Angkorian-HeritageObj: AI-Assisted Visual Condition Assessment of Khmer Stone Heritage Objects

Nimol Thuon[1,2*] and Panhapin Theang[2]

[1] University of Science and Technology of China, 230052 Hefei, China
[2] Université Paris Cité, 75013 Paris, France
tnimol@mail.ustc.edu.cn,panhapin.theang@etu.u-paris.fr

Abstract. Khmer stone heritage objects, including walls, carved reliefs, statues, architectural frames, and ornamental surfaces, suffer long-term deterioration from weathering, biological colonisation, cracking, surface deposits, discoloration, material loss, and past repair interventions. This paper studies AI-assisted visual condition assessment as a conservation-oriented computer vision problem for Khmer stone heritage. Unlike inscription transcription or document layout analysis, the goal is to support first-pass screening of visible condition severity and degradation attributes from ordinary RGB photographs. We introduce Angkorian-HeritageObj, a pilot collection of 200 real-world images collected under a small-data heritage setting. Each image is annotated with one severity label (good, moderate, or severe) and seven multi-label damage attributes (erosion, crack, biological growth, missing or broken part, discoloration, surface deposit, and repair mark). We evaluate three method families: (i) frozen foundation features from CLIP and DINOv2 combined with logistic regression or linear SVM, (ii) ImageNet-pretrained CNN transfer learning with EfficientNet-B0 and ResNet50, and (iii) zero-shot vision–language-model (VLM) prompting for structured severity, damage tagging, and condition description. Under five-fold stratified cross-validation, CLIP with logistic regression and EfficientNet-B0 both reach 85.50% accuracy on three-class severity classification, with macro-F1 of 86.33% and 86.20%, respectively. For multi-label damage recognition, CLIP with one-vs-rest logistic regression attains 90.8% micro-F1 and 87.9% macro-F1, whereas a zero-shot VLM yields only 53.0% severity accuracy and 54.7% damage-recognition micro-F1. These results indicate that lightweight foundation-model-based pipelines can support cultural heritage preservation through first-pass condition screening, while generic VLMs should be used as assistive description tools requiring expert review.

Keywords: Cultural heritage · Khmer heritage · Stone deterioration · Visual condition assessment · Foundation models · Vision–language models.

* Corresponding author

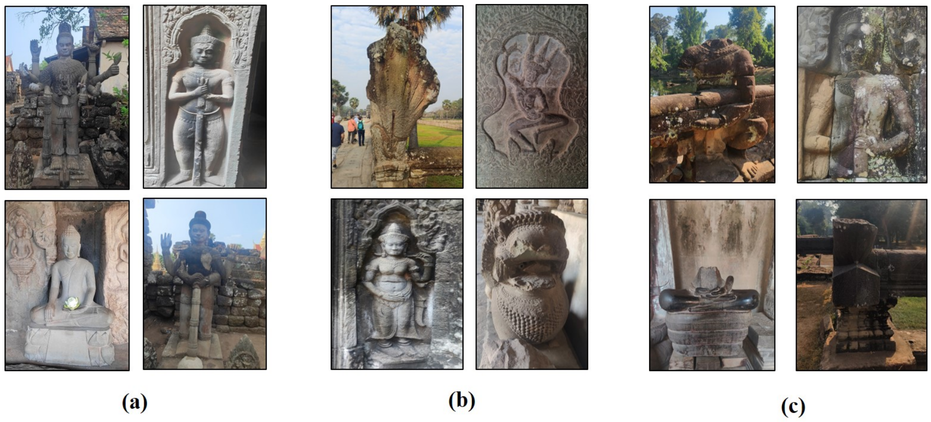

Fig. 1. Representative examples from the Angkorian-HeritageObj pilot collection illustrating the three condition classes: (a) good, (b) moderate degradation, and (c) severe degradation.

1 Introduction

Khmer stone heritage objects, including temple walls, sandstone blocks, lintels, bas-reliefs, statues, and architectural fragments, are central to Cambodia's cultural landscape. The Angkor World Heritage property, inscribed on the UNESCO World Heritage List in 1992, contains a large ensemble of temples and archaeological remains associated with several centuries of Khmer civilisation [23]. These objects are continuously affected by tropical wet–dry cycles, biological growth, salt weathering, cracking, surface deposits, and material loss [20, 1, 7, 22]. The ICOMOS-ISCS glossary provides a shared conservation vocabulary for describing such stone deterioration patterns [24].

