HOKKAIDO UNIVERSITY

Title	Study of Vision-Language Segment-Anything Models for Bridge Defect Recognition
Author(s)	Jiang, Bowen; Chun1, Pang-jo
Relation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026), June 28 - July 1, 2026, Sapporo, Japan
Citation	Proceedings of the 10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026)
Issue Date	2026-07-01
Doc URL	https://hdl.handle.net/2115/100473
Type	conference paper
File Information	C000230_1.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

10th International Symposium on Reliability Engineering and Risk Management (ISRERM 2026),
28 June-01 July 2026, Sapporo, Japan
Kitahara, T., Miyamori, Y. & Beer, M. (editors)

Study of Vision-Language Segment-Anything Models for Bridge Defect Recognition

Bowen Jiang[1], Pang-jo Chun[1, 2, *]

[1]*Department of Civil Engineering, The University of Tokyo, Japan*
E-mail: 1625041410@g.ecc.u-tokyo.ac.jp
[2]*Institute of Engineering Innovation, The University of Tokyo, Japan*
E-mail: chun@g.ecc.u-tokyo.ac.jp
**Corresponding Author*

As bridge inspection is often manual, labor-intensive, and inconsistent, this study proposes a language-guided defect analysis pipeline that integrates a fine-tuned Qwen3-VL-4B for structured reasoning, a YOLO-SAM hybrid model for object detection, and a fine-tuned SAM3 for prompt-based mask segmentation. By leveraging jointly annotated data—including boxes, polygons, and VQA-style pairs—the framework supports both semantic reasoning and pixel-level localization. Results demonstrate that the fine-tuned Qwen3-VL-4B achieves high reliability in structured prediction, with damage and component accuracy reaching 1 and 0.965, respectively. Comparative analysis indicates that keyword-guided prompting from Qwen3-VL-4B generates more spatially consistent and robust masks than detector-driven prompting, particularly for irregular defects like cracks and corrosion. The proposed framework provides a practical and interpretable direction for task-based bridge inspection and automated defect assessment.

Keywords: Bridge, Maintenance management, Damage detection, VLLMs.

1 Background

Routine maintenance schedule, bridge inspections, as well as post-repair evaluation, is usually labor intensive, and report compilation and maintenance plan consumes scarce time. Computer vision advances have significantly improved detection and estimate of bridge damage, and recent years traditional models for static images are transferring gradually to multi-modal models, such as Visual Question Answering(VQA) and Large Language Model(LLM), as well as random segmentation, such as real-time segment and segment anything.

2 Literature Review

Former researches considered the transfer learning on small-batch samples for task-specified language models in domain data. Also, platforms have been developed for data sharing and edge computation, motivating full digitization of bridge visual inspection.

However, there are still limitations existing. Previous research tends result-oriented evaluation, shows weak linkage with reasoning process, leading to inconsistent and irrelevant output. At the same time, traditional attention module only focus on the rough areas in the image, few for accurate matching text with mask region. The pixel-level segmentation for area specification is required in current inspection process and subsequent investigation.

3 Purpose and Research Method

This study aims to provide a feasible pipeline for damage identification and semantic segmentation by prompt engineering by LLM reasoning. Furthermore, the pixel perception is highlighted for image comprehension capability to locate the targets by visualization, strengthening the result explainability and reproducibility.The pipeline of this study is shown as Figure 1.

