the qualitative experiments (Section 2 and 3) were previously presented in a preliminary form in the Summaries of Technical Papers of Annual Convention of AIJ [11]. In this paper, we extend our previous work by providing a more detailed methodology and by adding quantitative experiments (Section 4).

This study proposes a novel multi-model approach that integrates Florence-2, Llama3.2-Vision, and SAM2 for zero-shot object detection and segmentation at construction sites without additional training. The objective is to develop a robust method for handling variable lighting conditions and occlusions. The effectiveness of the proposed method is validated through qualitative and quantitative experiments.

2. Proposed Method

2.1. Overview of Models

2.1.1. Florence-2

Florence-2 [8] is a vision-language model that integrates image and text inputs to estimate object regions, outputting bounding boxes, segmentation masks, and labels. It supports three main functionalities:

- OBJECT DETECTION (OD): Detect all objects in an input image without textual descriptions.
- CAPTION TO PHRASE GROUNDING: Detect objects in the input image that match specific textual descriptions.

- REFERRING EXPRESSION SEGMENTATION: Identifies segmentation masks in the input image that correspond to a textual description.

Figures 1, 2, and 3 illustrate examples of input images and the results generated using these functionalities. The input images included objects such as excavators, beams, step stools, flexible container baskets, and workers. The results demonstrate that region estimation is generally successful under challenging conditions, including occlusions and varying lighting environments.

However, the following limitations were observed:

- OBJECT DETECTION (Fig. 1): Missed detections occurred for large objects such as beams and visually ambiguous objects such as flexible container baskets.
- CAPTION TO PHRASE GROUNDING (Fig. 2): Beams and flexible container baskets were successfully detected with textual descriptions. However, occasionally, misclassified objects, such as excavators, were labeled as basket trolleys when multiple predefined categories were included in the textual descriptions, which sometimes contained categories not present in the image.
- REFERRING EXPRESSION SEGMENTATION (Fig. 3): Displayed poor performance in generating segmentation masks.

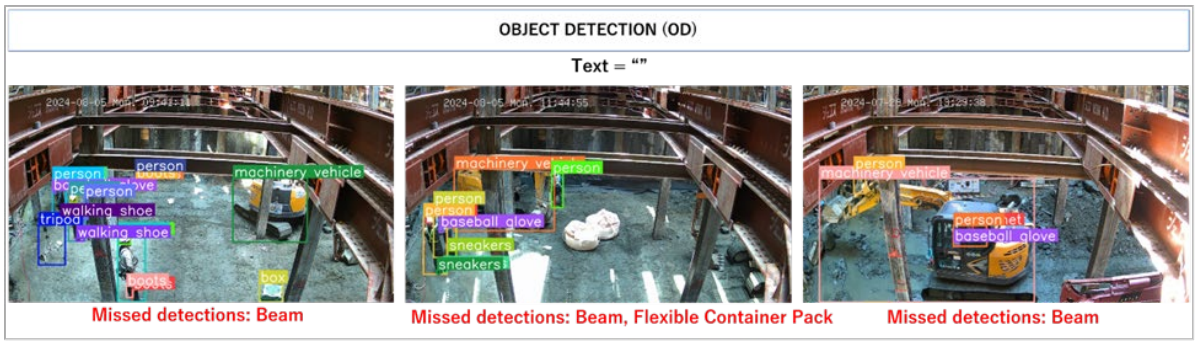

Fig. 1. Qualitative validation of Florence-2 OBJECT DETECTION for construction site objects.

Fig. 2. Qualitative validation of Florence-2 CAPTION TO PHRASE GROUNDING for construction site objects.

Fig. 3. Qualitative validation of Florence-2 REFERRING EXPRESSION SEGMENTATION for construction site objects.

In summary, Florence-2 exhibited strong region estimation capabilities, particularly when supported by textual descriptions. However, its limitations in label prediction and segmentation accuracy highlight the need for complementary models to enhance overall detection performance. Additionally, OBJECT DETECTION has shown instances of missed detections, and to address this issue, CAPTION TO PHRASE GROUNDING has been proven to be an optimal approach for preventing such omissions.