Condition assessment is usually performed by trained conservators through field inspection, photographic documentation, and, where available, advanced sensing such as photogrammetry, laser scanning, or 3D reconstruction [18]. Although expert inspection remains essential, it is time-consuming and difficult to scale across large heritage landscapes. This motivates conservation-oriented computer vision methods that can support cultural heritage preservation through first-pass screening, while leaving final interpretation to domain experts. Recent studies have explored computer vision and deep learning as decision-support tools for heritage inspection and damage assessment, with expert review remaining essential [12, 11, 16].

For Khmer cultural heritage, most computational studies have focused on document-like materials, scene text, palm-leaf manuscripts, inscription analysis, or script recognition [10, 13, 14]. These tasks are important but differ from visual condition assessment. A model designed to read text or analyse document layout is not necessarily designed to determine whether a stone object is visually stable, moderately degraded, or severely degraded. This paper therefore focuses on the visible physical condition of Khmer stone heritage objects rather than inscription reading.

We present Angkorian-HeritageObj, a preliminary study on AI-assisted visual condition assessment using a restricted pilot collection of 200 real-world RGB images. The collection is not publicly released at this stage because the images involve culturally sensitive heritage material and access constraints. This restricted small-data heritage setting is common in cultural heritage preservation, where large public datasets are difficult to construct. Each image is annotated with one condition severity label and seven multi-label damage attributes. We evaluate frozen CLIP and DINOv2 features with shallow classifiers, EfficientNet-B0 and ResNet50 transfer learning, and zero-shot Qwen2.5-VL prompting for severity classification, damage recognition, and structured condition description.

The main contributions are:

- We formulate AI-assisted visual condition assessment of Khmer stone heritage objects as a conservation-oriented computer vision problem for first-pass condition screening, distinct from inscription transcription and document layout analysis.
- We construct a restricted 200-image pilot collection with image-level labels for three-class severity and seven visible damage attributes informed by the ICOMOS-ISCS vocabulary.
- We benchmark foundation features, CNN transfer learning, and zero-shot VLM prompting under a restricted small-data heritage setting, and analyse their class-wise and damage-specific behaviour.

2 Related Work

2.1 Stone Deterioration and Conservation Terminology

The weathering of Khmer sandstone has been studied for several decades through international conservation programmes. Previous studies at Angkor have reported contour scaling, sanding, biological colonisation, salt weathering, material loss, and stone recession under tropical climate conditions [20, 1, 2]. Salt-weathering and petrophysical studies further show that groundwater, bat guano, porosity, and capillary behaviour can contribute to the deterioration of sandstone monuments [7, 22, 25]. These studies motivate the visible condition attributes considered in this paper, including erosion, biological growth, cracks, missing parts, discoloration, surface deposits, and repair marks.

Conservation practice also depends on consistent terminology. The ICOMOS-ISCS illustrated glossary [24] provides standardised terms and visual examples for stone deterioration patterns such as cracking, deformation, detachment, material loss, discoloration, deposits, and biological colonisation. Our work adapts this conservation vocabulary to a lightweight computer-vision setting. Rather than attempting expert-level diagnosis, we annotate image-level attributes that are visually identifiable from ordinary RGB photographs.

2.2 AI-Assisted Visual Inspection for Cultural Heritage

AI-assisted visual inspection is increasingly used to support cultural heritage documentation and monitoring. Recent reviews highlight the role of computer vision and deep learning in damage assessment, anomaly detection, and inspection workflows, while emphasising that AI outputs should remain decision-support evidence for expert interpretation [12, 11, 16]. Several studies have applied object detection or segmentation to heritage deterioration, including Mask R-CNN-based stone deterioration mapping, gold-foil shedding detection, limestone spalling detection, and surface-defect recognition on heritage buildings [5, 8, 4, 19].

These works demonstrate the potential of deep learning for cultural heritage inspection, but they often require larger annotated datasets, focus on specific damage types, or address masonry contexts different from Khmer sandstone objects. Khmer stone heritage presents a distinct visual setting because of tropical weathering, sandstone texture, biological growth, carved ornamentation, and the coexistence of walls, reliefs, statues, and architectural fragments. To our knowledge, no prior work has benchmarked image-level condition severity and multi-label damage recognition specifically for Khmer stone heritage under a small-data restricted-access setting.

2.3 Pretrained Vision Models and Vision–Language Models

Pretrained visual representations are attractive for small-data heritage applications because expert annotations are costly to obtain. CLIP learns transferable visual representations from natural language supervision and has shown strong zero-shot and linear-probe transfer [17], while DINOv2 provides robust self-supervised visual features that can be used without task-specific fine-tuning [15]. Conventional transfer learning also remains an important baseline, with ResNet and EfficientNet widely used for image classification under limited-data conditions [6, 21].

