

Seeing the Unseen: Towards Zero-Shot Inspection for Wind Turbine Blades using Knowledge-Augmented Vision Language Models

Yang Zhang, Postdoc Researcher

School of Mechanical, Aerospace and Manufacturing Engineering
University of Connecticut

Qianyu Zhou, Ph.D. Student

School of Mechanical, Aerospace and Manufacturing Engineering
University of Connecticut

Farhad Imani, Assistant Professor

School of Mechanical, Aerospace and Manufacturing Engineering
University of Connecticut

Jiong Tang[†], Pratt & Whitney Endowed Chair Professor

School of Mechanical, Aerospace and Manufacturing Engineering
University of Connecticut

191 Auditorium Road, Unit 3139, Storrs, CT 06269, USA

Phone: +1 (860) 486-5911; Email: jiong.tang@uconn.edu

[†] Corresponding author

Seeing the Unseen: Towards Zero-Shot Inspection for Wind Turbine Blades using Knowledge-Augmented Vision Language Models

Yang Zhang, Qianyu Zhou, Farhad Imani, Jiong Tang[†]

School of Mechanical, Aerospace and Manufacturing Engineering, University of Connecticut

ABSTRACT

Wind turbine blades operate in harsh environments, making timely damage detection essential for preventing failures and optimizing maintenance. Drone-based inspection and deep learning are promising, but typically depend on large, labeled datasets, which limit their ability to detect rare or evolving damage types. To address this, we propose a zero-shot-oriented inspection framework that integrates Retrieval-Augmented Generation (RAG) with Vision-Language Models (VLM). A multimodal knowledge base is constructed, comprising technical documentation, representative reference images, and domain-specific guidelines. A hybrid text-image retriever with keyword-aware reranking assembles the most relevant context to condition the VLM at inference, injecting domain knowledge without task-specific training. We evaluate the framework on 30 labeled blade images covering diverse damage categories. Although the dataset is small due to the difficulty of acquiring verified blade imagery, it covers multiple representative defect types. On this test set, the RAG-grounded VLM correctly classified all samples, whereas the same VLM without retrieval performed worse in both accuracy and precision. We further compare against open-vocabulary baselines and incorporate uncertainty Clopper-Pearson confidence intervals to account for the small-sample setting. Ablation studies indicate that the key advantage of the framework lies in explainability and generalizability: retrieved references ground the reasoning process and enable the detection of previously unseen defects by leveraging domain knowledge rather than relying solely on visual cues. This research contributes a data-efficient solution for industrial inspection that reduces dependence on extensive labeled datasets.

Keywords: wind turbine blade, damage detection, vision language model, retrieval augmented generation, zero-shot inspection.

[†] Corresponding author.

*Initial submission to journal on: Apr 28 2025; First revision submission on: Sep 8 2025; Second revision submission on: Sep 20 2025.

1 Introduction

Wind energy has become a key component of the global transition toward sustainable power. With rapid growth in installed capacity and the increasing scale of wind farms, ensuring the reliability and durability of wind turbine components is more important than ever. Among these components, turbine blades are particularly vulnerable due to exposure to harsh environmental conditions, including ultraviolet radiation, rain, hail, lightning strikes, and temperature cycling. These stressors lead to diverse set of failure modes and damage patterns, such as leading-edge erosion, surface cracks, damaged lightning receptors, and delamination, among others. If undetected, such damage can compromise structural integrity, increase downtime, reduce energy output, and raise maintenance costs. Timely detection and assessment are therefore essential to maintain operational efficiency and safety [1-3]. A wide range of inspection and monitoring techniques has been explored for wind turbine blade evaluation, including acoustic emission sensing [4], ultrasonic testing [5-6], and infrared thermography [7]. These traditional methods are reliable in many scenarios and are increasingly enhanced by artificial intelligence. In contrast, vision-based approaches offer complementary advantages through their non-contact operation, high spatial resolution, full-surface coverage, and potential for near real-time defect identification. Machine vision also offers enhanced interpretability, allowing visual outputs to be directly assessed by human operators or automated systems. Furthermore, the advances in drone technologies enable flexible, automated, and scalable image acquisition from multiple viewing angles without requiring turbine shutdown [8-9]. Combined with AI-powered image analysis, this approach provides a safe, efficient, and non-intrusive solution for large-scale blade inspection, aligned with the practical needs of modern wind operations.

Recent advances in computer vision and deep learning have accelerated the development of automated blade inspection systems. Robust feature extraction methods, as demonstrated in studies on omnidirectional outdoor imagery [10], are particularly important for handling diverse conditions in field inspections. Multiple architectures have shown promise, with object detection frameworks such as YOLOv8 widely adapted via specialized enhancements for damage detection [11-13]. Multimodal approaches are also effective, with researchers integrating optical-thermal video fusion [14], visible-infrared image fusion [15], and hyperspectral imaging with 3D CNNs [16] to improve recognition under challenging field conditions. Attention-based models have emerged as powerful tools, with Vision Transformers outperforming traditional CNNs in surface defect classification [17] and various attention mechanisms being incorporated into existing architectures to enhance feature extraction [18]. Beyond supervised learning, unsupervised techniques, such as memory-aided denoising autoencoders [19] and reverse knowledge distillation [20], show promise in limited labeled settings, demonstrating the field's continued evolution toward robust and practical deployments. A comprehensive list of related studies is provided in Table 1. Despite this progress,

effectiveness still depends heavily on access to well-labeled, balanced, and high-quality datasets, a condition rarely met in realistic wind turbine environments. In practice, collecting representative and consistent training data poses significant challenges. Factors such as environmental variability, seasonal shifts, inconsistent drone viewpoints, and inspection scheduling contribute to uneven and often noisy datasets. Certain defects (e.g., surface stains) occur frequently and are easily captured, whereas more critical damage types (e.g., cracks, delamination, or lightning-induced erosion) occur infrequently and are harder to document. The result is a persistent and dynamic data imbalance, which undermines the performance and generalization capability of conventional deep learning pipelines.

Table 1. Wind turbine blade inspection using deep learning and vision techniques.

Papers	Techniques	Applications
[11, 12, 13, 18, 21, 22, 23]	YOLOv8, YOLOv7, YOLOv5 with SE attention, GSConv, EMA, GA optimization	Wind turbine blade damage detection
[15]	YOLOv7 with RGB-IR feature fusion	Multimodal wind turbine defect detection
[24]	YOLOv5s with semi-supervised learning	Blade defect detection with limited labeled data
[17]	Vision Transformers (ViT)	Surface defect detection in renewable energy assets
[14, 19, 25]	AQUADA-Seg, Memory-Aided Denoising Autoencoder with Swin Transformer U-Net, Siamese CNN with similarity learning	Blade segmentation and damage detection
[16]	3D CNN with hyperspectral imaging	Fault detection (cracks, erosion, ice)
[26, 27]	ResNet-50, Mask R-CNN	Blade crack detection and automated damage detection
[28]	Coarse-to-fine stitching with regression-based shape optimization	Drone-based image stitching for defect analysis
[20]	ResNet architectures with reverse knowledge distillation	Structural anomaly detection and localization
[29]	RARNN (Receptive Attention Recurrent Neural Network)	Digital twin for dynamic impact identification
[30]	Spatio-temporal attention model	Ice formation detection on wind turbine blades

The challenges above have prompted growing interest in zero-shot approaches for blade inspection and health monitoring. Recent vision-language models, including AnomalyCLIP [31] and FiLo [32], demonstrate promising capabilities. AnomalyCLIP leverages object-agnostic text prompts to detect anomalies across diverse domains, while FiLo incorporates large language models (LLMs) to provide fine-grained descriptions with enhanced localization. Similarly, GAN-based zero-shot transfer learning has shown strong performance for structural health monitoring, where researchers reported F1 scores of 0.978 [33] through domain adaptation and spectral mapping. More recent work has explored multi-source transfer learning [34] and autoencoder-based domain adaptation frameworks [35], with some approaches achieving high accuracy even with unseen damage classes [36]. All zero-shot related anomaly or damage detection studies are listed in Table 2. However, these methods have limitations. GANs and similar generative models

often demand extensive computational resources and still require a minimum amount of real data to generate realistic outputs. Zero-shot vision-language models may struggle to generalize to industrial domains underrepresented in pretraining data, and their outputs can lack grounding in domain-specific context. In practice, these limitations constrain scalability, robustness, and interpretability, especially in data-scarce, variable, or evolving inspection environments. There is therefore a pressing need for frameworks that make effective use of limited labeled data and expert knowledge, while remaining adaptable without retraining.

Table 2. Anomaly detection using zero-shot learning.

Papers	Techniques	Applications
[31]	Vision-language model fine-tuning with object-agnostic prompt optimization	Zero-shot anomaly detection
[32]	Vision-language model training with Grounding DINO	Zero-shot anomaly detection with fine-grained descriptions
[33, 34]	GAN-based data generation with feature alignment	Zero-shot structural damage detection
[35]	Zero-shot CNN with domain adaptation	Cross-domain damage diagnosis
[36]	Generalized zero-shot learning (GZSL) with CNN backbones (ResNet, VGG, DenseNet)	Structural damage assessment with unseen classes
[37]	Image-text alignment with LVLM inference	Zero-shot industrial anomaly detection

Recently, large language models have emerged as a compelling alternative, offering strong generalization capabilities with minimal supervision. LLMs are pretrained on massive corpora and can perform a wide range of tasks using only natural language prompts, often without additional fine-tuning. These tasks include question answering, summarization, and anomaly detection, often without requiring additional fine-tuning. This flexibility makes LLMs attractive for rapid deployment in domains with limited labeled data. However, a fundamental limitation remains: LLMs are trained on broad, general-purpose data, and often lack the domain-specific grounding required for high-stakes applications such as structural health monitoring or wind turbine inspection. This gap can lead to factual inaccuracies or so-called hallucinations in industrial applications [38], where models generate fluent but incorrect or unsupported outputs. To address these limitations, Retrieval-Augmented Generation (RAG) has emerged as a promising framework. By augmenting LLMs with external knowledge retrieved at inference time, RAG enables the model to ground its outputs in task-specific information without modifying its underlying parameters. This approach has been applied successfully in several recent studies. For instance, SafeLLM [39] introduces a domain-specific safety monitoring framework for offshore wind maintenance. It leverages LLMs together with statistical techniques to identify potentially unsafe or hallucinated responses. Pastoriza et al [40] developed a retrieval-augmented anomaly detection system, which incorporates human-in-the-loop feedback for continuous error correction. Similarly, Thimonier et al [41] applied retrieval-augmented learning to deep

anomaly detection in tabular data, using transformer-based reconstructions grounded in retrieved context. Other promising approaches include AnomalyGPT [37], which leverages large vision-language models for industrial anomaly detection and reports 86.1% accuracy on benchmark datasets. Another example is LLM-DSKB [42], which integrates LLMs with domain-specific knowledge bases for industrial equipment operation and maintenance. A summary regarding LLM and RAG in industrial applications is listed in Table 3.

Table 3. RAG and LLM in industrial applications.

Papers	Techniques	Applications
[39, 43]	Statistical safety measures with Wasserstein distance and cosine similarity using Universal Sentence Encoder	Domain-specific safety monitoring for offshore wind maintenance
[40]	Retrieval-augmented post-processing	Anomaly detection adjustment
[41]	Transformer-based anomaly reconstruction with retrieval-enhanced scoring	Anomaly detection in structured tabular data
[42]	LLM embeddings with vector retrieval	Domain-adapted industrial equipment maintenance
[44]	Time-series to text conversion with LLM prompting and forecasting	Zero-shot time series anomaly detection
[45]	Multimodal LLM approach for contextual understanding and information extraction	Fault detection and diagnostics in hydrogenator

In addition to these industrial applications, recent studies have investigated RAG-grounded VLMs in broader visual tasks. For example, Visual RAG [46] demonstrates how multimodal large models can expand visual knowledge without fine-tuning by retrieving relevant exemplars at inference time. Bhat et al [47] integrated RAG with VLMs for scientific visual question answering, significantly improving factual grounding. Similarly, Dong et al [48] introduced semantic document layout analysis to enhance visually rich RAG tasks, while Khan et al [49] applied retrieval-augmented multimodal reasoning to open-vocabulary species recognition, achieving notable gains on unseen categories. These developments underscore the growing interest in RAG-grounded VLMs for visual understanding tasks [50]. Wind turbine blades present unique inspection challenges that differ from conventional anomaly detection tasks. Their large structural scale, diverse damage modalities (e.g., cracks, corrosion, peeling, and composite delamination), and highly variable environmental conditions complicate the design of robust detection systems. Traditional supervised learning approaches struggle to keep pace with these evolving and heterogeneous failure patterns, because they require continuous data collection and retraining. Vision-language models, when augmented with retrieval mechanisms, are well positioned to address these challenges by linking visual semantics from inspection imagery with structured domain knowledge, thereby enabling more adaptive, interpretable, and generalizable inspection capabilities. Nevertheless, despite their potential, the integration of RAG-grounded VLMs into the wind energy sector remains largely unexplored.