2.1.2. Llama3.2-Vision

Llama3.2-Vision [9] is a multimodal artificial intelligence (AI) model that excels in advanced reasoning across visual and textual inputs. It predicts accurate labels for the detected regions using textual queries, improves semantic consistency, and reduces the number of false positives. Additionally, it refines and filters the bounding boxes to enhance the detection precision. However, this approach does not support segmentation tasks.

2.1.3. SAM2

SAM2 [10] is a segmentation model that generates high-precision segmentation masks even under challenging conditions, such as occlusion or variable lighting. Its strength lies in its ability to delineate object boundaries with exceptional accuracy. However, it fails to predict labels, thereby necessitating integration with other models to form a complete detection pipeline.

2.2. Multi-Model Method

To address the limitations of these individual models, we proposed a multi-model approach that combines the strengths of Florence-2, Llama3.2-Vision, and SAM2:

- 1. Bounding Box Detection using Florence-2

  Florence-2 identifies the object regions and generates bounding boxes based on the input image and text. This forms a robust foundation for the subsequent steps, particularly the detection of previously unseen objects.

- 2. Label Refinement with Llama3.2-Vision

  Outputs from Florence-2 (images cropped from the bounding boxes) are refined using Llama3.2-Vision, which predicts accurate labels for the detected regions. This step minimizes false positives and ensures semantic alignment.

- 3. High-precision Segmentation using SAM2

  SAM2 generates high-precision segmentation masks for refined bounding boxes and accurately delineates object boundaries under diverse environmental conditions.

This integrated workflow addresses critical challenges, including occlusions, misclassifications, and segmentation inaccuracies. The framework provides flexibility, robustness, and scalability, making it ideal for the complex and dynamic conditions encountered at construction sites. Figure 4 illustrates the functionality and workflow of the proposed method.