Vision–language models (VLMs) can generate structured visual outputs such as labels, captions, and short descriptions. Qwen2.5-VL, for example, supports visual recognition and structured visual understanding [3]. Recent inspection studies suggest that VLMs are useful in zero-shot or few-shot settings, but their predictions are often poorly calibrated for specialised damage categories [9, 26, 27]. We therefore evaluate VLMs not only as classifiers but also as assistive description tools. Related historical document and palm-leaf manuscript benchmarks address important Khmer and Southeast Asian cultural materials, but they mainly target text recognition or document analysis rather than the physical condition of stone heritage [13, 10, 14].

3 Pilot Collection and Annotation Protocol

3.1 Scope and Data Governance

The Angkorian-HeritageObj pilot collection contains 200 RGB images of Khmer stone heritage objects captured under real-world temple conditions. The collection was developed as part of the AngkorianAI project[3]. The images cover walls, stone blocks, carved reliefs, statues, architectural frames, and other stone surfaces. The collection is restricted and is not publicly released at this stage because the images involve culturally sensitive heritage sites and access constraints. To reduce location-related sensitivity, exact site names and GPS coordinates are excluded, and each image is instead associated with an anonymised site group. The collection is intended as a small-scale feasibility study rather than a comprehensive public benchmark. Figure 1 shows representative examples from the three condition classes used in the pilot collection.

3.2 Annotation Design

The annotation protocol is image-level and intentionally lightweight, reflecting the preliminary nature of the study. Pixel-level masks and bounding boxes are not used; instead, each image receives one overall condition severity label and seven binary damage-attribute labels. The severity label describes the overall visual condition of the object, while the damage attributes provide multi-label evidence for visible degradation. Thus, one image is assigned exactly one severity class but may contain several damage attributes simultaneously.

Images were first organised into rough condition folders for initial sorting. The final labels used in all experiments were stored in a CSV annotation file rather than inferred from folder names. This allows initially misplaced images to be corrected during manual review without physically moving the image files. Each CSV row contains the image identifier, image path, anonymised site group, object type, condition severity, seven binary damage attributes, and optional notes for ambiguous cases. Before running the experiments, the annotation file was checked for valid severity labels, valid binary damage values, duplicate entries, and missing image paths.

For dataset statistics, Table 1 summarises the severity and damage-attribute distributions. The final severity labels comprise 51 good-condition images, 81 moderately degraded images, and 68 severely degraded images. Since the damage attributes are multi-label, their counts are not mutually exclusive. The most frequent attributes are erosion/weathering, biological growth, and dirt or surface deposit, each appearing in more than 140 images.

3.3 Operational Label Definitions

For label definition, Table 2 defines the severity classes and damage attributes used during annotation. The severity label is an image-based screening label,

[3] https://angkorianai.github.io/

Table 1. Dataset and annotation distribution in the Angkorian-HeritageObj pilot collection. Damage attributes are multi-label and therefore not mutually exclusive.

Group	Label / attribute	Count
Severity label	Good	51
	Moderate degradation	81
	Severe degradation	68
	Total	**200**
Damage attribute	Erosion / weathering	148
	Crack / fracture	72
	Biological growth	147
	Missing / broken part	60
	Discoloration / stain	56
	Dirt / surface deposit	148
	Repair mark	68

not a formal structural diagnosis. For example, an image is labelled moderate if damage is clearly visible but the object remains interpretable, and severe if deterioration strongly affects interpretability or if missing parts, heavy biological growth, extensive erosion, or major cracking dominate the image. The damage attributes are derived from the ICOMOS-ISCS glossary [24] but simplified to be recognisable from a single photograph.

4 Task Definition

We define three tasks. First, *condition severity classification* predicts one label, good, moderate, or severe, from an RGB image x_i. This task evaluates the overall visible preservation status of a stone heritage object. Second, *multi-label damage recognition* predicts a binary vector $\mathbf{a}_i \in \{0,1\}^7$ over the seven damage attributes in Table 2. This task is multi-label because one image may contain several degradation cues simultaneously, such as erosion, biological growth, and surface deposits. Third, *VLM-based structured condition description* asks a vision–language model to return one severity label, a list of visible damage attributes, and a short condition description from an image and prompt.

For severity classification, we report accuracy, macro-F1, and weighted-F1. For multi-label damage recognition, we report micro-F1, macro-F1, mean average precision (mAP), and per-class F1. For the VLM setting, the predicted severity and damage attributes are evaluated quantitatively, while the generated descriptions are analysed qualitatively.

5 Methods

For the experimental pipeline, Figure 2 summarises the three method families evaluated in this study: foundation features with shallow classifiers, CNN transfer learning, and zero-shot VLM prompting.

