

The Emerging Role of Vision-Language Models in the Automation of Railway Asset Management: A Review and Future Perspective

Ashley Varghese^{1,2}, Mohammadjavad Ghorbanalivaki², Gunho Sohn²

¹ Canadian National Railway, Canada – ashley.varghese@cn.ca

² Dept. of Earth and Space Science and Engineering, York University, Toronto, Ontario, Canada
(ashleyva, mvakili, gsohn)@yorku.ca

Keywords: Vision-Language Models (VLM), Railway Asset Management, Predictive Maintenance, Rail Automated Inspection, Open-vocabulary Detection

Abstract

The safety, efficiency, and longevity of global railway networks are directly linked to the rigorous inspection and management of their vast inventory of physical assets. Over the past decade, the field has progressed from manual surveys to automated systems leveraging imagery from track-based or aerial platforms. These systems predominantly built on traditional Computer Vision (CV) models have proven effective at detecting a pre-defined set of common assets. However, this progress has exposed a fundamental architectural and operational ceiling: the *closed-world assumption*. Current models are constrained to a fixed catalogue of classes defined during their training. It makes the model incapable of identifying novel objects or adapting to environmental changes without costly and continuous cycles of data re-annotation, retraining, and redeployment. This review paper argues that Vision-Language Models (VLMs), a paradigm whose rapid maturation is evidenced by recent comprehensive surveys offer a transformative solution. We provide a focused overview of the limitations of current CV systems and map the mechanics of a VLM-powered approach specifically Open-Vocabulary Detection and Reasoning Segmentation directly to the outstanding challenges in rail asset management. Ultimately, the literature suggests that the adoption of VLMs could catalyze a fundamental shift in railway infrastructure management that serves as a key enabler for next-generation Predictive Maintenance and autonomous Digital Twins.

1. Introduction

The global railway network constitutes the backbone of sustainable terrestrial transport facilitating the movement of billions of passengers and gigatons of freight annually. As nations strive to decarbonize their logistics chains and shift modal share away from road and air, the demand for rail capacity continues to rise driven by the global imperative for sustainable logistics (Kljaić et al., 2023). However, the safety, efficiency, and economic viability of these expanding networks are inextricably linked to the structural health of their physical assets. Railway infrastructure is a complex hierarchical system-of-systems comprising superstructural elements such as rails, sleepers, fasteners, and switches; substructural foundations and critical electrification systems including catenary wires and pantographs. The degradation of these assets is inevitable driven by the intensification of traffic loads, higher operating speeds, and the increasing severity of climate-change-related events. Infrastructure Managers (IMs) are consequently faced with a dual challenge: they must maintain an aging asset base to rigorous safety standards while simultaneously minimizing service disruptions caused by maintenance windows (Thaduri

et al., 2015). Historically, this balance was managed through corrective or preventive maintenance strategies. However, to meet modern availability demands, the industry is currently undergoing a decisive paradigm shift towards Predictive Maintenance (PdM) (Bianchi et al., 2025). PdM is a proactive strategy that relies on continuous high-frequency monitoring of asset conditions to forecast failures before they occur.

To support the data-intensive requirements of PdM, the railway domain has aggressively adopted remote sensing technologies. This fundamentally altering the nature of infrastructure inspection. The era of the “track-walker” relying solely on human visual inspection has largely been superseded by automated data acquisition platforms. Modern inspection campaigns routinely deploy high-speed diagnostic trains (such as the Doctor Yellow class in Japan) and Unmanned Aerial Vehicles (UAVs) equipped with synchronized sensor arrays (Olivier et al., 2025). A notable industry example is the Canadian National Railway’s (CN) Autonomous Track Inspection Program (ATIP), which utilizes automated boxcars equipped with advanced optical and laser sensors to inspect track geometry and component conditions at operational speeds (Canadian National Railway, 2022). Initiatives like ATIP and wayside inspection portals have resulted in a data deluge where IMs generate petabytes of visual data annually. Although data acquisition capabilities have matured, the analytical pipelines required to process these data remain a significant bottleneck. As noted by (Tang et al., 2022), the sheer volume of imagery renders manual review impossible. This leads to the use of Automated Defect Detection (ADD) systems to filter and prioritize maintenance needs.