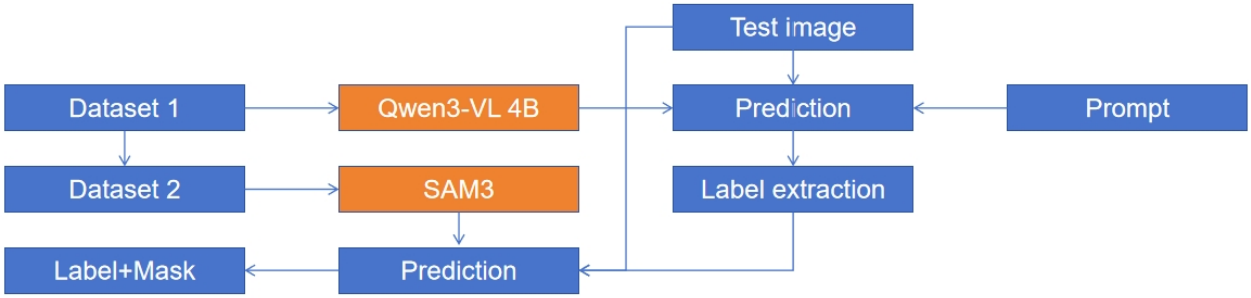


Figure 1. Overview of pipeline

As a multi-modular LLM, Qwen3-VL-4B is fine-tuned by Dataset 1 paired by question and answers with oriented image and test image and prompt are passed to the model weigh using the checkpoint for prediction. The output sentences would be structuralized and the keywords would be extracted for standby. On the other hand, the mask area of subjects of Dataset 1 in relative images are annotated and added to prepare the Dataset 2 paired by labels and masks. Then Dataset 2 is used for fine-tuning of SAM3 and YOLO model when the keywords are used for prompting to get the mask area of damages and members.

Table 1. Definition of damage grade

Grade	Characteristics	Measure
A~B	Fine corrosion, cracks with width less than 1mm or around 2mm but sparsely distributed, small spalling	Record but no special measure
C	Middle grained corrosion, crack with width between 2~3mm, small bricks of spalling, small fracture	Continue to monitor the deterioration process
D~E	Stratified corrosion, dense crossing cracks, large bricks of open-style spalling, severe free lime, water leakage	Appropriate repair or reinforcement

The study divides the common bridge damages into five grades, as shown in Table 1, according to current inspection code in Japan. While Grade A indicates finest defect, Grade E usually implies the heaviest broken for each damage type. At the same time, Grade E damages are always accompanied with malfunction such as deterioration of corrosion resistance and functional impairment of the support system,which should be paid much attention. This grading definition aims to provide a primary diagnosis for partial health status of the bridge based on current findings.

4 Dataset

The Dataset 1 consists of damage images and prompt and template answers is prepared for this task. Typical damages frequently detected on the superstructure except the pavement are used in this dataset, including corrosion, spalling, fracture, bolt detachment, free lime and water leakage. The distribution of the dataset by category is shown in Table 2, as others includes guard rail, bolt and joint and etc.

Table 2. Summary of the Dataset 1

Category	Number	Category	Number
Corrosion	397	Bearing+Cushion	40+39
Spalling	175	Deck Slab	364
Crack	158	Pier	156
Free Lime	207	Girder	292
Water leakage	216	Stiffener	41
Fracture	18	Abutment	169
Bolt Detachment	16	Others	86

For classification task of the components and damage, a dependence rule is established for two purpose, as Table 3 shows. One is conditional constraints. As the normal flatten deck slab is made of concrete, the possible damages occur in life expectation includes such as crack, spalling and free lime. While composite deck slab is with profiled steel sheet at the bottom, acting as module while concrete is poured onside, damages like corrosion and fracture might be observed. Therefore corresponding relationship was formed between the component and damage type by physical properties. Anther one is enriched description. Instead of abstract definition, the structure style and specific material type of component are added, aiming for detail recovery and following study for analysis of mutual effects or load transfer mechanism.

Table 3. Dependence relationship

	Father class	Children class	Correlation
Component	Bearing	Bearing Main body Bearing Concrete Cushion	Relative location
	Deck Slab	Flatten Deck Slab Composite Deck Slab with Profiled steel sheet and concrete	Structure style
Material	Steel	Coating Steel Weathering steel	Surface treatment
Damage	Spalling	Normal Spalling(A~E) Structural Penetration(F)	Severity grade

Dataset 1 combines the question style from both of VQA and VLLM, as shown in Table 4. On the one hand, the structural question is raised for explanation of the image, consists of multiple subjects.On the other hand, leading

question is used for modularized learning for single subject, as one known condition has been given as the knowledge background, so the model could concentrate on the specific area.

Table 4. Prompt template

Purpose	Subject	Question	Additional conditions
Leading question	Damage type	What is the damage type detected on {Member}?	Reply with only damage type label.
	Component type	Which component is affected by {Damage} in this image? .	Reply with only component label
	Severity grade	What is the damage grade on the {Member}?	Reply with one of A/B/C/D/E/unknown
Structural question	Joint	Provide a concise structured answer with exactly three lines: damage_type: <label>\component: <label>\grade: <label>	

For each image in the database, the structural question is attached and one leading question is selected randomly from three question templates. This aims to improve output consistency and generalization with minimal data expansion. The structured prompt stabilizes parseable field extraction, while guided-prompt randomness reduces wording overfitting and strengthens field-specific recognition.