Table 2. Severity labels and damage attributes used in Angkorian-HeritageObj.

Label	Definition
Condition severity labels	
Good	Mostly preserved; details remain clear with little visible damage.
Moderate	Visible damage is present, but the main form remains interpretable.
Severe	Heavy deterioration, missing parts, or unclear object details.
Binary damage attributes	
Erosion / weathering	Surface wear, rounded edges, or loss of detail.
Crack / fracture	Linear split, fissure, or visible fracture.
Biological growth	Moss, lichen, algae, roots, or vegetation-like coverage.
Missing / broken part	Lost, broken, incomplete, or detached stone element.
Discoloration / stain	Uneven colour, darkening, whitening, or stain.
Dirt / surface deposit	Dust, soil, mineral crust, or foreign surface material.
Repair mark	Cement-like fill, patching, joint repair, or restoration trace.

5.1 Foundation Features with Shallow Classifiers

For feature-based baselines, each image is encoded with a pretrained model and classified using a shallow classifier. We evaluate CLIP ViT-B/32 image features [17] and DINOv2-small features [15]. For severity classification, features are fed into logistic regression and linear SVM classifiers; for multi-label damage recognition, into one-vs-rest logistic regression. All classifiers use class balancing when applicable.

5.2 CNN Transfer Learning

EfficientNet-B0 [21] and ResNet50 [6] are used as supervised transfer-learning baselines. Both models are initialised with ImageNet-pretrained weights. To reduce overfitting on the 200-image pilot collection, the backbone is frozen and only the final classification layer is trained, with class-balanced cross-entropy loss for severity classification.

5.3 Zero-Shot VLM Prompting

The VLM baseline uses Qwen2.5-VL [3] in a zero-shot setting. The prompt asks the model to classify the condition as good, moderate, or severe; identify damage attributes from a fixed list; and return a one-sentence description. Figure 3 presents the structured prompt used for evaluation. Strict output formatting reduces parsing errors, and the VLM produces valid structured predictions for all 200 images.

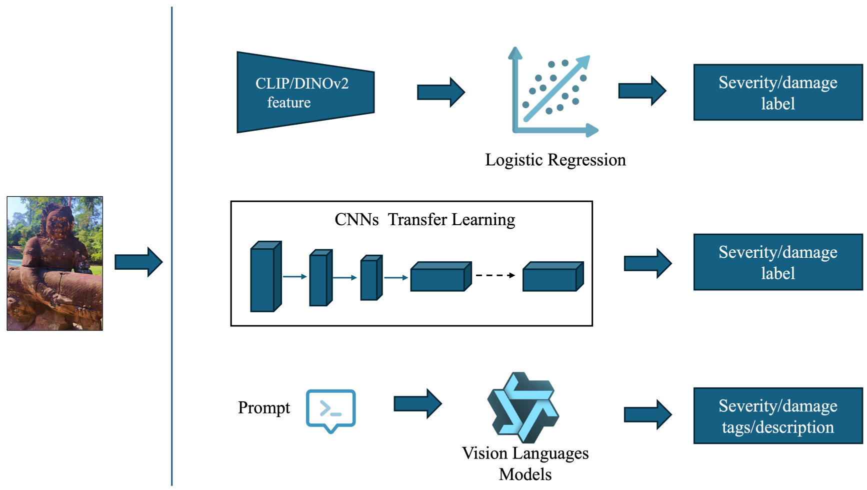

Fig. 2. Overview of the AI-assisted condition assessment pipeline. Frozen CLIP and DINOv2 features are combined with shallow classifiers for severity classification and damage recognition. EfficientNet-B0 and ResNet50 are evaluated as CNN transfer-learning baselines, while Qwen2.5-VL is used for zero-shot structured condition description.

6 Experimental Protocol

6.1 Data Split and Evaluation

All supervised and feature-based classification experiments use five-fold cross-validation. For severity classification, stratified folds preserve the class distribution of good, moderate, and severe samples. Each fold uses 160 images for training and 40 for testing. We report the mean and standard deviation across folds. For damage attribute recognition, the same five-fold setting is used with one-vs-rest logistic regression on top of the frozen features. Table 5 reports the mean and standard deviation of fold-level scores, while the per-class results in Table 6 are computed from pooled out-of-fold predictions across all folds. The VLM is evaluated on all 200 images in a zero-shot setting, without training or calibration. Therefore, the VLM tables report single aggregate scores rather than cross-validation mean and standard deviation.