For the past decade, the automation of visual inspection has relied predominantly on Deep Learning (DL) models. The commonly used DL models are specifically Convolutional Neural Networks (CNNs) such as the YOLO, ResNet, and Mask R-CNN families (Oh et al., 2022). These supervised learning models have achieved state-of-the-art performance in extracting complex rail paths (Ghorbanalivakili and Sohn, 2025) and detecting specific well-defined defects such as broken fasteners, surface cracks, or missing clips. However, despite their success in controlled tests, these traditional computer vision systems are fundamentally constrained by the closed-world assumption. This architectural limitation assumes that all possible classes of objects and defects are known *a priori* and are adequately represented in the training data. In the highly dynamic and unstructured environment of a railway corridor, this assumption fails for several critical reasons. First, the long-tail problem implies that common defects are abundant in training datasets. Conversely, rare anomalies appear too

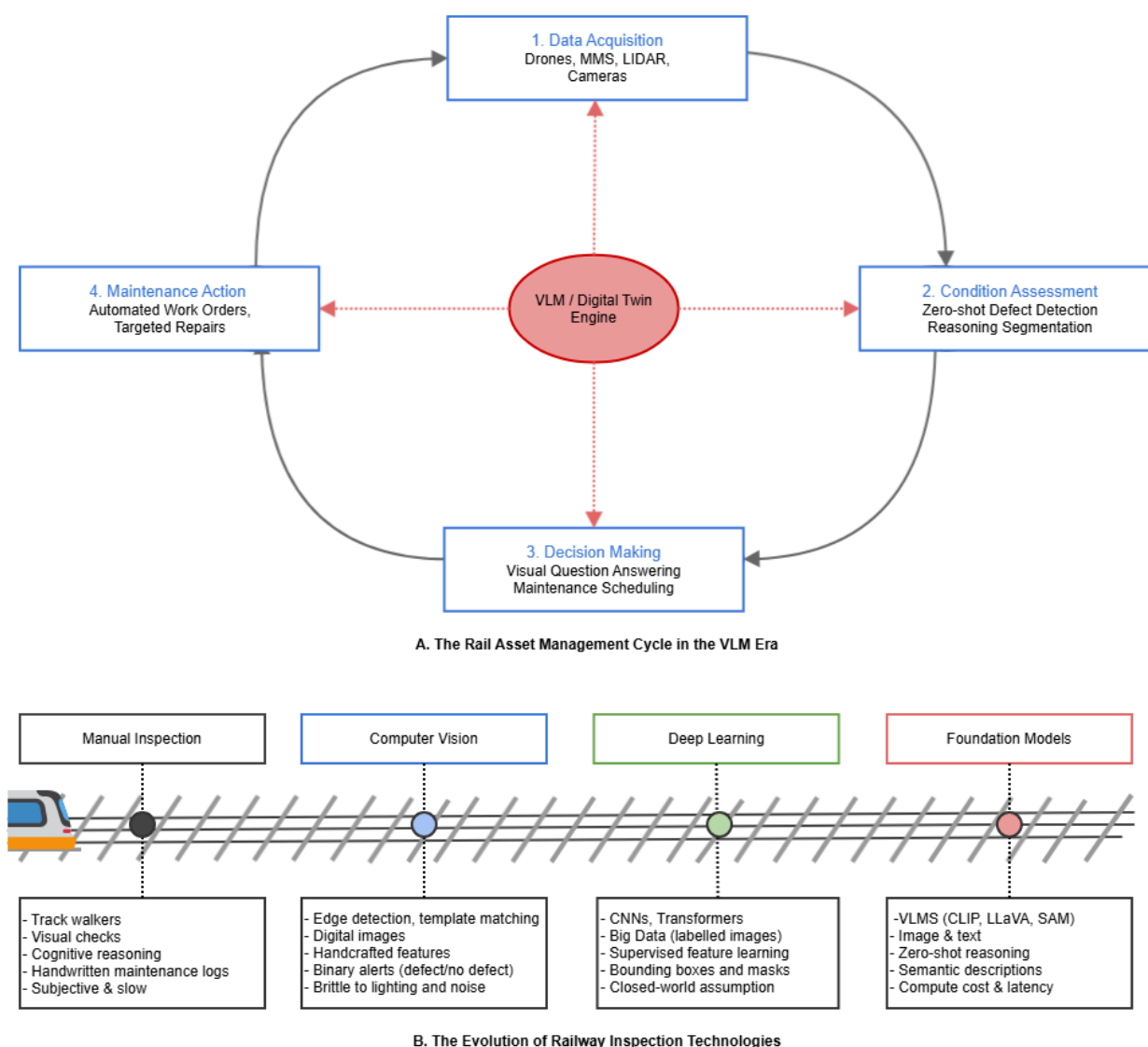


Figure 1. **The Paradigm Shift in Railway Asset Management.** (A) A conceptual framework for next-generation asset management wherein a central Vision-Language Model (VLM) and Digital Twin engine process multimodal inputs to drive continuous condition assessment, automated decision-making, and targeted maintenance. (B) The evolutionary timeline of railway inspection technologies illustrating the progression from subjective manual surveys and closed-world Deep Learning to open-vocabulary Foundation Models capable of zero-shot reasoning and semantic reporting.

infrequently to train robust supervised models (Weng et al., 2025). These rare instances include specific types of foreign object debris, unusual vegetation encroachment, or structural deformation caused by a landslide. Second, CNNs exhibit semantic rigidity; they perform pixel-level classification but lack contextual understanding. A standard model might identify a metal bar, but it cannot reason that a “metal bar lying *across* the tracks” represents a critical safety hazard, whereas a “metal bar *beside* the tracks” does not. Finally, the costs of adaptation are prohibitive. Adapting a closed-world model to a new environment such as a new climate zone with snow or a new type of sleeper material typically requires collecting large-scale annotated datasets to retrain the model from scratch (Wang and Hong, 2025). Without this extensive retraining standard deep learning models suffer from catastrophic forgetting where the adaptation to the new domain degrades performance on the

original distribution (Lourenço et al., 2026)