To fill this gap, we propose a multimodal retrieval-grounded visual reasoning framework for wind turbine blade inspection, enabling interpretable toward-zero-shot diagnosis. Unlike conventional text-only RAG, our system integrates textual and visual knowledge within a structured hybrid knowledge base that includes damage descriptions, turbine information, maintenance guidelines, and annotated exemplars. Knowledge entries are embedded with dual encoders (text and image) and stored in a vector database for efficient cross-modal retrieval. A domain-aware reranking mechanism further refines the results to ensure each inspection is guided by the most relevant evidence. The retrieved context is incorporated into a dynamic prompt that balances general expertise with visually similar reference cases, guiding a vision-language model to produce structured diagnostic reports covering blade count, damage presence, type, severity, and explanatory rationale. The architecture requires no task-specific retraining and supports on-the-fly updates through simple additions to the knowledge base. Importantly, the system preserves transparency by tracing which knowledge items informed each decision, creating a clear provenance from evidence to conclusion. In experiments, the framework demonstrates clear advantages over both supervised detectors and open-vocabulary vision models, particularly for rare or complex damage categories where labeled data are scarce. Although our evaluation set is necessarily limited due to the difficulty of obtaining verified blade images, it spans diverse defect types, and results are supported by baseline comparisons and uncertainty analysis. Together, these elements highlight the promise of RAG-grounded VLMs for creating adaptive, data-efficient, and interpretable inspection systems tailored to the unique challenges of the wind energy sector.

The remainder of this paper is organized as follows. Section 2 presents the overall architecture of the proposed visual-text RAG framework, including the design of the hybrid knowledge base, embedding strategy, retrieval mechanism, and prompt construction. Section 3 details the experimental setup, including dataset preparation, baseline models, and implementation specifics. Section 4 reports and discusses the evaluation results, highlighting the effectiveness of the proposed method under limited data conditions. Finally, Section 5 concludes the paper and outlines future directions for expanding domain adaptability and real-world deployment.

2 Framework of Wind Turbine Blade Inspection with RAG-Grounded VLM

2.1 Overall architecture

Traditional inspection methods struggle with the complex nature of wind turbine blade damage detection under diverse environmental conditions. To address these challenges, we design a multimodal RAG-grounded VLM framework that integrates textual and visual knowledge for accurate damage assessment. As shown in Figure 1, the architecture follows an end-to-end pipeline with four main stages: (1) Data Collection, where drone-captured blade images are uploaded and prepared for analysis; (2)

Knowledge Base Preparation, in which expert documents and annotated reference images are encoded and indexed for efficient retrieval; (3) Retrieval-Augmented Inference, where damage queries trigger cross-modal similarity search and domain-aware reranking to assemble the most relevant evidence; and (4) Vision-Language Reasoning, where the enriched context and input image are processed by a VLM to generate structured diagnostic reports, including damage detection, type, severity, and descriptive explanation. This pipeline enables interpretable towards zero-shot assessment by grounding visual analysis in domain knowledge.

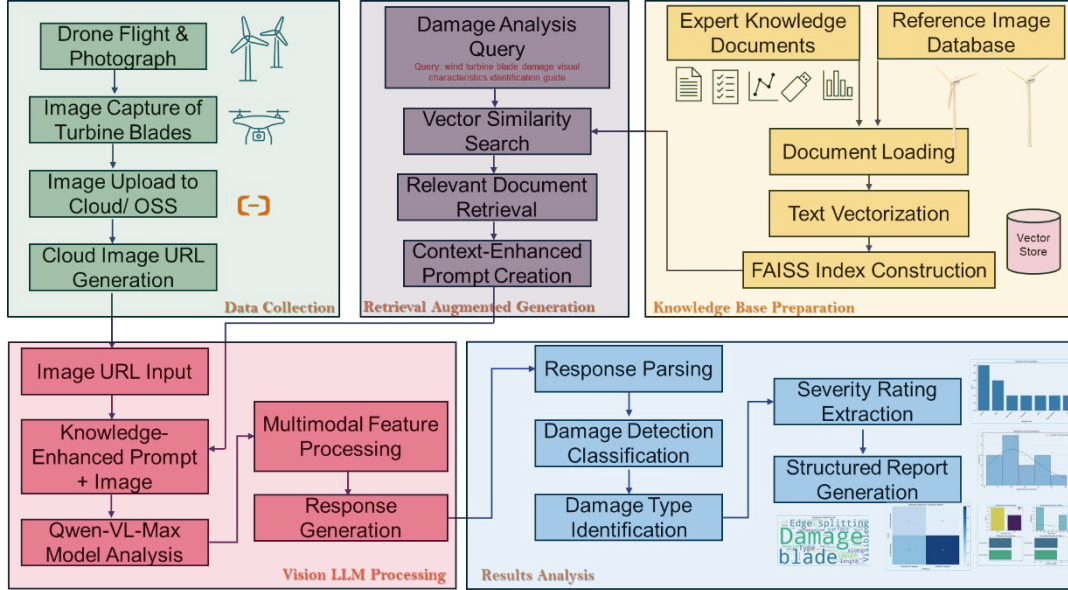


Figure 1. General flowchart for wind turbine blade inspection using RAG and VLM.

2.2 Retrieval augmented generation

RAG is the foundation of our framework, allowing the vision-language model to ground its analysis in domain-specific knowledge that pre-trained models alone cannot provide. By retrieving relevant information at inference time, RAG enables accurate assessment without task-specific fine-tuning. The implementation extends conventional RAG by incorporating multimodal retrieval, combining textual descriptions (e.g., classification criteria) with visual exemplars (e.g., images of damage types) to provide complementary evidence for blade inspection. The RAG pipeline comprises three components: (1) a structured knowledge base with domain-specific documentation and reference images, (2) a vector database for efficient embedding-based retrieval, and (3) a similarity search with reranking to identify the most relevant context for each query. Figure 2 shows the schematic flow of this system, and the following subsections describe each component in detail.

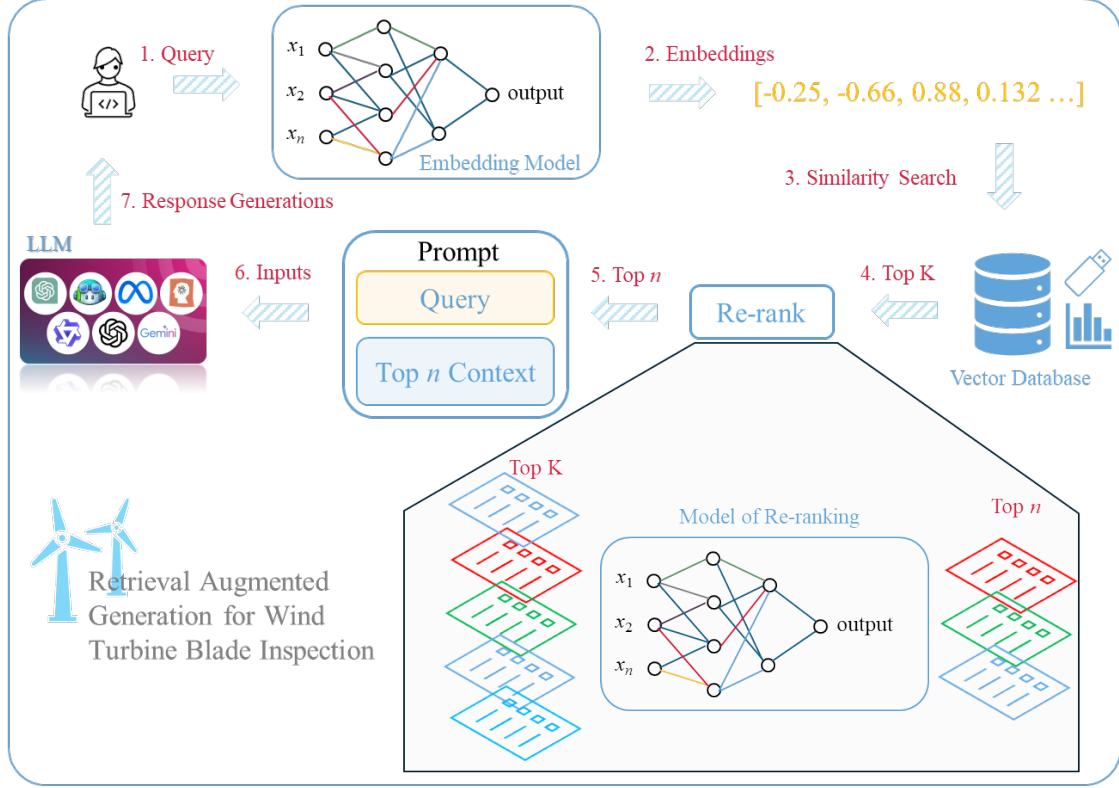


Figure 2. Schematic diagram of retrieval augmented generation.

2.2.1 Knowledge base

The knowledge base serves as the repository of domain-specific information that the system can access during analysis. We designed a dual-modal knowledge base that incorporates both textual and visual information relevant to wind turbine blade damage assessment. The textual component of the knowledge base contains several types of information. Lists of contents from knowledge base are provided in Table 4. There are totally 4 knowledge bases, with each functioning in different roles. For example, the damage description knowledge base (Textual 1) stores the descriptions of commonly seen damage types in wind turbine blades. The descriptions detail how the damage should look like, what the color of that damage will be, and what the shape of that damage type is, etc. Additionally, we provide the basic information of the wind turbine itself (Textual 2), such as what material the blade is made of, what the blade looks like under healthy condition (painted white), etc. Therefore, when VLM fetches this knowledge, it can help the VLM distinguish the damage (which usually causes color change) from the healthy state. Furthermore, we provide maintenance logs (Textual 3), which record past events. In each event, the identified damage is labeled with a severity level. Although this is not an essential feature, it is often useful to obtain a rough estimate of damage severity, which can assist in maintenance planning and logistics. We also incorporate an image-textual database (Image-Textual Metadata) in this study. Since the datasets we use involve images that are

taken in different visual conditions, such as sunny daytime, cloudy daytime, nighttime, and dusk, this multimodal knowledge base is essential for robust performance across varying lighting conditions. The image-textual database provides concrete visual examples that help the model recognize damage patterns despite visual variations due to weather, time of day, or camera position. This point is critical, as certain damage types exhibit different visual characteristics under varying lighting conditions, and textual descriptions alone are often inadequate for accurate identification. The visual component of the knowledge base consists of reference images showing various types of wind turbine blade damage. Each image is associated with metadata including **Description**: A textual explanation of the specific damage features visible in the image; **Image path**: A reference to the stored image file. In embeddings, the image will be fetched through the path. This metadata is stored in JSON format, enabling seamless integration with the textual knowledge base while maintaining the relationships between images and their corresponding damage descriptions.

Table 4. Lists of expert knowledge base about wind turbine blades.

Knowledge Types	Descriptions (Dynamic)
Image-Textual Metadata	• Images with known damage. The texts describe the images from different perspectives.
Textual 1	• Texts describing damage types: color, shape, location etc.
Textual 2	• Texts describing features of wind turbine, material, color, vortex generator, etc.
Textual 3	• Texts describing maintenance logs and damage severity levels.

To enhance retrieval efficiency, the textual documents are processed using a chunking mechanism based on the *RecursiveCharacterTextSplitter* from the LangChain library [51], which is specifically designed for document processing in retrieval-augmented generation systems. The *RecursiveCharacterTextSplitter* implements an intelligent recursive splitting algorithm that attempts to maintain the semantic integrity of content through hierarchical boundary detection. The *RecursiveCharacterTextSplitter* operates by attempting to split text at the most semantically meaningful boundaries first, working through the provided separators list in order. In implementation of this study, it first attempts to split at paragraph breaks ($\n\backslash\n$), then at line breaks ($\backslash\n$), then at sentence boundaries ($\.$), and finally at word boundaries ($\$). Only if no suitable boundaries are found will it resort to character-level splitting ($\text{"\text{"}$). This hierarchical approach preserves the natural structure of the text as much as possible, which is crucial for maintaining the contextual meaning of technical documentation. This approach divides documents into manageable segments while maintaining semantic coherence by preferentially splitting at natural boundaries such as paragraph breaks. The chunk size parameter (set to 1000 characters) and chunk

overlap parameter (set to 200 characters) are selected to balance retrieval precision with computational efficiency. The overlap between adjacent chunks ensures that concepts spanning chunk boundaries are not lost during retrieval. The knowledge base is designed to be extensible (as marked as Dynamic in Table 4), allowing new documents and reference images to be added through the adding document and adding reference image methods, respectively. This extensibility is critical for real-world deployment, as it enables the system to continuously incorporate new knowledge and examples as they become available, improving performance over time without requiring model retraining.

2.2.2 Embeddings and vector base

To enable efficient retrieval of relevant information from the knowledge base, we implemented a dual-embedding approach that captures both textual and visual semantic relationships. This section details the embedding models selected for each modality and the vector storage solution used for similarity search.