6.2 Implementation Details

CLIP and DINOv2 features are L_2-normalised before classification. Logistic regression is trained with a maximum of 5000 iterations. Linear SVM uses class-balanced weighting. EfficientNet-B0 and ResNet50 are trained for up to 100 epochs per fold with early stopping based on validation macro-F1. Data augmentation includes resized cropping, horizontal flipping, mild rotation, and colour

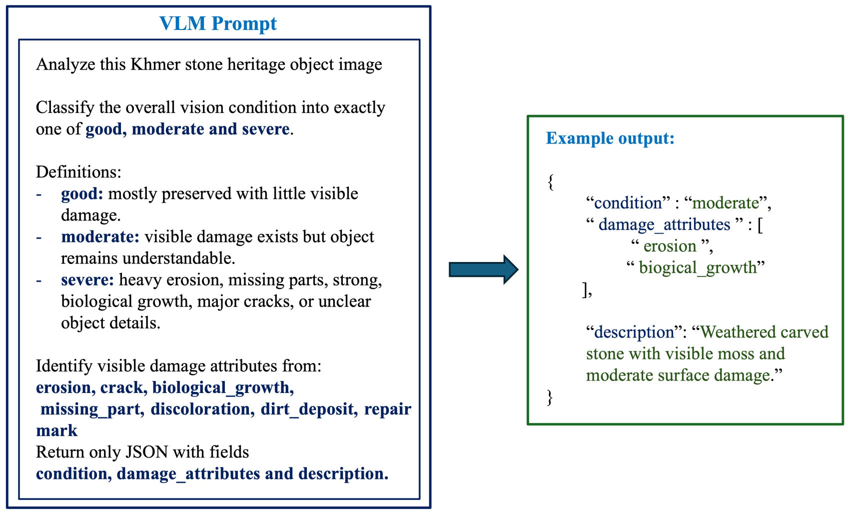

Fig. 3. Structured prompt template used for zero-shot VLM condition assessment with Qwen2.5-VL. The constrained label set and JSON-style output format allow direct comparison between VLM predictions and human annotations.

jitter. The batch size is 16. The CNN backbone is frozen and only the final classification layer is optimised. The VLM uses a reduced image resolution to fit within GPU memory and generates deterministic outputs with sampling disabled.

7 Results

7.1 Condition Severity Classification

CLIP with logistic regression achieves the strongest severity classification performance, reaching 0.855 accuracy and 0.863 macro-F1. EfficientNet-B0 matches the best accuracy at 0.855 and obtains a very close macro-F1 of 0.862, showing that lightweight transfer learning is also effective in this small-data setting. DINOv2-based classifiers remain competitive, while ResNet50 performs slightly lower. In contrast, the zero-shot VLM is much weaker, with 0.530 accuracy and 0.526 macro-F1, indicating that generic VLM prompting is not sufficient for reliable severity classification.

Class-wise results further show that EfficientNet-B0 performs best on the good class (F1 = 0.9320), while moderate and severe remain more difficult. The VLM obtains high recall for good cases (0.9608) but low recall for severe cases (0.3235), suggesting that it often underestimates damage severity. This pattern supports the need for task-specific supervision rather than relying on zero-shot VLM predictions alone.

Table 3. Condition severity classification results. Feature-based and CNN baselines are evaluated using five-fold stratified cross-validation; the VLM is evaluated zero-shot on all 200 images.

Method	Acc.	Macro-F1	Weighted-F1
CLIP + Log. Reg.	**0.855 ± 0.068**	**0.863 ± 0.063**	**0.856 ± 0.067**
CLIP + Linear SVM	0.845 ± 0.058	0.853 ± 0.054	0.846 ± 0.057
DINOv2 + Log. Reg.	0.840 ± 0.060	0.848 ± 0.055	0.841 ± 0.059
DINOv2 + Linear SVM	0.830 ± 0.051	0.839 ± 0.046	0.831 ± 0.050
EfficientNet-B0	**0.855 ± 0.037**	0.862 ± 0.033	0.855 ± 0.037
ResNet50	0.825 ± 0.045	0.830 ± 0.046	0.822 ± 0.046
VLM zero-shot	0.530	0.526	0.508

Table 4. Class-wise severity classification results for representative methods. Support is the number of images per class.