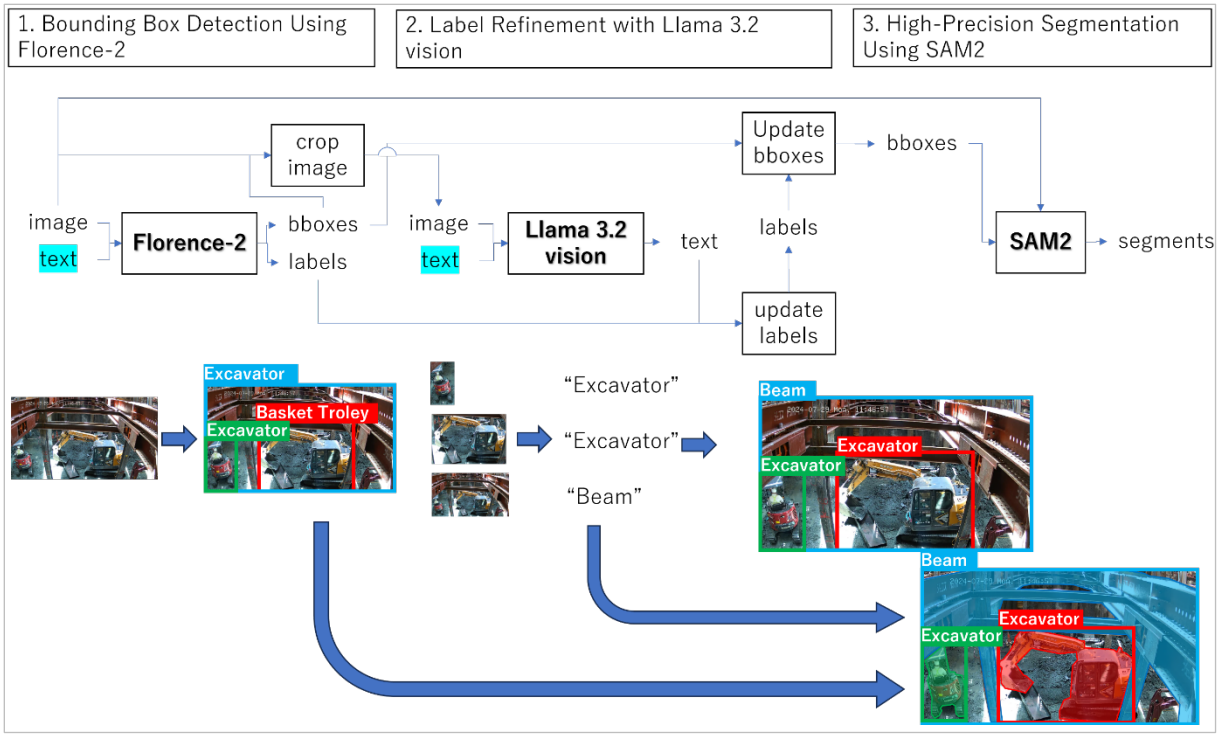

Fig. 4. Proposed multi-model method.

3. Qualitative Experiment and Results

3.1. Dataset and Objective

An image dataset was collected from a civil engineering tunnel excavation site to qualitatively evaluate the proposed multi-model method. A dataset was created using fixed cameras installed at the site to capture diverse scenarios. Representative images are shown in Fig. 5(a). These images include various features, including different lighting conditions (morning, afternoon, and evening), occlusions caused by struts, and differences in construction processes. The images contain objects such as struts, construction machinery, workers, and materials but do not have corresponding annotation data.

These images did not contain the corresponding annotation data. Therefore, we conducted a qualitative experiment by visually inspecting the detection and segmentation results for a small number of images. This approach allowed us to assess the effectiveness of the proposed multi-model method under real-world conditions, such as varying lighting and occlusions, even in the absence of ground-truth annotations for our specific dataset.

3.2. Experimental Setup

The experiments were conducted using Windows 11 as the operating system and a G-Force RTX6000 GPU. The models used were “microsoft/Florence-2-large-ft” for Florence-2, “Llama3.2-Vision:11b-instruct-fp16” for Llama3.2-Vision, and “sam2_hiera_large.pt” for SAM2.

3.3. Bounding Box Detection Using Florence-2

The first step of the proposed method involved using Florence-2 to estimate object regions and generate bounding boxes based on textual descriptions. The input images included categories such as Excavator, Beam, Column, Step Stool, Worker, Flexible Container Basket, and Basket Trolley. The CAPTION TO PHRASE GROUNDING functionality of Florence-2 was employed using the following text input:

"Excavator, Beam, Green_Step_Stool, Worker, Big_Sacks, Basket_Trolley, and Other"

Visual inspection of the region estimation results for 32 images revealed several key observations.

Fig. 5(b) shows the detection results. Florence-2 demonstrated strong region estimation capabilities, particularly for detecting previously unseen objects. However, several issues were identified. False

positives occurred when bounding boxes were generated for objects not present in the image, especially for less frequently appearing objects such as Step Stools and Basket Trolleys, while false positives may be mitigated in the subsequent label prediction stage.

3.4. Label Refinement with Llama3.2-Vision

The second step involved refining the bounding boxes generated by Florence-2 using Llama3.2-Vision for label prediction. The cropped images were input into Llama3.2-Vision along with the following text query:

> *"Return ensurely the name of the main object from the image from only one of the following captions:*
>
> *- Excavator*
>
> *- Beam*
>
> *- Green_Step_Stool*
>
> *- Worker*
>
> *- Big_Sacks*
>
> *- Basket_Trolley*
>
> *- Other"*

Llama3.2-Vision significantly improved label prediction accuracy compared to Florence-2 and successfully reduced false positives. For instance, as shown in Fig. 5(c), the mislabeling of the Excavator as Basket Trolley and Step Stool as Big Sacks by Florence-2 was corrected using Llama3.2-Vision, ensuring accurate label assignment. However, Llama3.2-Vision occasionally assigned labels to background objects that appeared in the cropped region, resulting in false detections.

3.5. High-Precision Segmentation Using SAM2

In the final step, segmentation was performed using SAM2 to generate high-precision masks for the bounding boxes refined using Llama3.2-Vision. As illustrated in Fig. 5(d), SAM2 demonstrated robust performance even under challenging conditions, such as occlusions and varying lighting environments.