Text embeddings

For encoding textual information, we select the Sentence-BERT all-MiniLM-L6-v2 model [52]. This model is chosen because it achieves a good balance between computational efficiency and embedding quality, making it suitable for resource-constrained industrial environments. It also performs well in capturing semantic relationships between texts, which is important for retrieving conceptually related information even when keywords do not match exactly. In addition, although the model is pre-trained on general text, it has demonstrated strong transferability to technical domains without requiring task-specific fine-tuning. The text embedding process then transforms each document chunk into a dense vector representation. Each document chunk is encoded into a fixed-dimensional vector (384 dimensions, referring to the dimensionality of the vector representation created by the model) that captures its semantic content. These vectors enable the system to identify conceptually similar documents even when they use different terminology, addressing a key limitation of traditional keyword-based retrieval systems.

Image embeddings

For encoding visual information, we employ the CLIP (Contrastive Language-Image Pre-training) model [53], specifically the openai/clip-vit-base-patch32 variant. CLIP is well suited to our framework because it learns a joint representation of text and images, which enables effective cross-modal retrieval. The Vision Transformer (ViT) architecture underlying CLIP captures rich visual features that are highly relevant for damage detection. Moreover, CLIP’s training strategy provides strong zero-shot generalization, which is crucial for recognizing rare or unusual damage patterns. Following extraction, a normalization procedure is applied to ensure consistency in similarity measurements during the retrieval process. The resulting embeddings are 512-dimensional vectors that capture the visual characteristics of each reference image. The choice of 512 dimensions for visual embeddings complements our text embedding

dimensionality (384) while providing sufficient capacity to represent complex visual features. The higher dimensionality of visual data reflects its intrinsic complexity relative to textual content, enabling richer representation of spatial patterns, textures, and other fine-grained characteristics critical for effective damage analysis.

Vector storage with FAISS

To enable efficient similarity search, we implemented vector storage using the FAISS library [54]. FAISS is selected for its computational efficiency, scalability and customization options. Specifically, FAISS implements optimized algorithms for high-dimensional similarity search, enabling rapid retrieval even with large knowledge bases. In addition, the library supports both in-memory and disk-based indices, allowing the system to scale to large collections of documents and images. Furthermore, FAISS offers various index types optimized for different retrieval scenarios, and in our implementation, we select the basic IndexFlatL2, which performs exact nearest neighbor search using the L2 (Euclidean) distance metric, due to its precision and the moderate size of our knowledge base.

We create separate indices for textual and visual embeddings. Specifically, we select the basic IndexFlatL2 structure due to its precision and the moderate size of our knowledge base. The text index is configured to handle the 384-dimensional text embeddings, while the image index is designed specifically for the 512-dimensional visual embeddings. Both indices store their respective embeddings as float32 arrays, enabling efficient similarity searches across both modalities. The L2 distance metric is chosen for similarity calculations as it provides intuitive distance measurements in the embedding space and is compatible with the normalized embeddings produced by the models. The mathematical expression for L2

is $L_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$. Here $\mathbf{x}$ and $\mathbf{y}$ are the vectors being compared, with x_i and y_i representing their respective components at position i . This dual-embedding approach, combined with efficient vector storage, forms the foundation of our retrieval system, enabling the integration of diverse information sources during the damage assessment process.

2.2.3 Similarity search and rerank

The effectiveness of an RAG system is fundamentally determined by its ability to retrieve the most relevant information from the knowledge base. The approach proposed implements a two-stage retrieval process consisting of an initial similarity search followed by a reranking step that further refines the results.

Hybrid similarity search

To leverage both textual and visual information during retrieval, we implement a hybrid similarity search mechanism that operates across both modalities simultaneously. This parallel retrieval approach is

a key point in our system, as it allows for the integration of complementary information types during analysis. The similarity search process begins with a textual query formulation. For damage assessment tasks, we define a default query that captures the essential information needs: *“comprehensive wind turbine blade damage assessment guidelines including technical documentation. The image to be analyzed may be taken at cloudy, night or dusk with bad vision.”* This query is intentionally designed to retrieve broad contextual information about damage assessment while also accounting for challenging imaging conditions that are common in real-world wind turbine inspections. The query is encoded using the same text embedding model used for the knowledge base, transforming it into a 384-dimensional vector representation. This embedding is then used to perform a k-nearest neighbors search in the text index, retrieving the most semantically similar documents from our knowledge base using $\text{sim}(\mathbf{q}, \mathbf{d}) = \mathbf{q} \cdot \mathbf{d} / \|\mathbf{q}\| \cdot \|\mathbf{d}\|$. $\mathbf{q}$ represents the query vector (the search input) and $\mathbf{d}$ represents the document vector (i.e., items in the knowledge base). In parallel, if an input image is provided, it is processed using the CLIP model to create a 512-dimensional visual embedding. This embedding is used to perform a similar k-nearest neighbors search in the image index, identifying visually similar reference images from our database. The result of these parallel searches is a set of potentially relevant text documents and a set of visually similar reference images. The parameter top_k controls the initial number of results retrieved from each modality, providing a balance between recall (retrieving all relevant items) and the computational cost of subsequent processing, as shown in Equations (1) and (2).

$$R_{\text{text}} = \text{topK}(\text{sim}(\mathbf{q}_{\text{text}}, \mathbf{d}_i)) \quad \forall \mathbf{d}_i \in D_{\text{text}} \quad (1)$$

$$R_{\text{image}} = \text{topK}(\text{sim}(\mathbf{q}_{\text{image}}, \mathbf{d}_j)) \quad \forall \mathbf{d}_j \in D_{\text{image}} \quad (2)$$

We set topK to 5 in our implementation to strike an optimal balance between retrieval comprehensiveness and computational efficiency, ensuring the system captures sufficient contextual information while maintaining responsiveness for real-time damage assessment applications.

Reranking algorithm

While embedding-based similarity search is effective at identifying broadly relevant information, it may not optimally prioritize the most useful documents for a specific task. To address this, we implement a reranking algorithm that refines the initial retrieval results. The reranking approach combines multiple signals to assess the relevance of each retrieved document. Specifically, the algorithm considers two main factors. First, documents containing more query keywords are assigned higher scores, which helps prioritize the most topically relevant information. Second, shorter, and more focused documents are given slightly higher priority, as they often contain concentrated relevant content compared with longer, more general documents. While the initial retrieval phase operates in the vector space where documents and queries are represented as embeddings, the reranking phase works directly with the retrieved document objects and

their textual content. In this second phase, we process the actual text rather than vector representations, allowing for content-based heuristics as shown in the following equations,

$$\text{keyword_score}(d) = \sum_{k \in \text{keywords}} \text{Ind}(k \in \text{content}(d)) \quad (3)$$

$$\text{length_factor}(d) = \frac{1}{0.1 + \frac{|\text{content}(d)|}{1000}} \quad (4)$$

$$\text{score}(d, q) = \text{keyword_score}(d) \cdot \text{length_factor}(d) \quad (5)$$

where Ind is an indicator function that returns 1 when keyword k appears in the document content and 0 otherwise; $|\text{content}(d)|$ represents the document length; q denotes the query keyword set, and d represents the document object. The division by 1000 normalizes document length to a practical scale, making the formula work effectively across documents of varying sizes. This approach is computationally efficient while still providing meaningful improvements over the initial embedding-based retrieval. The parameter top_n (set to 3 in the implementation) controls the final number of documents retained after reranking, focusing the context on the most relevant information using score obtained as show in Equation (6).

$$R_{\text{final}} = \text{top_n}(\text{score}(d, q)) \quad \forall d \in R_{\text{initial}} \quad (6)$$

The combination of hybrid similarity search with reranking enables our system to efficiently identify the most relevant textual and visual information for a given damage assessment task, providing a rich contextual foundation for the analysis of vision-language model.

2.3 Response generation and result extractions

The final component of our system is the response generation module, which leverages a vision-language model to analyze the input image in conjunction with the retrieved context and produce a structured damage assessment. This section details the prompt engineering approach, the VLM integration, and the result extraction technique.

Dynamic prompt construction

A critical aspect of our approach is the construction of effective prompts that guide the analysis of VLM. Rather than using static prompts, we adopt a dynamic prompt construction technique that incorporates retrieved contextual information. This approach enables the model to benefit from domain-specific knowledge while maintaining the flexibility to address diverse damage assessment scenarios. The prompt construction process starts with an initial base instruction: *“I need to utilize the knowledge base and observe the features of the anomaly on the wind turbine related components, and identify damage type.”* This foundation establishes the core analytical objective. To achieve dynamic prompting, we append a

transitional phrase: “*Using the following reference information to help with the analysis:*” In this process, two key elements are incorporated: relevant textual knowledge and similar reference images. Textual knowledge is derived from the top-ranked documents, where the most relevant damage type information is extracted and formatted. Metadata from similar reference images, when available, is also integrated to provide comparative examples. This dynamic composition ensures each prompt is uniquely tailored to the specific damage assessment task while maintaining a consistent framework that guides the VLM analysis process. By combining task-specific instructions with contextually relevant knowledge, the proposed approach enables more accurate and informed damage assessments across diverse scenarios. The prompt including specific analysis instructions is shown in Table 5:

Table 5. Dynamic prompt constructions.

Analytical Prompts
"Based on these descriptions and references, analyze the image and determine: 1. How many blades are visible in the image? 2. Is there visible damage on any of the turbine blades in the image? 3. If yes, what specific type of damage can be identified in this damage lists ('Missing Teeth of Vortex generators', 'Lightning Receptors', 'Crack', 'Corrosion', 'Erosion', 'Rust', 'Delamination', 'Fracture', 'Dent', 'Ice', 'Snow', 'Surface Peeling', 'Wear', 'Lightning Strike/Burning')? 4. Provide a detailed description of the damage observed, referencing the specific characteristics described above. 5. Rate the severity of the damage on a scale of 1-5, where 1 is minor and 5 is severe."

These structured questions serve multiple purposes. They provide a clear analytical framework that guides the assessment of VLM and ensure comprehensive coverage of key damage characteristics. In addition, they facilitate subsequent extraction of structured information from the responses generated by the model. When visually similar reference images are available, the prompt further includes adaptive guidance based on patterns in those references. This adaptive component enhances the system's intelligence in two ways. When all similar reference images show the same damage type, a targeted note is added to suggest that the VLM carefully check for this specific type. When the reference images display multiple damage types, these are presented as potential candidates, guiding the VLM to assess which, if any, are present.

Result extraction and structuring

Our system integrates with the Qwen-VL-Max [55] vision-language model through an OpenAI-compatible API interface. The implementation passes both the constructed prompt and the image URL to the model, enabling comprehensive analysis that considers both visual evidence and contextual information. The image is fetched from cloud storage where it is pre-uploaded. This mimics the engineering implementation loop where the drone captures images and transmits them to the terminal, where the system then analyzes the visual data. Since the output from VLM is natural language, we implement a systematic extraction approach that transforms the free-text response into a structured format containing key assessment metrics. The structured output captures six essential elements: the complete model response, a damage detection flag, identified damage types, severity rating, descriptive assessment, and a record of knowledge base elements used during analysis. A schematic diagram is given to show the structured answer extraction process in Figure 3.

The extraction process employs natural language processing techniques to identify meaningful patterns in the model response. For damage detection, we analyze the text for indicative phrases such as "damage is detected," "there is damage," or "signs of damage." This approach provides a Boolean indicator of whether damage is detected in the image. For damage type classification, we implement a comprehensive pattern recognition system that searches for mentions of specific damage categories from our taxonomy, including cracks, corrosion, delamination, and others. Importantly, our system incorporates negation handling to prevent false positives. For example, phrases like "no cracks" or "absence of corrosion" are correctly interpreted as the absence of those damage types rather than their presence. Severity assessment is extracted through targeted pattern matching that identifies numerical ratings within context. The system searches for phrases like "severity: 3" or "severity rating of 4" to establish a quantified assessment on our predefined 1-5 scale.

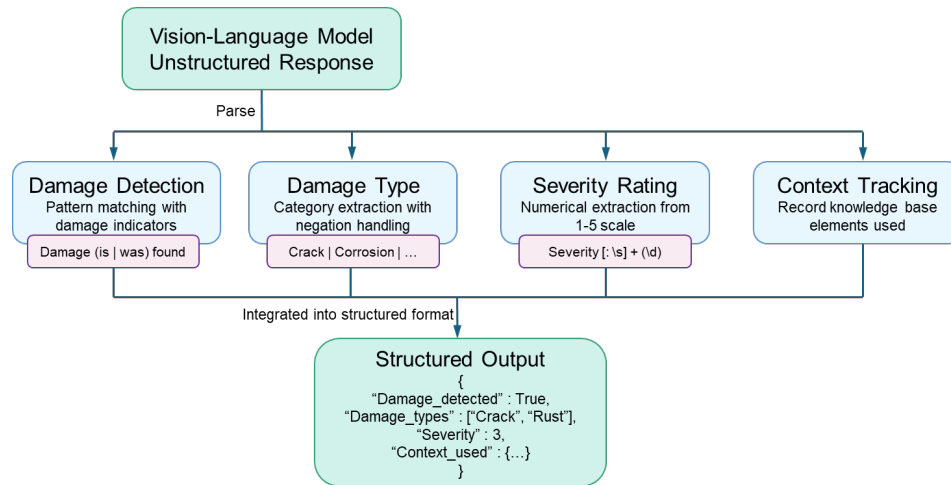


Figure 3. Structured output extraction process.