5 Research Method

In this study, built the fine-tuning set from the latest finished image directories rather than relying on legacy JSON paths, because images had been moved and new categories were added. To avoid train-validation leakage, we performed image-level splitting so that all QA variants of the same image stayed in the same split. Besides, domain rules are encoded for controlled equivalence mappings for selected component aliases. samples with invalid labels or missing images were removed, and targeted manual fixes were applied for known annotation edge cases.

To mitigate long-tail effects, we used targeted oversampling at the image-group level. In the final setting, minority damage classes such as fracture, bolt detachment, Spalling were upsampled to a fixed target count, and Water Leakage was moderately upsampled to a higher but controlled count. This strategy improved minority exposure without collapsing the distribution into fully uniform sampling, preserving realism for frequent classes. The final processed number of the data is 2812 pieces.

Qwen3-VL-4B is trained by using Low-Rank Adaptation(LoRA) under a Parameter-Efficient Fine Tuning(PEFT) setup on Oumi Platform. Instead of updating all model parameters, this study injected low-rank adapters into key projection modules in the language stack, while keeping 96% pretrained weights frozen, including frozen visual layers in the setting. Typical LoRA hyperparameters included low-rank dimension, scaling factor, and non-zero adapter dropout for regularization. Training used mixed precision and gradient checkpointing to reduce memory usage.

Besides traditional metrics, the fine tuned weights are used for evaluation on validation data. Deterministic greedy decoding is adopted and the semantic accuracy is calculated by Eq. (1):

$$Acc = \frac{N_{Correct} + N_{Exempt}}{N_{val}} \tag{1}$$

In Eq. (1), the $N_{correct}$ means the number of data pieces that hits the truth answer, N_{exempt} means that was false but exempted by rules, such as results with specified grade or no specific classification standard.

SAM3 is fine tuned under light LoRA configuration as well, and m-size weight of YOLO26 is used for fine-tuning of YOLO-SAM model for most common types of five damages as Dataset 2. For former model, Detection Transformer(DETR), text encoder and mask decoder are mainly applied, while keeping the vision encoder frozen for stability. Training runs with mixed precision, small batch size plus gradient accumulation, AdamW optimization, and cosine learning-rate scheduling for 50 epoches. The latter model is run under default settings.

6 Results

The fine-tuned Qwen3-VL-4B model shows strong and stable performance as results are shown in Table 5. Training loss converged to 0.0490, while validation loss reached 0.1081. Mean token accuracy (MTA) was 0.9798 on training data and 0.9673 on validation data, indicating stable optimization with a limited gap between train and validation.

Table 5. Quantified metrics of Qwen model

Class	Item	Value	Item	Value
Original metrics	Loss	0.0490	Eval_loss	0.1081
	MTA	0.9798	Eval MTA	0.9673
Customized metrics	Damage accuracy	1.0000	Component accuracy	0.9651
	Grade accuracy	0.9535	Joint accuracy	0.9244

For structured task evaluation, damage accuracy reached 1, component accuracy reached 0.9651, grade accuracy reached 0.9535, and joint accuracy reached 0.9244. These results indicate highly reliable damage-type recognition and robust component and grade prediction. The lower joint score is expected because joint correctness requires all fields to be simultaneously correct. Overall, the results support the effectiveness of the data cleaning, normalization, and class-balancing pipeline for multi-attribute bridge damage prediction.

A qualitative example supports the quantitative findings as shown in Figure 2. On a representative bridge image, the fine-tuned model predicts “Corrosion, Bearing, Grade E” and generates the actionable comment “Repair needed,” while the reasoning output identifies relevant structural regions. The consistency between structured labels and textual rationale suggests that the model provides reliable and interpretable outputs for practical inspection workflows.