Method	Class	Precision	Recall	F1-score	Support
EfficientNet-B0	Good	0.9231	0.9412	0.9320	51
	Moderate	0.8023	0.8519	0.8263	81
	Severe	0.8710	0.7941	0.8308	68
ResNet50	Good	0.8448	0.9608	0.8991	51
	Moderate	0.8243	0.7531	0.7871	81
	Severe	0.8088	0.8088	0.8088	68
VLM zero-shot	Good	0.5158	0.9608	0.6712	51
	Moderate	0.4430	0.4321	0.4375	81
	Severe	0.8462	0.3235	0.4681	68

7.2 Multi-Label Damage Attribute Recognition

For damage-attribute recognition, CLIP with one-vs-rest logistic regression achieves the best F1 performance, with 0.908 micro-F1 and 0.879 macro-F1. DINOv2 performs similarly and obtains the highest mAP. The zero-shot VLM is substantially weaker, reaching only 0.547 micro-F1 and 0.504 macro-F1. Its outputs are relatively precise when damage tags are produced, but many visible attributes are missed, especially for repair marks and dirt deposits.

Erosion, biological growth, and dirt or surface deposits are recognised most reliably by the supervised feature-based models. These attributes often cover large regions and produce strong texture or colour cues. Missing parts, discoloration, and repair marks are more challenging because they require object context and conservation-specific interpretation. The VLM performs reasonably on erosion and cracks, but performs poorly on repair marks and dirt deposits.

7.3 Qualitative Analysis

The qualitative examples in Fig. 4 confirm the quantitative trend. CLIP and EfficientNet-B0 recover both severity and major damage attributes in represen-

Table 5. Multi-label damage attribute recognition results. CLIP and DINOv2 use one-vs-rest logistic regression under five-fold cross-validation; the VLM is evaluated zero-shot.

Method	Micro-F1	Macro-F1	mAP
CLIP + OVR Log. Reg.	**0.908 ± 0.021**	**0.879 ± 0.032**	0.914 ± 0.034
DINOv2 + OVR Log. Reg.	0.902 ± 0.028	0.868 ± 0.037	**0.917 ± 0.035**
VLM zero-shot	0.547	0.504	–

Table 6. Per-class F1 scores for multi-label damage recognition. "Pos." denotes the number of positive samples for each damage attribute.

Damage attribute	Pos.	CLIP	DINOv2	VLM
Erosion / weathering	148	0.966	0.966	0.782
Crack / fracture	72	0.870	0.818	0.692
Biological growth	147	0.960	0.963	0.513
Missing / broken part	60	0.806	0.797	0.552
Discoloration / stain	56	0.821	0.800	0.460
Dirt / surface deposit	148	0.943	0.947	0.324
Repair mark	68	0.809	0.809	0.208

tative good, moderate, and severe cases. The zero-shot VLM, however, tends to downgrade moderate or severe degradation when the object structure remains visually clear.

In the moderate example, the VLM predicts good and returns only erosion, missing biological growth and dirt deposit. In the severe example, it predicts moderate and recovers only part of the visible damage evidence. These failures are consistent with Tables 3 and 5, where zero-shot VLM performance is much lower than feature-based and transfer-learning baselines. This suggests that VLM descriptions may be useful for initial documentation, but their outputs require expert review before conservation-oriented use.

8 Discussion

8.1 Small-Data Learning Is Feasible

A small, carefully annotated image collection can support meaningful AI-assisted visual condition assessment. CLIP, DINOv2, and EfficientNet-B0 all achieve strong severity classification performance despite the limited data size, supporting the use of pretrained representations for conservation-oriented tasks where data collection is expensive and public release may be restricted. This is consistent with broader observations in industrial and structural-inspection settings where pretrained vision features or VLM components have been shown to transfer well to few-shot defect recognition [9, 26].