This extraction method is designed to be robust against variations in the model response format. By carefully considering contextual cues and linguistic constructions, we ensure accurate capture of assessment metrics regardless of the specific phrasing used by the model. The inclusion of negation handling is particularly important for technical assessments, as it prevents misinterpretation when the model explicitly notes the absence of damage characteristics. The resulting structured output significantly enhances the system integration capabilities with downstream applications such as maintenance management systems, inspection databases, or automated reporting tools. Furthermore, by tracking which knowledge base elements influenced the assessment through the "context_used" field, we provide transparency that supports explainability and audit capabilities. This balanced approach to information extraction retains the model's nuanced analysis and allows for both automated processing and human review of damage assessment results. Moreover, the structured format supports efficient data processing, trend analysis, and maintenance prioritization while preserving the contextual richness of the original assessment.

3 Wind Turbine Blade Inspection Case Study

In this section, we examine the feasibility and effectiveness of the proposed zero-shot approach for wind turbine blade inspection. We will detail each step of the framework, from dataset selection and knowledge base preparation to inspection results.

3.1 Datasets overview

This study utilizes two open-source datasets: one from Chen [56-57] and another from Foster et al [21]. The Chen dataset includes both optical and thermal imagery. All blade videos are captured using either DJI Zenmuse H20T or DJI Mavic 2 Enterprise Advanced drones while wind turbines operated normally. For thermal imaging, the fusion color palette is selected. The data collection protocol involves positioning the drone approximately 12 ± 4 meters horizontally from the hub nose and 2 meters vertically (Figure 4). To avoid thermal interference from the turbine, the camera is tilted upward by 15 degrees before capturing paired optical and thermal videos. For longer blades, multiple segments horizontally or vertically are recorded, maintaining a 5-meter interval between filming positions. Videos are taken from both sides of the blades. To enhance data diversity and model robustness, footage from various angles and distances are also recorded. The final dataset comprises 36 videos, with both training and testing sets containing all video frames. Each frame has a resolution of 853×480 pixels. While thermal images were collected, they are not included in our current study. Since our goal is to achieve zero-shot inspection capability, we randomly selected images from both the training and testing datasets, rather than adhering to their original division. The second dataset originates from Shihavuddin and Chen [58], with original images at 5280×2970 -pixel resolution. In a recent study, Foster et al [21] subdivided these original images into 72 smaller segments of

586×371 pixels each to accommodate YOLOv5 input requirements. For our inspection, we use this more recent 586×371-pixel version.

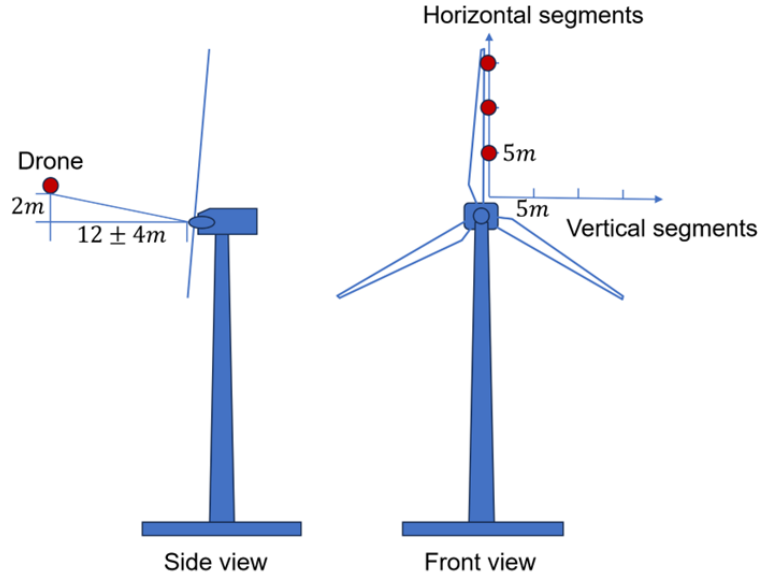


Figure 4. Schematic diagram of photography of drone for wind turbine [56].

It is important to note that neither dataset comprehensively covers all possible damage types. The first dataset primarily contains healthy state examples and crack damage. The second dataset predominantly features corrosion, erosion, surface peeling, and dirt accumulation, though Foster et al [21] simplified their classification to just two categories: dirt and damage. To expand the range of identifiable damage types in our proposed approach, we supplemented the datasets with three additional damaged blade images exhibiting lightning strike and burning, icing and snow accumulation, and leading-edge erosion, as shown in Figure 5 from [59]. Consequently, our final testing dataset encompasses multiple conditions: healthy state, corrosion, erosion, surface peeling, icing and snow, lightning strike and burning, and cracks. As images in the second dataset often display multiple damage types simultaneously (such as combined corrosion/erosion with surface peeling), we further categorized all damage types into four mechanism-based groups: healthy state, surface damage (corrosion, erosion, surface peeling, rust), environmental damage (icing, snow, lightning strike, burning), and structural damage (fracture, cracks). Our confusion matrix will be based on these four categories. To address potential concerns about subjective classification, potentially inflating accuracy metrics, we will include the complete raw responses for all 30 testing images in our results discussion.

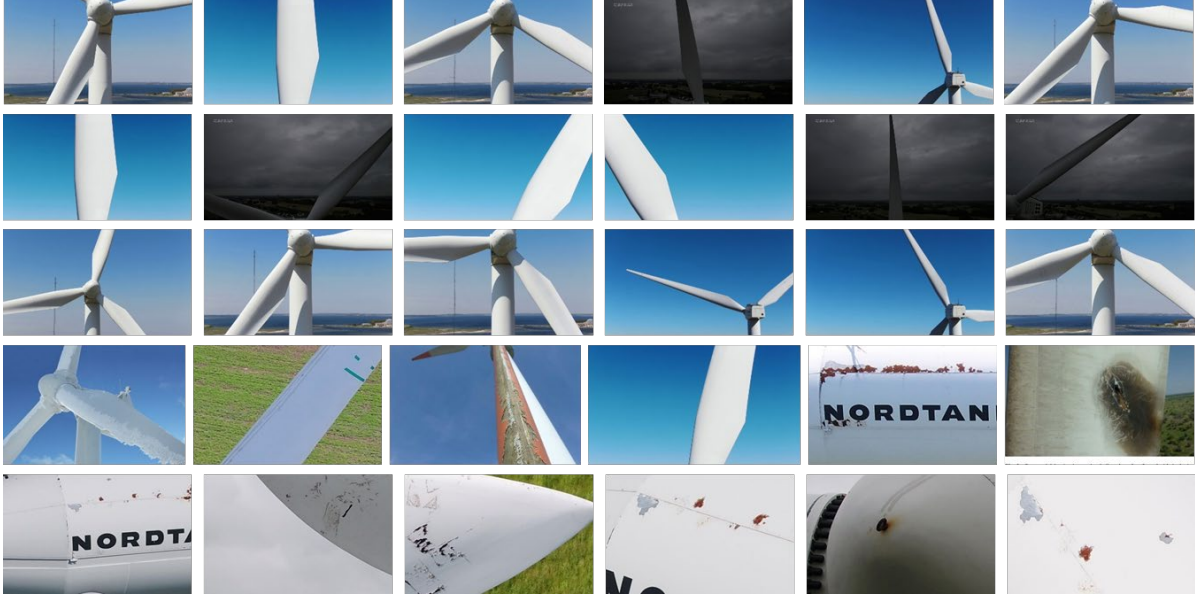


Figure 5. Testing wind turbine blade images from datasets [21, 56, 59].

3.2 Knowledge base preparations

We aim to achieve blade inspection towards zero-shot without training or fine-tuning large language models. Instead, we adopt a retrieval-augmented generation approach that relies on specialized domain knowledge. In this study, we construct four professional knowledge bases: three textual and one image–text database. The first contains descriptions of blade damage types, including visual appearance, color changes, typical locations, and other relevant cues (e.g., an example description of crack damage is given in Table 6). The second describes turbine blade structures, covering materials, surface coatings, markings, lightning conductors, and vortex generators that should not be misclassified as damage. The third consists of simulated maintenance logs summarizing past damage cases, severity levels, and approximate classifications; although secondary to our analysis, this resource supports rough severity estimation when damage is detected. The fourth knowledge base is multimodal, pairing blade photographs under different health conditions with textual descriptions. These descriptions specify lighting and weather conditions, the number of blades visible, and the type of damage if present. Together, the four knowledge bases are embedded and stored in a vector database. During inference, similarity search with reranking retrieves the most relevant entries, which are combined with a predefined prompt. The system then uses the top three retrieved references to support interpretable blade inspection. It should be noted that the reference images included in the knowledge base are drawn only from the Chen datasets [56-57], meaning that part of the test images originates from different sources and are not present in the knowledge base.

Table 6. Knowledge base preparations and examples.

Knowledge Types	Sample (Part of knowledge base)
Image/Textual	 The image was captured during daytime with clear blue-sky conditions. It shows three wind turbine blades with a visible crack at the end of the bottom right blade. The crack is oriented perpendicular to the longitudinal edge of the blade, which may indicate structural stress damage.
Descriptions of damage types	Cracks: The obvious features for the cracks are that they are perpendicular to the length of the blade. Crack damage looks like linear fractures on the blade surface, often appearing as fine lines that can range from microscopic to several centimeters in length. They typically start at stress concentration points and may be straight, branched, or web-like. Fresh cracks appear as sharp, clean breaks with defined edges, while older cracks may have discoloration around the edges. They can be superficial (affecting only the outer layer) or structural (penetrating deeper into the blade material).
Descriptions of turbine	- Wind turbine blades are typically made of composite materials, primarily fiberglass reinforced polyester or epoxy, and sometimes carbon fiber for larger blades. - A load-bearing spar or spar caps running the length of the blade. - Leading and trailing edge reinforcements. - Outer shell or skin made of composite materials. - Protective coating and paint to shield against environmental elements. - Root section reinforced with metal for connection to the hub. - There is usually a seam line on the back side of the blade, along the length direction but it is not damage.
Maintenance logs	- Level 1 (Minor): Superficial damage that does not affect structural integrity or performance. Examples include minor surface erosion, small scratches, or minor coating damage. Monitoring recommended. - Level 2 (Low): Early-stage damage that may progress if not addressed but does not present immediate concern. Examples include small cracks less than 10cm, early stage leading edge erosion, or limited surface peeling. Scheduled repair recommended within 3-6 months.

3.3 Identification results

It is worth mentioning that our approach follows a setting towards zero-shot: the RAG-grounded VLM is not trained or fine-tuned on the target dataset, but directly applied to unseen images. Therefore, concepts such as cross-validation or train/test splits are not applicable in the conventional sense. Our evaluation is performed on 30 independent test images from open-source datasets, with ground-truth labels established by domain experts. After processing through the RAG-grounded VLM, we extract the model judgments from the originally generated responses and compare them with the actual damage conditions. The results are shown in Figure 6. From the confusion matrix in Figure 6(a), we can see that the proposed approach can accurately identify different types of damage. Additionally, we provide various metrics to measure the performance of the proposed method, as shown in Figure 6(b). Among these metrics, accuracy refers to the proportion of correctly predicted samples (including true positives and true negatives) out of the total number of samples; precision refers to the proportion of true positives among all samples predicted as positive; recall refers to the proportion of true positives that are correctly predicted as positive among all actual positive samples; and the F1 score is the harmonic mean of precision and recall, calculated as

$2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$. The proposed method achieves 100% accuracy across all test images. However, to provide statistical rigor and acknowledge the inherent uncertainty associated with finite sample sizes, we apply Clopper-Pearson exact confidence interval analysis to our performance metrics. The exact method is particularly suitable for proportion estimates, especially when dealing with extreme values such as perfect accuracy, as it avoids the limitations of normal approximation methods that can produce unrealistic bounds. The expression is given in Equation (7).

$$CI = \begin{cases} [0, \text{Beta}_{1-\alpha/2}(1, n)] & \text{if } x = 0 \\ [\text{Beta}_{\alpha/2}(x, 1), 1] & \text{if } x = n \\ [\text{Beta}_{\alpha/2}(x, n - x + 1), \text{Beta}_{1-\alpha/2}(x + 1, n - x)] & \text{if } 0 < x < n \end{cases} \quad (7)$$

where, x represents the number of correct classifications, n is the total sample size, α is the significance level (i.e., $\alpha = 1 - \text{confidence level}$), and $\text{Beta}_p(a, b)$ is the p -th quantile of the Beta distribution with shape parameters a and b . This method provides exact confidence intervals by leveraging the relationship between the binomial distribution and the Beta distribution, offering high accuracy compared to normal approximation methods, particularly when dealing with extreme proportions or small sample sizes. This approach embodies a fundamental statistical principle: when evidence is limited, estimates should reflect appropriate uncertainty, while as sample size increases, confidence in the estimates grows accordingly. As shown in Figure 6(b), the Clopper-Pearson 95% confidence intervals for our key metrics demonstrate robust performance with statistical transparency. All performance metrics achieved perfect scores with Clopper-Pearson 95% confidence intervals of [88.4%, 100%]. These intervals acknowledge the sample size limitations while demonstrating that even under conservative statistical assumptions, our framework maintains strong performance with lower bounds consistently above 88%, representing robust classification capability for wind turbine blade inspection tasks. It is worth noting that in this method, no model training or fine-tuning is performed.