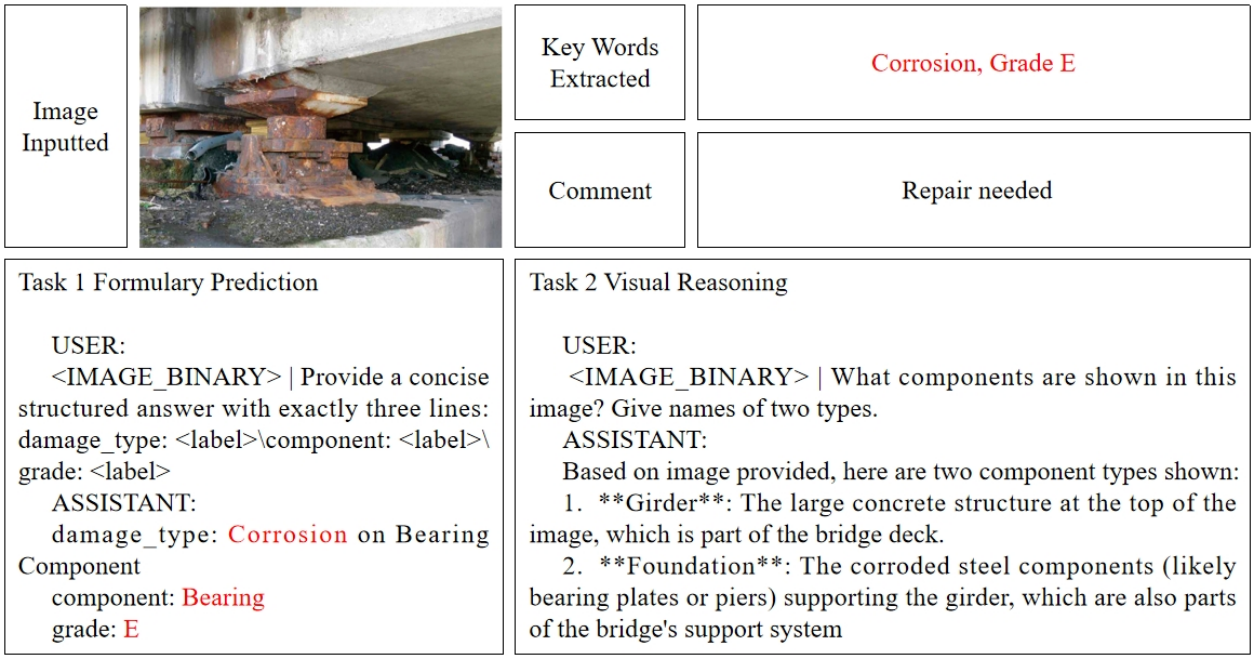


Figure 2. Output of Qwen model

The fine-tuning results of YOLO-SAM indicate moderate overall performance,shown as Figure 3 with clear class-wise variability. Among the five defect categories, Water Leakage and Spalling perform relatively better, while Free Lime and Corrosion remain more challenging, mainly due to lower recall and mAP. The confusion matrix further shows that corrosion is frequently confused with background. A likely reason is the annotation strategy for weathering steel components: labels were drawn around the entire component area. Because exposed weathering steel is rarely free of rust, the annotated regions often contain both corroded and non-corroded visual patterns, which weakens class boundaries and increases confusion with other predictions and background regions.

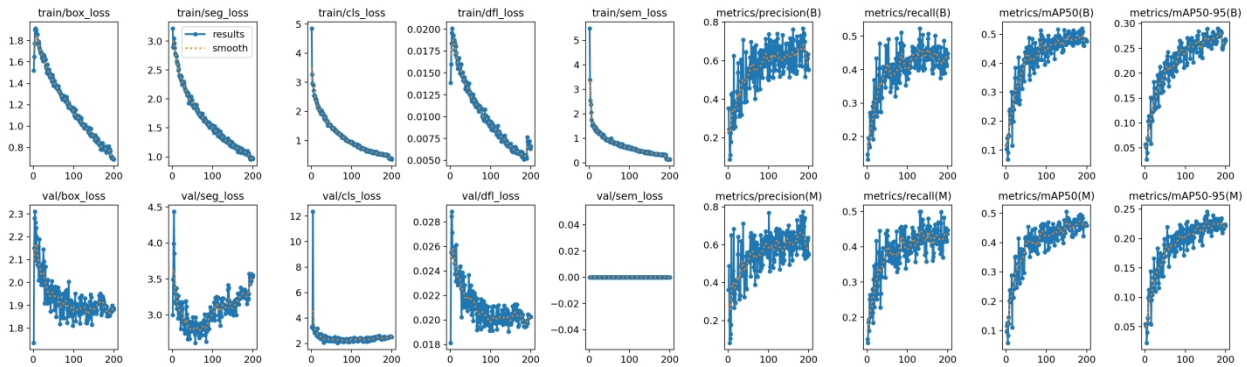


Figure 3. Result of YOLO-SAM hybrid model

SAM3 fine-tuning yields mAP50 to be 0.4251 (>0.4), showing acceptable practical segmentation, whereas mAP50-95, which is less than 0.3, indicates that fine boundary alignment is still challenging. In Figure 4, keyword of damage type guided by prompting successfully highlights the principal corrosion area on the bearing, producing an inspection-relevant mask with poly-curved edges. This supports using the model for defect localization, with future work for type extension and prioritization.