3.6. Discussion of Experimental Results

The integration of Florence-2, Llama3.2-Vision, and SAM2 proved to be effective for detecting and segmenting previously unseen objects in complex construction site environments. Florence-2 provides a strong foundation for region estimation, Llama3.2-Vision improves label accuracy and reduces false positives, and SAM2 delivers high-quality segmentation masks, even under adverse conditions.

However, Llama3.2-Vision occasionally assigns labels to background objects in cropped regions, resulting in false detections. Addressing this issue may require improved prompts or additional constraints to help the model focus on the main object.

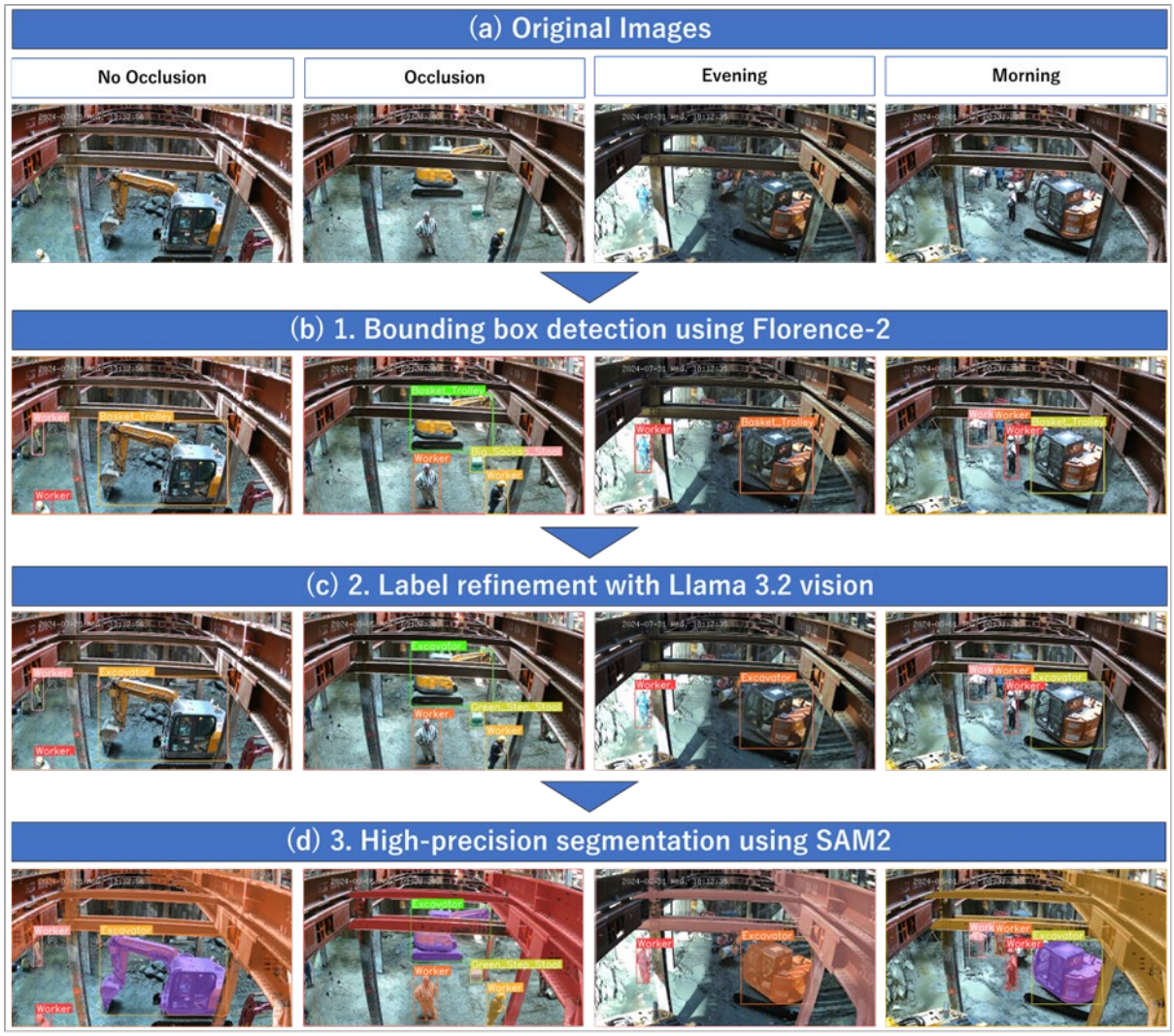

Fig. 5. Results of bounding box estimation (Florence-2), label prediction (Llama 3.2-Vision), and segmentation (SAM2).

4. Quantitative Experiment and Results

4.1. Dataset

The Alberta Construction Image Dataset (ACID) [12] was used to evaluate the multi-model method quantitatively. ACID is a comprehensive image dataset designed to train AI models for construction automation. The dataset comprises 10,000 labeled images collected from construction sites worldwide and contains 15,767 annotated construction machine instances. These annotations cover ten categories of construction equipment: dozer, backhoe loader, wheel loader, excavator, dump truck, grader, compactor, mobile crane, cement truck, and tower crane. Of the 10,000 images, 3,000 were used for evaluation.

4.2. Experimental Setup

All the experiments were conducted on a system running Windows 11 with a G-Force RTX6000 GPU. For quantitative evaluation, we compared three methods: YOLOv11, Florence-2, and our proposed multi-model approach. The evaluation was performed separately for object detection and semantic segmentation tasks.

4.2.1. Object Detection

The following methods were evaluated for object detection:

- YOLOv11: YOLOv11 (“yolo11n.pt”) was fine-tuned using 7,000 training images from the ACID dataset, excluding the 3,000 images reserved for evaluation.