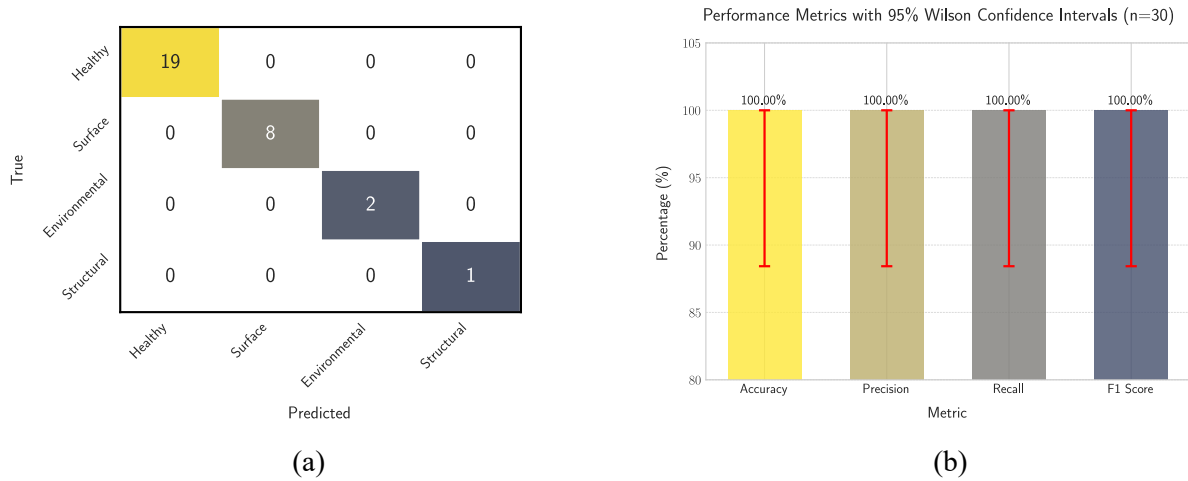


Figure 6. Confusion matrix (a) and evaluation metrics (b) for testing datasets.

Meanwhile, we have investigated the statistics on the damage severity levels across all test data, as shown in Figure 7(a). We can see that most instances are classified as level zero, corresponding to a healthy state. Among the other photographs with damage, the severity is mostly concentrated at level three, indicating moderate damage. One case shows particularly severe damage, which is a photograph containing lightning strikes and burning damage with a large hole, as shown in Figure 5. From these results, we can see that the proposed approach is not only capable of providing accurate damage identification but also reliable damage severity estimation, which offers meaningful reference for practical engineering applications. For example, recording the damage severity of different blades can help in efficiently arranging logistics, maintenance, and other activities, thereby reducing costs.

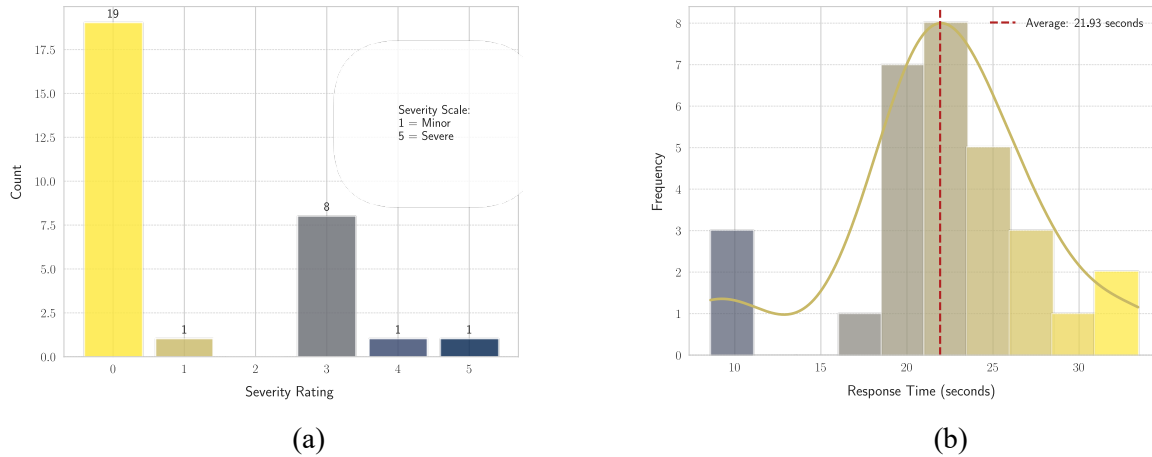


Figure 7. Severity distributions (a) and response time distributions (b) among testing datasets.

Another noteworthy aspect is the distribution of response times, as shown in Figure 7(b). For the 30 test samples, the average response time was 21.93 seconds on Google Colab equipped with 12.7GB system RAM and a T4 GPU with 15.0GB memory. This end-to-end latency includes several components: approximately 2–4 seconds for cloud storage access, 8–12 seconds for knowledge base retrieval, 6–10 seconds for VLM inference, and 3–5 seconds for structured response generation. Most responses cluster around 22 seconds, while a few are completed in as little as 10 seconds. This time distribution provides a practical reference for engineering applications where the real-time deployment is desired. For instance, Cloud I/O and retrieval currently account for nearly half of the total time, suggesting that local knowledge base deployment or improved caching strategies could substantially reduce latency. Similarly, inference time can be shortened by adopting model distillation or quantization techniques, or by deploying on dedicated GPU resources. Preliminary tests suggest that such optimizations could reduce the response time to 8–12 seconds per image. These observations provide practical guidance for improving efficiency and indicate the trade-off between computational cost and the benefits of zero-shot generalization offered by the proposed RAG-VLM framework.

Table 7 presents the complete generation process of query results for several different damage images. The RAG-grounded VLM provides formatted responses according to the user and prompt. When answering whether damage is present, clear judgments are required. Additionally, based on the description in the fourth question, we can see from the text how the model observed the damage, what it looks like, and thereby made reasonable judgments based on the observations.

Table 7. Examples of raw responses by VLM on test data.

Examples for Test Image Data Responses
Image Index: 1 Ground Truth: healthy Raw Response: "Based on the provided image and the reference information, here is the analysis: 1. How many blades are visible in the image? - One blade is visible in the image. 2. Is there visible damage on any of the turbine blades in the image? - No visible damage is observed on the turbine blade in the image. 3. If yes, what specific type of damage can be identified in this damage lists... - No specific type of damage is identified. 4. Provide a detailed description of the damage observed... - The blade appears to be in good condition with no visible signs of damage such as cracks, corrosion, erosion, rust, delamination, fractures, dents, ice, snow, surface peeling, wear, or lightning strikes/burning. The clear line along the longitudinal direction is a normal manufacturing feature rather than damage. 5. Rate the severity of the damage on a scale of 1-5, where 1 is minor and 5 is severe. - Severity rating: 1 (minor) since no damage is observed." Extracted Type: Healthy Extracted Severity: 0 Context Used: Text Docs: img_h10_chunk_0, img_h5_chunk_0, img_h15_chunk_0; Image Docs: img_h14, img_h13, img_crk2-3
Image Index: 12 Ground Truth: Lightning Strike/Burning/fracture Raw Response: "Based on the provided image and the reference information, here is the analysis: 1. How many blades are visible in the image? - The image shows one wind turbine blade. 2. Is there visible damage on any of the turbine blades in the image? - Yes, there is visible damage on the turbine blade. 3. If yes, what specific type of damage can be identified in this damage lists... - The damage observed appears to be a combination of delamination and lightning strike/burning. 4. Provide a detailed description of the damage observed... - The damage is characterized by a significant area where the surface material has been peeled away, revealing underlying layers. This is indicative of delamination. Additionally, there are signs of burning and charring around the damaged area, which suggests a lightning strike. The presence of blackened and burnt areas further supports this conclusion. 5. Rate the severity of the damage on a scale of 1-5, where 1 is minor and 5 is severe. - Based on the extent of the damage and the potential impact on the structural integrity of the blade, the severity of the damage is rated as 5 (severe)." Extracted Type: Burn Mark, Delamination, Lightning Strike/Burning

Extracted Severity: 5 Context Used: Text Docs: img_Unspecified_h10_chunk_0, img_Unspecified_h5_chunk_0, img_Unspecified_h15_chunk_0; Image Docs: img_Surface_Peeling_pel2, img_Unspecified_h12, img_Unspecified_h14
Image Index: 17 Ground Truth: crack Raw Response: "Based on the provided image and the reference information, here is the analysis: 1. How many blades are visible in the image? - Only one blade is visible in the image. 2. Is there visible damage on any of the turbine blades in the image? - Yes, there is visible damage on the turbine blade. 3. If yes, what specific type of damage can be identified... - The damage appears to be a crack. 4. Provide a detailed description of the damage observed... - The damage observed is a crack that is visible near the wide side of the blade. This is similar to the description in Similar Image #1, where a crack is noted near the tip of the blade. The lighting conditions in the image are poor, but the crack is still identifiable. 5. Rate the severity of the damage on a scale of 1-5, where 1 is minor and 5 is severe. - Based on the visual evidence, the severity of the damage is rated as 3." Extracted Type: Crack Extracted Severity: 3 Context Used: Text Docs: img_Unspecified_h10_chunk_0, img_Unspecified_h5_chunk_0, img_Unspecified_h15_chunk_0; Image Docs: img_Crack_crk2-5, img_Unspecified_hole4, img_Unspecified_hole3
Image Index: 6 Ground Truth: ice/snow Raw Response: "Based on the provided image and the reference information, here is the analysis: 1. Number of Blades Visible: The image shows three blades of a wind turbine. 2. Visible Damage: There is visible damage on one of the turbine blades. 3. Specific Type of Damage: The damage observed appears to be ice accumulation along the edge of one of the blades. This does not match the specific types of damage listed but it is closest to "Ice" and "Snow". 4. Detailed Description of Damage: The blade in question has a significant accumulation of ice along its edge. The ice formation is uneven and extends along the length of the blade, which can affect the aerodynamic performance of the turbine. The ice appears thick and jagged, indicating that it has built up over time due to cold and wet conditions. 5. Severity Rating: On a scale of 1-5, where 1 is minor and 5 is severe, the severity of the ice accumulation would be rated as a 4." Extracted Type: Wear, Ice Extracted Severity: 1 Context Used: Text Docs: img_Unspecified_h10_chunk_0, img_Unspecified_h5_chunk_0, img_Unspecified_h15_chunk_0; Image Docs: img_Crack_crk1-3, img_Crack_crk1-7, img_Unspecified_hole1
Image Index: 7 Ground Truth: peeling/corrosion/rust/erosion/crack Raw Response: "Based on the provided image and the reference information, here is the analysis: 1. Number of Blades Visible: The image shows a close-up of a single blade section. 2. Visible Damage: Yes, there is visible damage on the turbine blade in the image. 3. Specific Type of Damage: The damage observed appears to be a combination of Crack and Corrosion. 4. Detailed Description of Damage: - Crack: There are visible cracks along the edge of the blade, particularly noticeable near the tip where the material appears to be splitting. - Corrosion: There are areas showing signs of rust or corrosion, especially around the cracks and at the edges of the blade. These areas appear discolored and have a rough texture, indicating potential material degradation.

5. Severity Rating: Based on the extent and type of damage observed, the severity can be rated as 4 out of 5." Extracted Type: Corrosion, Rust, Crack Extracted Severity: 4 Context Used: Text Docs: img_Unspecified_h10_chunk_0, img_Unspecified_h5_chunk_0, img_Unspecified_h15_chunk_0; Image Docs: img_Unspecified_hole1, img_Crack_mul2, img_Unspecified_h13

3.4 Knowledge usage analysis

For each image analysis, we track which knowledge base components are used as references. As illustrated in Table 7 and Figure 8, the system selects the three most relevant text chunks and three reference images to support its assessment. In one example (Marked as index 1 in Table 7), the retrieved texts describe blades under different lighting conditions (e.g., dusk with visible damage, night with poor visibility, and cloudy daytime with no damage), helping the model account for contextual challenges. The visual references include two healthy blades and one cracked blade, which provide concrete examples for comparison. By combining these textual and visual cues, the RAG system grounds its reasoning in complementary knowledge sources and determines whether the query image more closely resembles a healthy blade or exhibits features of a specific damage type.