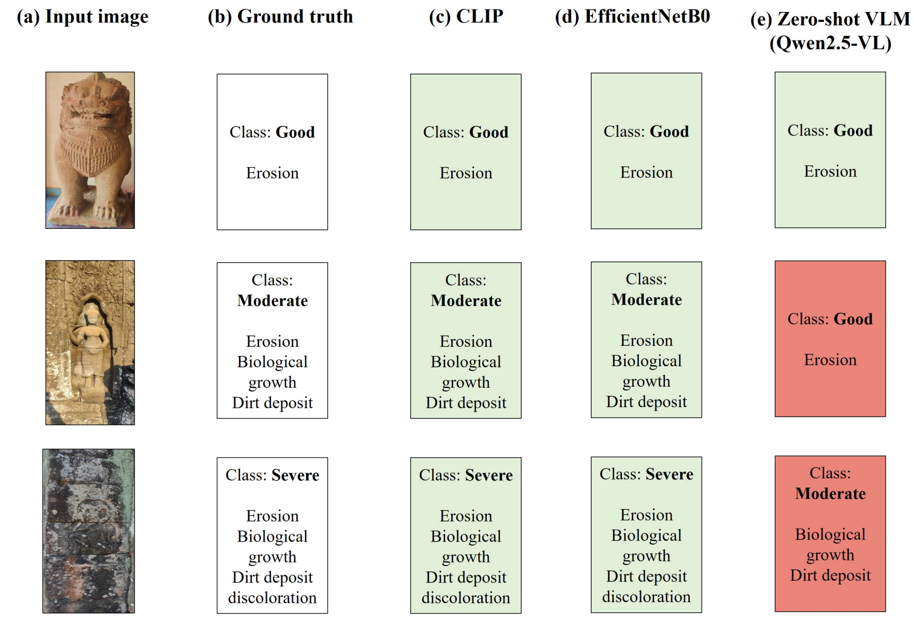

Fig. 4. Qualitative comparison of condition assessment results on three representative cases. Columns show (a) input images, (b) ground-truth annotations, (c) CLIP-based predictions, (d) EfficientNet-B0 predictions, and (e) zero-shot Qwen2.5-VL predictions. Green boxes indicate predictions consistent with the ground truth; red boxes indicate incorrect or incomplete predictions.

8.2 Damage Attributes Are Easier When Visual Cues Are Large

The multi-label damage results show especially strong performance for erosion, biological growth, and dirt deposits. These labels often correspond to broad colour or texture cues. In contrast, missing parts and repair marks require object-level reasoning: a missing finger, a broken edge, or a repair patch is obvious to a human familiar with the expected object shape but is less obvious to a model without contextual knowledge. Future work should therefore explore object-aware models, part-based annotations, expert-guided prompts, or retrieval grounding similar to recent VLM-with-RAG approaches in industrial inspection [27].

8.3 VLMs Are Useful but Not Sufficient

The VLM results show both promise and limitation. The model produces valid structured outputs for all images and can deliver short, human-readable descriptions, but its zero-shot classification scores are substantially below those of supervised feature-based baselines. It strongly favours the good class and misses many severe cases, suggesting a tendency to underestimate visual degradation in the absence of in-context examples or calibration. For practical conservation workflows, VLMs are therefore best treated as assistive description tools rather than final decision-makers, and prompts should be designed to make the model conservative about the good label and to encourage explicit reasoning about extent and dominance of damage.

8.4 Limitations

The main limitation is dataset size. Although 200 images are sufficient for a preliminary workshop study, they do not cover the full diversity of Khmer stone heritage conditions. The labels are image-level and do not localise damage regions; ordinary RGB images cannot determine subsurface deterioration, material chemistry, or structural stability. The severity labels are visual screening categories, not expert conservation diagnoses. Finally, the cross-validation protocol may be optimistic if visually similar images from the same object appear in different folds; future work will use object-level or site-level splits when more data become available.

9 Conclusion

This paper presented a preliminary study on AI-assisted visual condition assessment of Khmer stone heritage objects. Using a restricted pilot collection of 200 images, we defined image-level severity labels and seven multi-label damage attributes informed by the ICOMOS-ISCS vocabulary. Experiments show that CLIP and DINOv2 features with shallow classifiers, as well as EfficientNet-B0 transfer learning, are effective under small-data conditions: CLIP with logistic regression and EfficientNet-B0 both reach 85.50% accuracy for three-class severity classification, and CLIP with one-vs-rest logistic regression reaches 90.76% micro-F1 for multi-label damage recognition. Zero-shot VLM prompting produces valid structured descriptions but markedly weaker quantitative performance, indicating that VLMs are useful for assistance but require task-specific calibration and expert review.

The study suggests that lightweight AI methods can support first-pass visual screening of Khmer stone heritage condition. Future work will expand the collection, improve annotation reliability through double-annotation and Cohen's κ agreement, introduce site-level evaluation, explore damage localisation with bounding boxes or masks, and investigate human-in-the-loop workflows for conservation documentation.

References

1. André, M.F., Etienne, S., Mercier, D., Vautier, F., Voldoire, O.: Assessment of sandstone deterioration at ta keo temple (angkor): first results and future prospects. Environmental Geology **56**(3–4), 677–688 (2008). https://doi.org/10.1007/s00254-008-1408-8