- Florence-2: Florence-2 (“microsoft/Florence-2-large-ft”) was used with the CAPTION TO PHRASE GROUNDING functionality and the following text input for object detection: "dozer, backhoe_loader, wheel_loader, excavator, dump_truck, grader, compactor, mobile_crane, cement_truck, tower_crane, and other."
- Our proposed multi-model method:

  Step 1: Florence-2 (“microsoft/Florence-2-large-ft”) was used with the CAPTION TO PHRASE GROUNDING functionality and the specified text input to estimate object regions and generate bounding boxes.

  Step 2: Llama3.2-Vision (“Llama3.2-Vision:11b-instruct-fp16”) was used to refine the labels of the detected regions. The cropped region images were input with the following prompt for label refinement: “Return only the name of the main object in the image from one of the following captions: Excavator; Beam; Green_Step_Stool; Worker; Big_Sacks; Basket_Trolley; Other.”

This setup ensures a fair and reproducible comparison among the baseline YOLOv11, Florence-2, and the proposed multi-model methods. We used the official COCO evaluation toolkit (pycocotools) [13] to compute the mean Average Precision (mAP) and mean Average Recall (mAR) for all object sizes (small, medium, and large).

4.2.2. Semantic Segmentation

Florence-2 does not support semantic segmentation for multiple object classes via text input. Therefore, following two methods were evaluated:

- YOLOv11: YOLOv11 (“yolo11n-seg.pt”) was fine-tuned using 7,000 training images from the ACID dataset, excluding the 3,000 images reserved for evaluation.
- Our proposed multi-model method:

  Step 1: Florence-2 ("microsoft/Florence-2-large-ft") with CAPTION TO PHRASE GROUNDING and the specified text input was used to estimate object regions and generate bounding boxes (as described in Section 4.2.1).

  Step 2: Llama3.2-Vision ("Llama3.2-Vision:11b-instruct-fp16") was used to refine the labels of the detected regions, using the same prompt as in Section 4.2.1.

  Step 3: SAM2 (“sam2_hiera_large.pt”) was used to generate segmentation masks for the detected regions.

We used the official COCO evaluation toolkit (pycocotools) to compute the mean Average Precision (mAP) and mean Average Recall (mAR) for all object sizes (small, medium, and large).