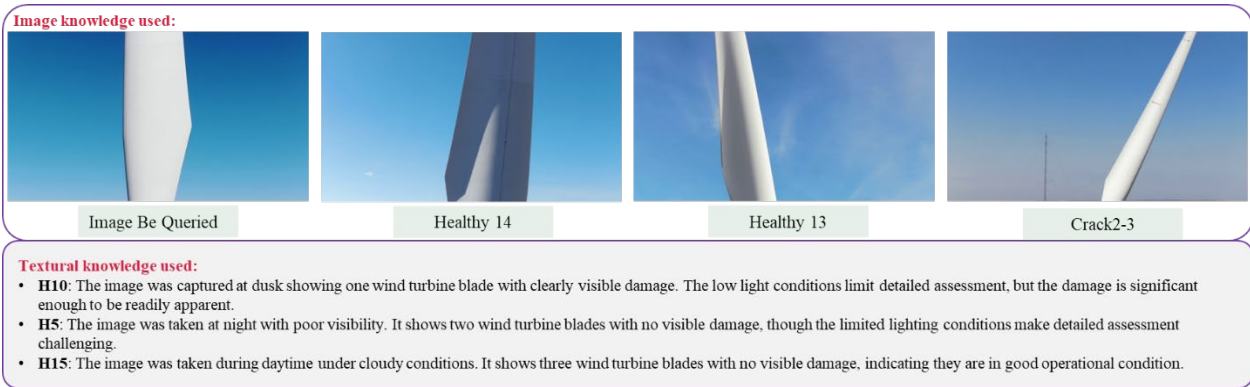


Figure 8. Knowledge usages of testing image 1.

This hybrid approach to knowledge retrieval offers several advantages. The textual knowledge provides valuable context about damage types, severity levels, and assessment challenges. The image references enable direct visual comparison, helping the model identify similar patterns or anomalies. By incorporating references from various lighting conditions, the system can better handle images taken in suboptimal environments. By tracking which knowledge sources were used, we gain insight into how the model arrived at its conclusions, making the system more transparent and trustworthy. Analysis of knowledge usage across our test dataset revealed that certain reference images and text descriptions are consistently retrieved for specific damage types, suggesting that the RAG system effectively identified visual and contextual

patterns relevant to accurate damage assessment. This consistency in knowledge retrieval also indicates that our embedding approach successfully captures both semantic meaning in text and visual features in images that are relevant to the turbine blade damage detection task.

4 Ablation Study and Interpretation

4.1 Performance comparisons to model without RAG

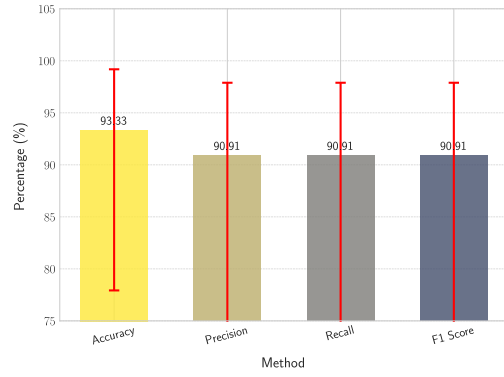
To comprehensively assess the contribution of the RAG mechanism in our wind turbine damage detection system, we conduct an ablation study by removing the RAG component while maintaining all other aspects of the model architecture. This experiment aimed to quantify the impact of the retrieval capabilities on accuracy of the model and determine whether the performance improvements justify the additional computational complexity introduced by RAG. We evaluate the non-RAG model on the same test dataset containing 30 wind turbine blade images with various damage conditions. The test set comprises 19 healthy samples and 11 damaged samples across different damage categories: Surface (8 samples), Environmental (2 samples), and Structural (1 sample). This consistent test environment enables direct comparison with the RAG-grounded VLM model.

Figure 9 presents the ablation study results for the non-RAG model. The confusion matrix reveals that the non-RAG model achieves 28 correct classifications out of 30 samples, correctly identifying 18 out of 19 healthy samples, all 8 Surface damage instances, and both Environmental damage cases. However, it fails to correctly classify the single Structural damage example and misclassifies one healthy sample as having Structural damage. The performance metrics with Clopper-Pearson confidence intervals demonstrate the statistical uncertainty in these estimates. The non-RAG model achieved an accuracy of 93.33%, [77.9%, 99.2%], while precision, recall, and F1 score all reached 90.91%, [73.5%, 97.9%]. These confidence intervals reflect the inherent uncertainty associated with the sample size and highlight the range within which the true population performance likely resides. Comparing the ablation results with RAG-grounded VLM reveals significant performance improvements. The primary limitations of the non-RAG model occur in structural damage classification, where it exhibits both false negative and false positive errors, underscoring the value of domain-specific knowledge integration through RAG.

True	Healthy	18	0	0	1
	Surface	0	8	0	0
	Environmental	0	0	2	0
	Structural	1	0	0	0
		Predicted			
		Healthy	Surface	Environmental	Structural

(a)

VLM Performance without RAG with 95% Wilson Confidence Intervals (n=30)



(b)

Figure 9. (a) Confusion matrix and (b) evaluation metrics for testing datasets.

We further analyze the two incorrectly classified images as shown in Figure 10. The first misclassification involves a false negative where a structural crack goes undetected; this image is captured in low-light conditions with poor visibility, causing the crack features to be obscured. The second misclassification is a false positive where surface contamination (dirt) is incorrectly identified as cracks. In this case, the surface irregularities introduced by the contamination create visual patterns that the model misinterprets as structural damage. These errors demonstrate the model vulnerability to challenging imaging conditions: poor visibility obscures the actual damage in the first case, while visual interference triggers false detection in the second case. These findings highlight how environmental, and imaging factors can impact detection performance in real-world applications.

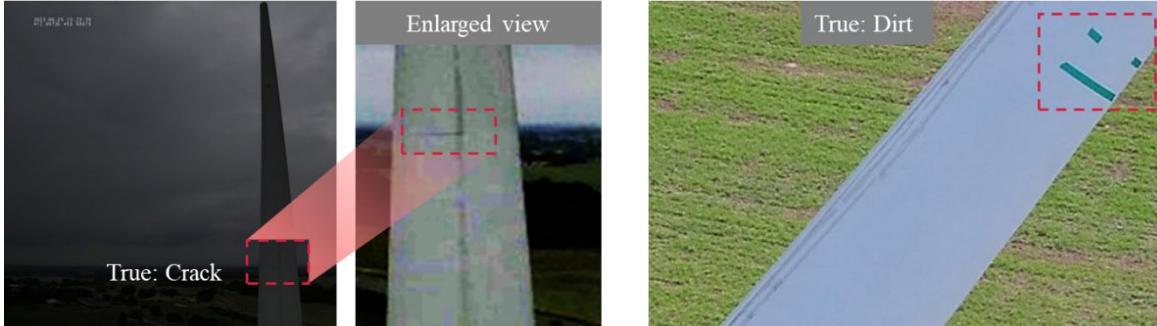


Figure 10. Images misidentified with the system without RAG.

We further examine how the proposed RAG-grounded VLM achieves correct judgments on cases missed by non-RAG methods. For example, in the response to Image 17 (Table 7), the system explicitly cites “Similar Image #1, where a crack is noted near the tip of the blade.” Here, “Image #1” refers to the most relevant retrieved reference rather than the original dataset numbering. As shown in Figure 11, this reference corresponds to Crack 2-5 in the knowledge base, an image captured at night with a visible crack near the tip. The other two retrieved references, captured in daytime with better visibility, depict healthy blades and serve as contrasts. In this case, the model’s final judgment is grounded in both the crack reference

and supporting textual description, while contrasting healthy references reinforce the classification decision. This illustrates that similarity search and reranking ensure the retrieval of relevant context, and that the final judgment reflects a comprehensive synthesis of retrieved knowledge.

Raw Responses for Image 17

Based on the image provided and the reference information, here is the analysis:

1. How many blades are visible in the image?
 - Only one blade is visible in the image.
2. Is there visible damage on any of the turbine blades in the image?
 - Yes, there is visible damage on the turbine blade.
3. If yes, what specific type of damage can be identified in this damage lists...?
 - The damage appears to be a crack.
4. Provide a detailed description of the damage observed, referencing the specific characteristics described above.
 - The damage observed is a crack that is visible near the wide side of the blade. This is similar to the description in **Similar Image #1**, where a crack is noted near the tip of the blade. The lighting conditions in the image are poor, but the crack is still identifiable.
5. Rate the severity of the damage on a scale of 1-5, where 1 is minor and 5 is severe.
 - Based on the visual evidence, the severity of the damage is rated as 3. The crack is significant enough to be readily apparent, but it does not appear to be extensive or deep enough to warrant the highest severity rating.



Figure 11. Knowledge usages of testing image 17.

The RAG approach enhances performance by enriching feature representation with contextualized domain knowledge. During inference, the model retrieves relevant documents and reference images, gaining access to specialized expertise and similar case examples that improve classification accuracy, especially for rare failure modes. These retrieval-augmented judgments justify the additional computational overhead, as the cost of misclassification in critical infrastructure far outweighs the retrieval cost. Overall, this study demonstrates that RAG-grounded VLMs provide substantial benefits for specialized visual inspection, particularly under imbalanced datasets with rare but high-risk defects. In summary, our

framework exemplifies the concept of seeing the unseen: subtle patterns under challenging conditions, rare modes without training examples, and defects beyond conventional thresholds can now be identified through knowledge-augmented zero-shot inspection, offering a reliable path toward safer infrastructure monitoring.

4.2 Performance comparisons to YOLO models

We next compare our RAG-grounded VLM with YOLO-based detectors to highlight the advantages of retrieval-augmented, zero-shot inspection over conventional supervised learning methods. YOLO [60] is chosen as the representative baseline because it is one of the most widely adopted and state-of-the-art object detection frameworks, and its zero-shot extension (YOLO-World) is directly aligned with our inspection scenario. Unlike anomaly detection models that primarily target texture-level surface irregularities, YOLO is designed for object- and region-level detection, making it more suitable for large structural components such as wind turbine blades where damages often occur at multiple scales. This makes YOLO-World a representative and relevant comparator for evaluating our zero-shot inspection framework. First, we use a YOLO world model for damage detection. At this point, since there is no training, we directly use it for detection, which is a zero-shot operation. The world model is YOLOv8s-WorldV2. We adopt the same 30 test images that span across 8 damage types. The final detection results are shown in Figure 12. From the results, we can draw two direct conclusions: First, from the object recognition level, the world model does not correctly identify the objects. Second, from the damage detection level, it also fails to identify any damage. Not only does it misidentify wind turbine blades as airplanes, but it also misclassifies some smaller damages as birds. Obviously, these detection results are unreasonable. Therefore, from the perspective of multi-scale damage detection, this world model cannot competently handle the detection task. This is reasonable, of course, because when the world model is trained, it uses community data with limited classifications. Since damage detection belongs to the second tier of multi-scale detection, it requires finer classification and expert knowledge for training.

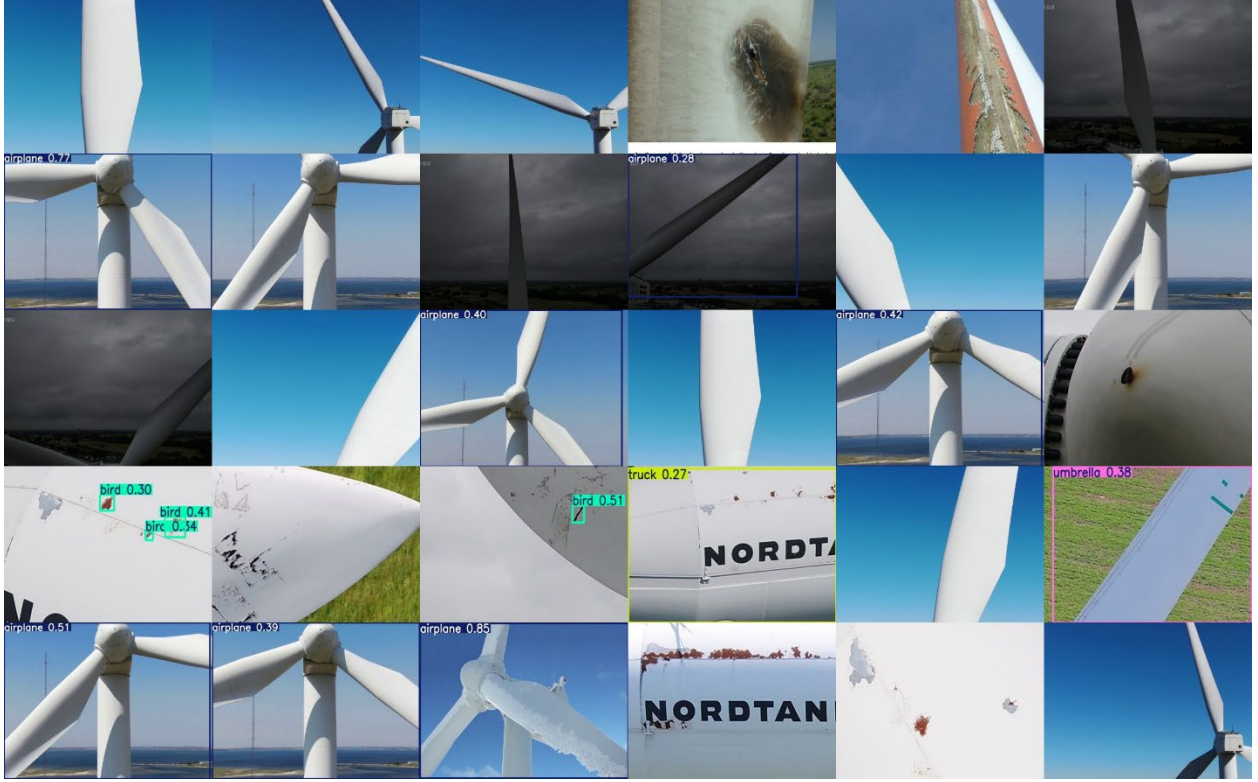


Figure 12. Detection results on testing images from YOLO world model.