Key Words Injected	Corrosion	Additional information	Grade E

Task 3 Mask Segmentation USER: <IMAGE_BINARY> | Where is the corrosion on the bearing? Provide a exact masked annotation. ASSISTANT: Sure, it is <SEG>	Image outputted	

Figure 4. Output of Qwen-SAM3 mixed model

Figure 5 compares two prompt-generation pipelines for SAM-based defect segmentation. Qwen-SAM3 produces masks that are closer to GT in shape continuity, boundary adherence, and coverage of irregular defect regions. Its predictions better preserve thin, curved, and fragmented structures, especially under low contrast and complex background textures. YOLO-SAM can localize major defect areas, but its masks are generally coarser, with more boundary leakage and occasional background over-segmentation for unnecessary context. Overall, the visual evidence indicates that Qwen-SAM3 provides more reliable fine-grained segmentation for structural defect assessment, while YOLO-SAM remains competitive for fast, coarse localization of damage regions.

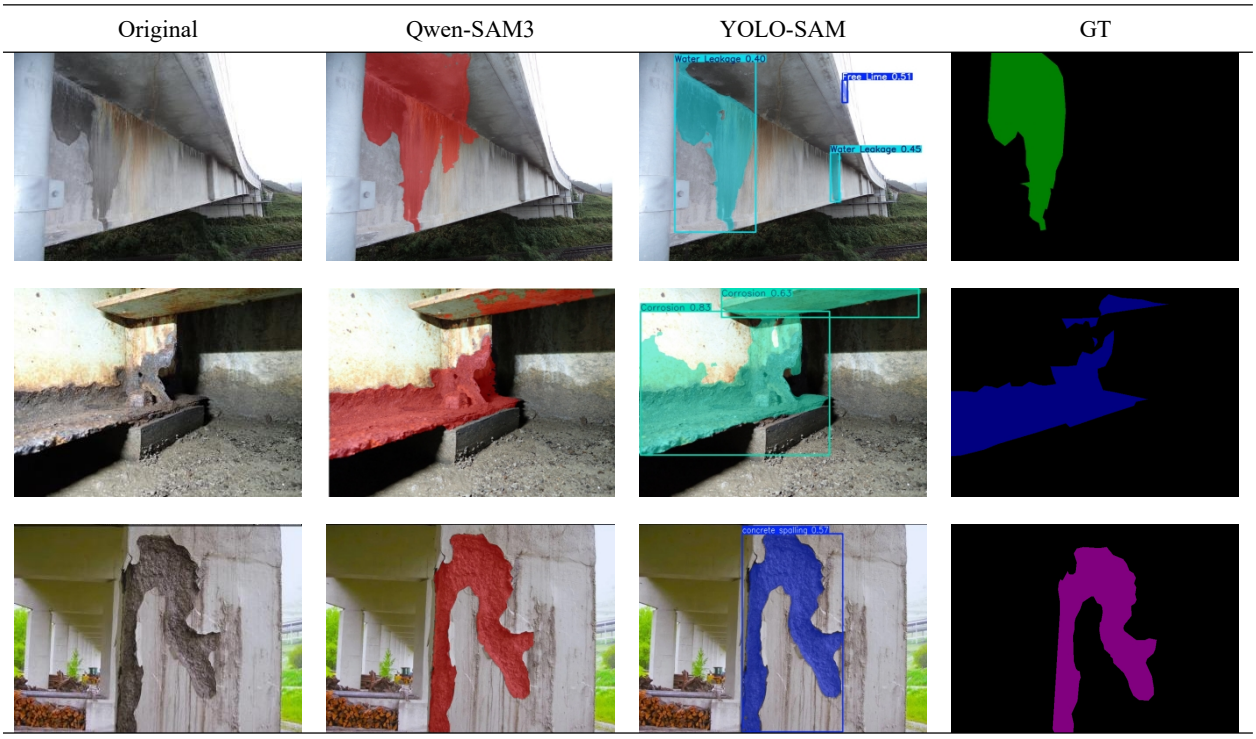
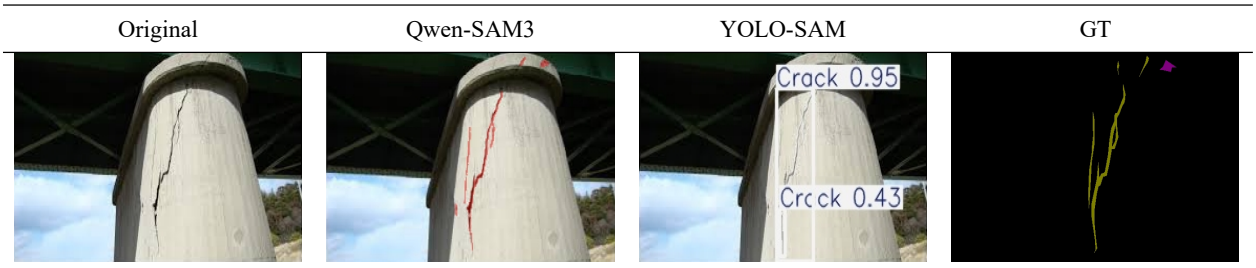