4.3. Results

4.3.1. Object Detection Results

Table 2 summarizes the object detection performances of the three methods on the ACID dataset. The evaluation metrics include mAP for all object sizes, including small, medium, and large objects, and mAR.

Table 2. Object detection performance comparison.

Model	mAP	mAP(small)	mAP(medium)	mAP(large)	mAR
YOLOv11	0.646	0.215	0.262	0.678	0.799
Florence-2	0.025	0.002	0.012	0.031	0.134
Ours	0.383	0.003	0.059	0.421	0.676

The YOLOv11 model fine-tuned on the ACID dataset achieved the highest overall object detection performance (mAP = 0.646, mAR = 0.799), outperforming the other methods across all object sizes, particularly for large objects. Florence-2 demonstrated a limited object detection capability, with a low overall mAP of 0.025 and an mAR of 0.134, particularly for small and medium objects. The proposed multi-model method outperformed Florence-2, achieving an mAP of 0.383 and an mAR of 0.676.

Although it did not achieve the performance level of YOLOv11, it demonstrated a significant improvement over Florence-2, particularly for large objects (mAP = 0.421). Furthermore, the proposed multi-model method achieved a higher mAR (0.676) than mAP (0.383).

4.3.2. Semantic Segmentation Results

Table 3 summarizes the semantic segmentation performances of the evaluated methods. The evaluation metrics include mAP for all object sizes, including small, medium, and large objects, and mAR.

Table 3. Semantic Segmentation performance comparison.

Model	mAP	mAP(small)	mAP(medium)	mAP(large)	mAR
YOLOv11	0.533	0.026	0.127	0.599	0.704
Florence-2	-	-	-	-	-
Ours	0.389	0.000	0.114	0.424	0.663

The YOLOv11 model fine-tuned on the ACID dataset achieved the highest overall semantic segmentation performance (mAP = 0.533, mAR = 0.704), outperforming the other methods across all object sizes, particularly for large objects. Florence-2 was not evaluated for semantic segmentation, as it does not provide this functionality. The proposed multi-model method achieved an mAP of 0.389 and an mAR of 0.663. Although it did not achieve the performance level of YOLOv11, it demonstrated a significant improvement over Florence-2, particularly for large objects (mAP = 0.424). As with object detection, the proposed multi-model method achieved a higher mAR (0.663) than mAP (0.389).

4.4. Discussion

While YOLOv11 achieved the best overall performance in both object detection and semantic segmentation, the proposed multi-model approach demonstrated substantial improvements over Florence-2, particularly for large objects. These results highlight the potential of integrating vision-language models for object detection and semantic segmentation in construction automation, although further work is needed to improve the results for small- and medium-sized objects.

We also observed that Llama 3.2-Vision occasionally misclassified background objects or misinterpreted certain categories, such as labeling only the tires as "wheel loader." Providing more explicit prompts or additional contextual information may help mitigate these issues.

In addition, there was no significant difference between the object detection results (i.e., accuracy of bounding box and label estimation) and the semantic segmentation results (i.e., accuracy of segmentation mask and label estimation), indicating that SAM2 provides high segmentation performance and is a reasonable choice as a segmentation model in this pipeline.

Finally, the observation that the proposed multi-model method achieved a higher mAR than mAP suggests that while most true objects were detected, the lower precision is likely attributable to false positives or label inaccuracies, rather than deficiencies in mask quality.

5. Conclusion

In this study, we proposed a novel multi-model approach for zero-shot object detection and segmentation in construction site environments by integrating Florence-2, Llama3.2-Vision, and SAM2. The proposed method addresses the limitations of traditional object detection models, such as YOLO, which struggle with uncommon objects, occlusions, and varying lighting conditions commonly found at construction sites.

Qualitative experiments were conducted using images collected from a civil engineering tunnel excavation site. These experiments demonstrated that the proposed approach can effectively detect and segment various construction materials and equipment under real-world conditions, even in the absence of annotation data. Florence-2 provided robust region estimation, Llama3.2-Vision improved label accuracy and reduced false positives, and SAM2 delivered high-quality segmentation masks under challenging scenarios.