From the performance of the world model above, it does not perform ideally on multi-scale damage detection tasks because the YOLO model has not been trained on damage data. In this second comparison model, we adopt 3000 images from dataset 2 to retrain the YOLO model. The model used at this time is YOLOv8n. This dataset has only two classifications: one is damage, and the other is dirt. The dataset has corresponding labels and annotations, therefore, this retraining is supervised learning. After training is completed, we test using the same 30 images. Compared to the world model, the retrained model at this time can detect damage in some images and perform boxing identification, as shown in Figure 13. However, there is a major problem: the test images where damage is identified basically belong to the same type of images as the retraining dataset (i.e., images with similar lighting conditions, angles, and damage patterns from dataset 2). Test images from different datasets (with different imaging conditions, blade types, or damage characteristics) are not identified for damage detection, indicating poor cross-domain generalization capability.

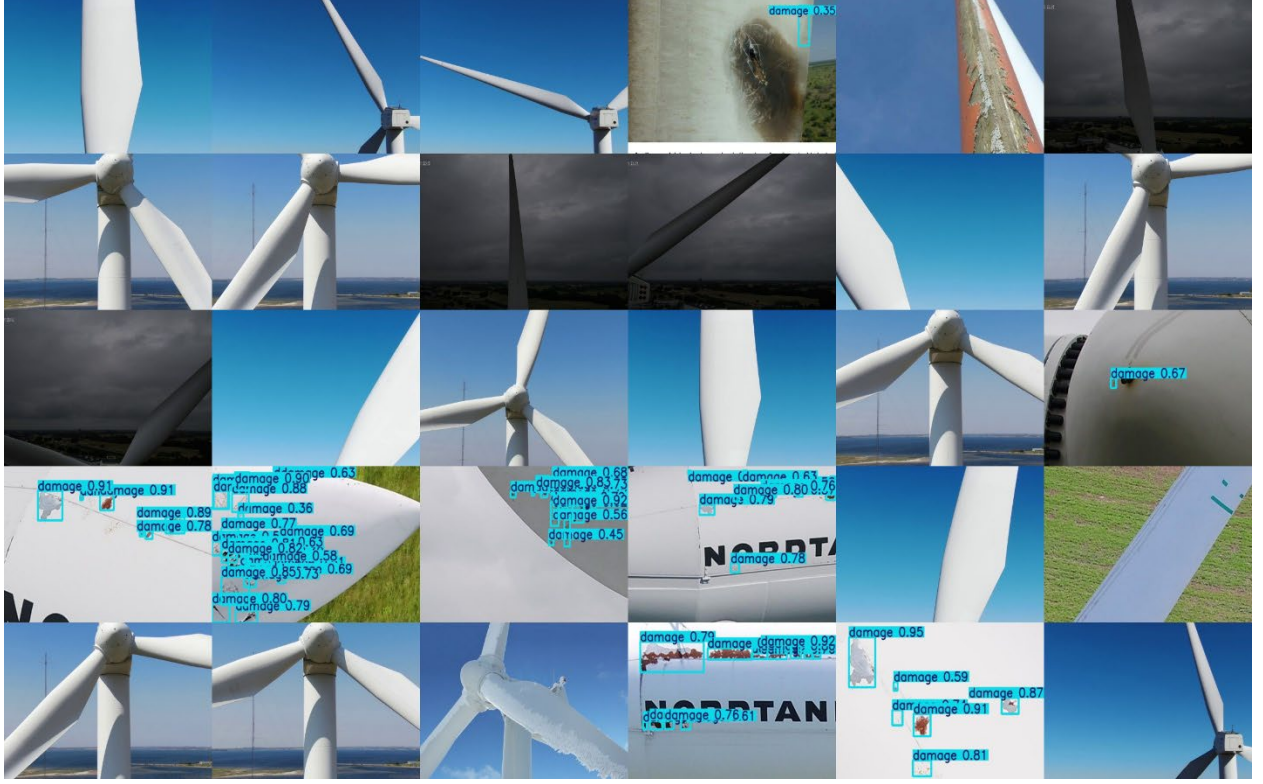


Figure 13. Detection results on testing images from retrained YOLO model.

The results above give us the insight that to classify different types of damage, we need to collect corresponding damage datasets and then retrain the model. This brings challenges to practical engineering applications. Some damage types occur rarely, and collecting sufficient datasets requires a very long cycle. In addition, damage evolves gradually, and when it evolves to a discoverable stage, we need to collect data and retrain the model again, repeating this cycle. This is why this study states that the proposed method is zero-shot, as it only requires pre-embedding expert knowledge (leveraging heterogeneous features such as damage shape, color, and possible size) and adding reference images to the database when available. At this stage, no retraining or additional data collection is required. We present the final identification results of our method in this study, as shown in Figure 14. The damage identifications appearing in each image are labeled above each image. Some images contain multiple damages, and the corresponding labels are also identified. From the results, we can see that the identified results match the actual damage. Since this is an open-source dataset, the true labels are known.

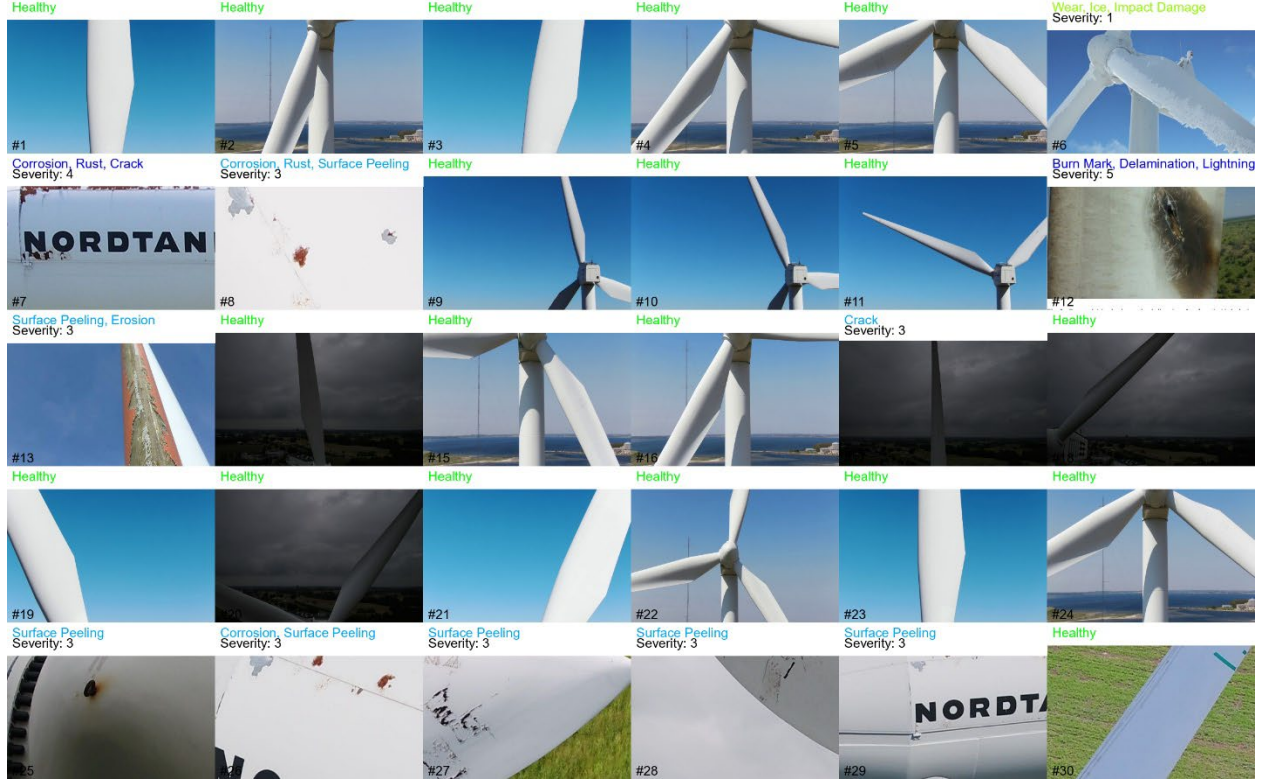


Figure 14. Detection results on testing images from proposed approach.

These comparative results validate our core hypothesis that RAG-grounded VLMs overcome fundamental limitations of existing approaches. Supervised methods like YOLOv8n fail catastrophically on unseen images and require collecting new datasets and retraining for each novel damage type, making them impractical for real-world deployment where damage patterns continuously evolve. While YOLO-World offers nominal zero-shot capability, it lacks domain expertise for specialized industrial tasks. Both YOLO models, on our 30-image test set, fail to recognize most damages, often skipping images entirely. As a result, although YOLO models achieve extremely fast inference speed (within 1 second per image), this advantage is of little practical value here because valid outputs were too few to allow a meaningful efficiency comparison. Moreover, the long-term cost of supervised pipelines remains high: dataset collection, annotation, and repeated retraining are unavoidable whenever new damage types appear, which greatly reduces sustainability in industrial practice. In contrast, our RAG-grounded VLM achieves towards zero-shot inspection: it operates without any blade-specific training, generalizes across previously unseen datasets and novel damage types, and produces comprehensive inspection reports rather than simple classifications, making it more suitable for industrial applications where adaptability and interpretability are crucial. Although the embedding models inherently leverage broad pretraining corpora, the inspection task itself remains zero-shot under the accepted VLM definition, where zero-shot denotes the absence of task-specific training rather than the exclusion of all domain-related knowledge. Overall, the framework

can scale effectively to larger datasets and continuously expand knowledge bases without retraining, and its modular design supports transfer to related industrial domains such as bridge inspection, corrosion monitoring, and manufacturing defect detection. These properties highlight the generalizability of the approach and its potential for broader deployment in knowledge-driven inspection and optimization tasks.

5 Conclusion

This study introduces an effective framework for wind turbine blade inspection towards zero-shot by integrating Retrieval-Augmented Generation with Vision Language Models. The framework demonstrates considerable advantages for specialized industrial inspection tasks, particularly in contexts where labeled training data is limited or challenging to obtain. Our experimental results reveal that the RAG-enhanced approach achieved perfect classification accuracy (100%) across all performance metrics when evaluated on a diverse dataset of 30 wind turbine blade images encompassing healthy blades, surface damage, environmental damage, and structural damage. In contrast, the baseline model without RAG achieved 93.33% accuracy with 90.91% precision, recall, and F1 score. Uncertainty analysis is also performed to assess the robustness of the framework under small-sample testing scenarios. The ablation study clearly identified the value added by the RAG component, particularly for detecting structural damage in challenging imaging conditions. The hybrid retrieval mechanism implemented in our framework effectively leverages both textual and visual information from a domain-specific knowledge base. The reranking algorithm further improves retrieval quality by prioritizing the most relevant context based on query relevance and document specificity. This approach enables the model to access specialized knowledge about damage characteristics without requiring extensive labeled examples. Analysis of misclassified cases in the non-RAG model revealed particular challenges with low-light imaging environments and surface contamination. The RAG enhancement effectively compensates for these limitations by providing relevant contextual information during inference, demonstrating the value of domain knowledge integration for challenging cases. While our framework shows promising results, future work should expand testing to larger and more diverse datasets, explore optimization of the retrieval mechanism for real-time applications, and investigate the minimum knowledge base requirements needed to maintain high performance. In addition, the use of bounding boxes for damage localization will be further investigated. Through the integration of retrieval-augmented generation with vision language models, we enable systems to see what is previously unseen, offering an effective solution for near zero-shot inspection that transcends the limitations of labeled data while ensuring critical infrastructure safety.

Data availability

All related data and implementations are open-sourced at <https://github.com/yangzhang10/Wind-Turbine-Inspection-with-VLM-and-RAG-ASOC-.git>.

Acknowledgment

This research is supported in part by the U.S. Economic Development Administration through the Connecticut Manufacturing Simulation Center (CMSC) at the University of Connecticut, and in part by the National Science Foundation under grant 2434519.