Figure 5. Comparison of different models

Shown as Figure 6, a notable failure case is the free lime sample at case 6-2, where YOLO-SAM does not produce a valid segmented region, while Qwen-SAM3 model still returns a plausible mask. This suggests that keyword-guided prompting from outputs of structured Qwen is more reliable than box-dependent prompting for challenging bridge defect tasks.



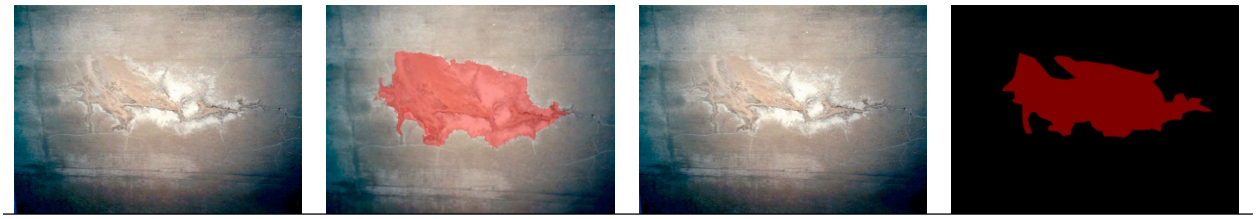


Figure 6. Failure examples

7 Summary

This study developed and evaluated a comprehensive bridge-defect analysis pipeline combining vision-language understanding with prompt-guided segmentation. The fine-tuned Qwen3-VL-4B model exhibits superior performance in extracting inspection-critical semantics, achieving a joint accuracy of 0.9244 across damage, component, and grade fields. While the YOLO-SAM model provides competitive results for fast localization, it shows limitations in classification variability and boundary precision, particularly for free lime and crack due to complex background and similarity interference.

In contrast, the Qwen-SAM3 mixed model produces masks with higher shape continuity and better adherence to ground truth boundaries. The results confirm that structured keyword prompting offers a more robust alternative to traditional box-dependent methods for bridge defect morphologies.

While this study focuses on keyword-driven prompting, future work will explore the integration of more advanced cross-modal segmentation backbones, such as XSAM, to further enhance the alignment between visual features and complex linguistic descriptions. Additionally, upcoming research will focus on incorporating spatial reasoning and causal analysis to evaluate the mutual effects of multiple structural damages, ultimately supporting more comprehensive bridge deterioration prediction.

References

Luo, Kui, et al. "Computer vision-based bridge inspection and monitoring: A review." Sensors 23.18 (2023): 7863.
Zhang, Chenhong, et al. "Automatic bridge inspection database construction through hybrid information extraction and large language models." Developments in the Built Environment 20 (2024): 100549.
Chun, Pang-Jo, Tatsuro Yamane, and Yu Maemura. "A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage." Computer-Aided Civil and Infrastructure Engineering 37.11 (2022): 1387-1401.
Carion, Nicolas, et al. "SAM 3: Segment anything with concepts." arXiv preprint arXiv:2511.16719 (2025).
Bai, Shuai, et al. "Qwen3-vl technical report." arXiv preprint arXiv:2511.21631 (2025).
Sapkota, Ranjan, and Manoj Karkee. "Ultralytics YOLO evolution: An overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 object detectors for computer vision and pattern recognition." arXiv preprint arXiv:2510.09653 (2025).