For a quantitative evaluation, we used the ACID dataset to compare our method with YOLOv11 and Florence-2. The results revealed that our multi-model method outperformed Florence-2 alone and significantly improved the object detection and segmentation performance for large objects. However, it did not reach the accuracy of a fine-tuned YOLOv11 model. Some challenges remain, such as the limited performance for small- and medium-sized objects and the occasional mislabeling of background elements by Llama3.2-Vision.

Our findings highlight the potential of integrating advanced vision-language and segmentation models for flexible, training-free object detection and segmentation in complex construction environments. Future work will focus on further improving the detection accuracy for small objects, enhancing the label assignment robustness, and extending the approach to other real-world scenarios.

Acknowledgements

The authors would like to thank the ACID team for providing access to the Alberta Construction Image Dataset (ACID), which was essential for the quantitative evaluation conducted in this study.

References

[1] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, “Dataset and benchmark for detecting moving objects in construction sites,” Automation in Construction, vol. 122, p. 103482, Feb. 2021, doi: 10.1016/j.autcon.2020.103482.
[2] B. Xiao and S.-C. Kang, “Development of an Image Data Set of Construction Machines for Deep Learning Object Detection,” Journal of Computing in Civil Engineering, vol. 35, no. 2, Mar. 2021, doi: 10.1061/(asce)cp.1943-5487.0000945.
[3] M. Baubriaud, S. Derrode, R. Chalon, and K. Kernn, “Accelerating Indoor Construction Progress Monitoring with Synthetic Data-Powered Deep Learning,” Proceedings of the 41st International Symposium on Automation and Robotics in Construction, Jun. 2024, doi: 10.22260/isarc2024/0103.
[4] S. Han, W. Park, K. Jeong, T. Hong, and C. Koo, “Utilizing synthetic images to enhance the automated recognition of small-sized construction tools,” Automation in Construction, vol. 163, p. 105415, Jul. 2024, doi: 10.1016/j.autcon.2024.105415.
[5] Y. Ding, M. Zhang, J. Pan, J. Hu, and X. Luo, “Robust object detection in extreme construction conditions,” Automation in Construction, vol. 165, p. 105487, Sep. 2024, doi: 10.1016/j.autcon.2024.105487.
[6] T. Kim, M. Koo, J. Hyeon, and H. Kim, “Zero-shot Learning-based Polygon Mask Generation for Construction Objects,” Proceedings of the 41st International Symposium on Automation and Robotics in Construction, Jun. 2024, doi: 10.22260/isarc2024/0012.
[7] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” Computer Vision and Pattern Recognition, 2021. https://doi.org/10.48550/arXiv.2103.00020
[8] B. Xiao et al., “Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks,” 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4818–4829, Jun. 2024, doi: 10.1109/cvpr52733.2024.00461.
[9] Meta, “Llama 3.2: Revolutionizing edge AI and vision with open,” customizable models https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices, Accessed: Apr. 30, 2025.
[10] N. Ravi, V. Gabeur, Y. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K.V. Alwala, N. Carion, C. Wu, R. Girshick, P. Dollár, C. Feichtenhofer, “SAM2: Segment Anything in Images and Videos,” October. 2024, https://doi.org/10.48550/arXiv.2408.00714.
[11] A. Tarutani, F. Himuro, “Zero-Shot Object Detection and Semantic Segmentation on Construction Sites by Integrating Multiple Models (in Japanese),” Summaries of Technical Papers of Annual Meeting, Architectural Institute of Japan, Kyushu, to appear, Sep. 2025, in Japanese.
[12] B. Xiao and S.-C. Kang, “Development of an Image Data Set of Construction Machines for Deep Learning Object Detection,” Journal of Computing in Civil Engineering, vol. 35, no. 2, Mar. 2021, doi: 10.1061/(asce)cp.1943-5487.0000945.
[13] "COCO API," GitHub repository, https://github.com/cocodataset/cocoapi, Accessed: Apr. 30, 2025.