References

- [1] Mishnaevsky Jr, L., Hasager, C.B., Bak, C., Tilg, A.M., Bech, J.I., Rad, S.D. and Fæster, S., 2021. Leading edge erosion of wind turbine blades: Understanding, prevention and protection. *Renewable Energy*, 169, pp.953-969.
- [2] Kong, K., Dyer, K., Payne, C., Hamerton, I. and Weaver, P.M., 2023. Progress and trends in damage detection methods, maintenance, and data-driven monitoring of wind turbine blades—A review. *Renewable Energy Focus*, 44, pp.390-412.
- [3] Gohar, I., Yew, W.K., Halimi, A. and See, J., 2025. Review of state-of-the-art surface defect detection on wind turbine blades through aerial imagery: Challenges and recommendations. *Engineering Applications of Artificial Intelligence*, 144, p.109970.
- [4] Jiang, L., Zhang, S.P., Shen, G.Q. and Zhou, L., 2025. Acoustic Emission-based wind turbine blade icing monitoring using deep learning technology. *Renewable Energy*, p.122980.
- [5] Meng, M., Chua, Y.J., Wouterson, E. and Ong, C.P.K., 2017. Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. *Neurocomputing*, 257, pp.128-135.
- [6] Mendikute, J., Carmona, I., Aizpurua, I., Bediaga, I., Castro, I., Galdos, L. and Lanzagorta, J.L., 2025. Defect detection in wind turbine blades applying convolutional neural networks to ultrasonic testing. *NDT & E International*, 154, p.103359.
- [7] Wang, C. and Gu, Y., 2022. Research on infrared nondestructive detection of small wind turbine blades. *Results in Engineering*, 15, p.100570.
- [8] Memari, M., Shakyia, P., Shekaramiz, M., Seibi, A.C. and Masoum, M.A.S., 2024. Review on the advancements in wind turbine blade inspection: Integrating drone and deep learning technologies for enhanced defect detection. *IEEE Access*, 12, pp.33236-33282.
- [9] Zhang, S., He, Y., Gu, Y., He, Y., Wang, H., Wang, H., Yang, R., Chady, T. and Zhou, B., 2024. UAV based defect detection and fault diagnosis for static and rotating wind turbine blade: A review. *Nondestructive Testing and Evaluation*, pp.1-39.
- [10] Aggarwal, A.K. and Chauhan, A.P.S., 2025. Robust feature extraction from omnidirectional outdoor images for computer vision applications. *International Journal of Instrumentation and Measurement*, 10.

- [11] Liu, L., Li, P., Wang, D. and Zhu, S., 2024. A wind turbine damage detection algorithm designed based on YOLOv8. *Applied Soft Computing*, 154, p.111364.
- [12] Wu, Z., Zhang, Y., Wang, X., Li, H., Sun, Y. and Wang, G., 2024. Algorithm for detecting surface defects in wind turbines based on a lightweight YOLO model. *Scientific Reports*, 14(1), p.24558.
- [13] Hang, X., Zhu, X., Gao, X., Wang, Y. and Liu, L., 2024. Study on crack monitoring method of wind turbine blade based on AI model: Integration of classification, detection, segmentation and fault level evaluation. *Renewable Energy*, 224, p.120152.
- [14] Jia, X. and Chen, X., 2024. AI-based optical-thermal video data fusion for near real-time blade segmentation in normal wind turbine operation. *Engineering Applications of Artificial Intelligence*, 127, p.107325.
- [15] Zhou, W., Wang, Z., Zhang, M. and Wang, L., 2023. Wind turbine actual defects detection based on visible and infrared image fusion. *IEEE Transactions on Instrumentation and Measurement*, 72, pp.1-8.
- [16] Rizk, P., Rizk, F., Karganroudi, S.S., Ilinca, A., Younes, R. and Khoder, J., 2024. Advanced wind turbine blade inspection with hyperspectral imaging and 3D convolutional neural networks for damage detection. *Energy and AI*, 16, p.100366.
- [17] Dwivedi, D., Babu, K.V.S.M., Yemula, P.K., Chakraborty, P. and Pal, M., 2024. Identification of surface defects on solar PV panels and wind turbine blades using attention based deep learning model. *Engineering Applications of Artificial Intelligence*, 131, p.107836.
- [18] Li, W., Zhao, W. and Du, Y., 2025. Large-scale wind turbine blade operational condition monitoring based on UAV and improved YOLOv5 deep learning model. *Mechanical Systems and Signal Processing*, 226, p.112386.
- [19] Jia, X. and Chen, X., 2025. Unsupervised wind turbine blade damage detection with memory-aided denoising reconstruction. *IEEE Transactions on Industrial Informatics*, 21(1), pp.762-770.
- [20] Lei, X., Sun, M., Zhao, R., Wu, H., Zhou, Z., Dong, Y. and Sun, L., 2024. Unsupervised vision-based structural anomaly detection and localization with reverse knowledge distillation. *Structural Control and Health Monitoring*, 2024(1), p.8933148.
- [21] Foster, A., Best, O., Gianni, M., Khan, A., Collins, K. and Sharma, S., 2022. Drone footage wind turbine surface damage detection. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pp.1-5. IEEE.
- [22] Ataei, S.T., Zadeh, P.M. and Ataei, S., 2025. Vision-based autonomous structural damage detection using data-driven methods. *arXiv preprint, arXiv:2501.16662*.

- [23] Zhang, Y., Wang, L., Huang, C. and Luo, X., 2025. Wind turbine blade defect detection based on the genetic algorithm-enhanced YOLOv5 algorithm using synthetic data. *IEEE Transactions on Industry Applications*, 61(1), pp.653-665. doi: 10.1109/TIA.2024.3481190.
- [24] Ye, X., Wang, L., Huang, C. and Luo, X., 2024. Wind turbine blade defect detection with a semi-supervised deep learning framework. *Engineering Applications of Artificial Intelligence*, 136, p.108908.
- [25] Sheiati, S., Jia, X., McGugan, M., Branner, K. and Chen, X., 2024. Artificial intelligence-based blade identification in operational wind turbines through similarity analysis aided drone inspection. *Engineering Applications of Artificial Intelligence*, 137, p.109234.
- [26] Iyer, A., Nguyen, L. and Khushu, S., 2022. Learning to identify cracks on wind turbine blade surfaces using drone-based inspection images. *arXiv preprint, arXiv:2207.11186*.
- [27] Nguyen, L., Iyer, A. and Khushu, S., 2022. An automated system for detecting visual damages of wind turbine blades. *arXiv preprint, arXiv:2205.10954*.
- [28] Yang, C., Liu, X., Zhou, H., Ke, Y. and See, J., 2023. Towards accurate image stitching for drone-based wind turbine blade inspection. *Renewable Energy*, 203, pp.267-279.
- [29] Li, T., Luan, Y., Pang, Z. and Zhang, W., 2024. Structural digital twin modeling and adaptive pretrain-finetune learning for dynamic impact identification on wind turbine blades. *IEEE Transactions on Industrial Informatics*, 20(8), pp.10292-10303.
- [30] Jiang, G., Yue, R., He, Q., Xie, P. and Li, X., 2023. Imbalanced learning for wind turbine blade icing detection via spatio-temporal attention model with a self-adaptive weight loss function. *Expert Systems with Applications*, 229, p.120428.
- [31] Zhou, Q., Pang, G., Tian, Y., He, S. and Chen, J., 2023. AnomalyClip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint, arXiv:2310.18961*.
- [32] Gu, Z., Zhu, B., Zhu, G., Chen, Y., Li, H., Tang, M. and Wang, J., 2024a. FILO: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp.2041-2049.
- [33] Soleimani-Babakamali, M.H., Soleimani-Babakamali, R., Nasrollahzadeh, K., Avci, O., Kiranyaz, S. and Taciroglu, E., 2023. Zero-shot transfer learning for structural health monitoring using generative adversarial networks and spectral mapping. *Mechanical Systems and Signal Processing*, 198, p.110404.
- [34] Soleimani-Babakamali, M.H., Soleimani-Babakamali, R., Kashfi-Yeganeh, A., Nasrollahzadeh, K., Avci, O., Kiranyaz, S. and Taciroglu, E., 2025. Multi-source transfer learning for zero-shot structural damage detection. *Applied Soft Computing*, 169, p.112519.
- [35] Xiong, Q., Kong, Q., Xiong, H., Chen, J., Yuan, C., Wang, X. and Xia, Y., 2024. Zero-shot knowledge transfer for seismic damage diagnosis through multi-channel 1D CNN integrated with autoencoder-based domain adaptation. *Mechanical Systems and Signal Processing*, 217, p.111535.

- [36] Chen, M., Mangalathu, S. and Jeon, J.S., 2024a. Rapid damage state identification of structures using generalized zero-shot learning method. *Earthquake Engineering & Structural Dynamics*, 53(14), pp.4269-4286.
- [37] Gu, Z., Zhu, B., Zhu, G., Chen, Y., Tang, M. and Wang, J., 2024b. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3), pp.1932-1940.
- [38] Chen, W., Yan-yi, L., Tie-zheng, G., Da-peng, L., Tao, H., Zhi, L., Qing-wen, Y., Hui-han, W. and Ying-you, W., 2024b. Systems engineering issues for industry applications of large language model. *Applied Soft Computing*, 151, p.111165.
- [39] Walker, C., Rothon, C., Aslansefat, K., Papadopoulos, Y. and Dethlefs, N., 2024a. SafeLLM: Domain-specific safety monitoring for large language models: A case study of offshore wind maintenance. *arXiv preprint, arXiv:2410.10852*.
- [40] Pastoriza, S., Yousfi, I., Redino, C., Vucovich, M., Rahman, A., Aguinaga, S. and Nandakumar, D., 2025. Retrieval augmented anomaly detection (RAAD): Nimble model adjustment without retraining. *arXiv preprint, arXiv:2502.19534*.
- [41] Thimonier, H., Popineau, F., Rimmel, A. and Doan, B.L., 2024. Retrieval augmented deep anomaly detection for tabular data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp.2250-2259.
- [42] Wang, H. and Li, Y.F., 2023. Large language model empowered by domain-specific knowledge base for industrial equipment operation and maintenance. In *2023 5th International Conference on System Reliability and Safety Engineering (SRSE)*, pp.474-479. IEEE.
- [43] Walker, C., Rothon, C., Aslansefat, K., Papadopoulos, Y. and Dethlefs, N., 2024b. Using large language models to recommend repair actions for offshore wind maintenance. In *Journal of Physics: Conference Series*, 2875(1), p.012025. IOP Publishing.
- [44] Alnegheimish, S., Nguyen, L., Berti-Equille, L. and Veeramachaneni, K., 2024. Large language models can be zero-shot anomaly detectors for time series?. *arXiv preprint, arXiv:2405.14755*.
- [45] Jose, S., Nguyen, K.T., Medjaher, K., Zemouri, R., Lévesque, M. and Tahan, A., 2024. Advancing multimodal diagnostics: Integrating industrial textual data and domain knowledge with large language models. *Expert Systems with Applications*, 255, p.124603.
- [46] Bonomo, M. and Bianco, S., 2025. Visual RAG: Expanding MLLM visual knowledge without fine-tuning. *arXiv preprint arXiv:2501.10834*.
- [47] Bhat, N.N., Mondal, J. and Sarkar, S., 2025, July. ExpertNeurons at SciVQA-2025: Retrieval Augmented VQA with Vision Language Model (RAVQA-VLM). In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)* (pp. 221-229).

- [48] Dong, Y., Ueda, N., Boros, K., Ito, D., Sera, T. and Oyamada, M., 2025. SCAN: Semantic Document Layout Analysis for Textual and Visual Retrieval-Augmented Generation. arXiv preprint arXiv:2505.14381.
- [49] Khan, F.F., Chen, J., Mohamed, Y., Feng, C.M. and Elhoseiny, M., 2025. VR-RAG: Open-vocabulary Species Recognition with RAG-Assisted Large Multi-Modal Models. arXiv preprint arXiv:2505.05635.
- [50] Zheng, X., Weng, Z., Lyu, Y., Jiang, L., Xue, H., Ren, B., Paudel, D., Sebe, N., Van Gool, L. and Hu, X., 2025. Retrieval augmented generation and understanding in vision: A survey and new outlook. arXiv preprint arXiv:2503.18016.
- [51] Chase, H., 2022. LangChain. Available at: <https://github.com/langchain-ai/langchain>
- [52] Reimers, N. and Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv preprint, arXiv:1908.10084.
- [53] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning, pp.8748-8763. PMLR.
- [54] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L. and Jégou, H., 2024. The FAISS library. arXiv preprint, arXiv:2401.08281.
- [55] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C. and Zhou, J., 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint, arXiv:2308.12966.
- [56] Chen, X., 2023. Drone-based optical and thermal videos of rotor blades taken in normal wind turbine operation. IEEE Dataport. doi: <https://dx.doi.org/10.21227/yzs5-1067>.
- [57] Chen, X., 2024. Dataset for AI-based optical-thermal video data fusion for near real-time blade segmentation in normal wind turbine operation. Mendeley Data, V1. doi: 10.17632/9rcf5p89zn.1
- [58] Shihavuddin, A. and Chen, X., 2018. DTU - Drone inspection images of wind turbine.
- [59] Wang, W., Xue, Y., He, C. and Zhao, Y., 2022. Review of the typical damage and damage-detection methods of large wind turbine blades. *Energies*, 15(15), p.5672.
- [60] Ultralytics. (2023). YOLOv8. GitHub repository. <https://github.com/ultralytics/ultralytics>