2. André, M.F., Voldoire, O., Roussel, E., Vautier, F., Phalip, B., Peou, H.: Contrasting weathering and climate regimes in forested and cleared sandstone temples of the angkor region. Earth Surface Processes and Landforms **37**(5), 519–532 (2012). https://doi.org/10.1002/esp.2265
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025), `https://arxiv.org/abs/2502.13923`
4. Fu, X., Angkawisittpan, N.: Detecting surface defects of heritage buildings based on deep learning. Journal of Intelligent Systems **33**(1), 20230048 (2024)
5. Hatır, M.E., İnce, İ., Korkanç, M.: Intelligent detection of deterioration in cultural stone heritage. Journal of Building Engineering **44**, 102690 (2021). https://doi.org/10.1016/j.jobe.2021.102690
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Hosono, T., Uchida, E., Suda, C., Ueno, A., Nakagawa, T.: Salt weathering of sandstone at the angkor monuments, cambodia: identification of the origins of salts using sulfur and strontium isotopes. Journal of Archaeological Science **33**(11), 1541–1551 (2006). https://doi.org/10.1016/j.jas.2006.01.018
8. Hou, M., Huo, D., Yang, Y., Yang, S., Chen, H.: Using mask r-cnn to rapidly detect the gold foil shedding of stone cultural heritage in images. Heritage Science **12**(1), 1–14 (2024). https://doi.org/10.1186/s40494-024-01158-9
9. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19606–19616 (2023)
10. Kesiman, M.W.A., Valy, D., Burie, J.C., Paulus, E., Suryani, M., Hadi, S., Verleysen, M., Chhun, S., Ogier, J.M.: Icfhr 2018 competition on document image analysis tasks for southeast asian palm leaf manuscripts. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 483–488. IEEE (2018)
11. Mishra, M., Barman, T., Ramana, G.V.: Artificial intelligence-based visual inspection system for structural health monitoring of cultural heritage. Journal of Civil Structural Health Monitoring **14**(1), 103–120 (2024). https://doi.org/10.1007/s13349-023-00683-8
12. Mishra, M., Lourenço, P.B.: Artificial intelligence-assisted visual inspection for cultural heritage: state-of-the-art review. Journal of Cultural Heritage **66**, 536–550 (2024). https://doi.org/10.1016/j.culher.2024.01.005
13. Nikolaidou, K., Seuret, M., Mokayed, H., Liwicki, M.: A survey of historical document image datasets. International Journal on Document Analysis and Recognition (IJDAR) **25**(4), 305–338 (2022)
14. Nom, V., Bakkali, S., Luqman, M.M., Coustaty, M., Ogier, J.M.: Khmerst: a low-resource khmer scene text detection and recognition benchmark. In: Proceedings of the Asian Conference on Computer Vision. pp. 1777–1792 (2024)
15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

16. Pathirage, C.S.N., Li, J., Li, L., Hao, H., Liu, W., Ni, P.: Structural damage identification based on autoencoder neural networks and deep learning. Engineering Structures **172**, 13–28 (2018). https://doi.org/10.1016/j.engstruct.2018.05.109
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Rakha, T., Gorodetsky, A.: Review of unmanned aerial system (uas) applications in the built environment: towards automated building inspection procedures using drones. Automation in Construction **93**, 252–264 (2018). https://doi.org/10.1016/j.autcon.2018.05.002
19. Saravanan, T.J., Bhaskar, B.: Automated evaluation of degradation in stone heritage structures utilizing deep vision in synthetic and real-time environments. Journal of Building Engineering **98**, 111117 (2024)
20. Siedel, H., Pfefferkorn, S., von Plehwe-Leisen, E., Leisen, H.: Sandstone weathering in tropical climate: results of low-destructive investigations at the temple of angkor wat, cambodia. Engineering Geology **115**(3–4), 182–192 (2010). https://doi.org/10.1016/j.enggeo.2009.06.001
21. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
22. Uchida, E., Ogawa, Y., Maeda, N., Nakagawa, T.: Deterioration of stone materials in the angkor monuments, cambodia. Engineering Geology **55**(1–2), 101–112 (1999). https://doi.org/10.1016/S0013-7952(99)00110-3
23. UNESCO World Heritage Centre: Angkor. World Heritage List, Site 668, `https://whc.unesco.org/en/list/668/`, last accessed 2026
24. Vergès-Belmin, V. (ed.): ICOMOS-ISCS Illustrated Glossary on Stone Deterioration Patterns, Monuments and Sites, vol. XV. ICOMOS, Paris (2008), `https://openarchive.icomos.org/id/eprint/2073/`
25. Wedekind, W., Gross, C.J., Hoffmann, A., Siegesmund, S.: Damage phenomenon and petrophysical properties of sandstones at the phnom bakheng temple (angkor, cambodia): first investigations and possible conservation measures. Environmental Earth Sciences **77**(21), 735 (2018)
26. Yong, G., Jeon, K., Gil, D., Lee, G.: 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 (2023)
27. Zhang, Y., Zhou, Q., Imani, F., Tang, J.: Seeing the unseen: Towards zero-shot inspection for wind turbine blades using knowledge-augmented vision language models (2025)