This rigidity results in high false-negative rates for edge cases and prevents the full automation of asset management. Human experts must still intervene to interpret complex or novel scenarios. Consequently, we posit that the railway domain stands on the precipice of a transformative shift driven by Vision-Language Models (VLMs). Departing from the fixed-label constraints of traditional CNNs, foundation models such as CLIP (Radford et al., 2021) and its successors are pre-trained on webscale datasets of image-text pairs. This allows them to learn a joint embedding space for visual concepts and natural language enabling zero-shot detection. In a railway context, this means an inspection system is no longer limited to a fixed catalogue of defects. Instead, an operator can query the system using natural language prompts such as “Identify any

obstruction on the track” or *“Locate areas with heavy ballast fouling”* and the model can identify these features without explicit prior training on those specific categories (Abdelhamed et al., 2024). This capability moves the field from rigid pattern recognition to flexible semantic understanding.

This review paper provides a comprehensive analysis of the potential for VLMs to revolutionize railway asset management. We begin by reviewing the architectural limitations of current railway Computer Vision (CV) systems. This review identifies the “closed-world” constraint as the primary barrier to fully automated inspection. We then synthesize recent VLM literature (Sun et al., 2024; Zhang et al., 2024a) and discuss the applications of VLM. Finally, we propose a forward-looking perspective on reasoning segmentation (Lai et al., 2024). This outlines how VLMs can be used to generate automated descriptive condition reports. Ultimately, we envision that such reporting mechanisms could integrate directly into Digital Twin ecosystems to support the next generation of predictive maintenance.

2. Rail Infrastructure Management and Monitoring

The management of railway infrastructure represents a complex challenge in civil engineering. A railway network acts as a distributed system composed of heterogeneous components. These include rails, sleepers, fasteners, and complex electrification grids. These elements interact continuously under varying operational loads and environmental conditions (Tang et al., 2022). The complexity is exacerbated by the vast geographical scale of national networks. Tracks traverse diverse geological terrains and climate zones. Each zone introduces unique degradation mechanisms such as thermal expansion in heat or brittle fracture in sub-zero environments (Bianchi et al., 2025).

Consequently, infrastructure monitoring cannot be treated as a series of isolated inspection tasks. Instead, it requires a holistic management framework. This framework must integrate technical health with operational objectives. It must account for interdependencies between physical infrastructure and service availability. The primary goal is to ensure the system’s dependability by rigorously analyzing its reliability, availability, maintainability, and safety (RAMS). RAMS is a framework essential for determining optimal maintenance intervals and minimizing lifecycle costs (Hidirov and Guler, 2019).

2.1 The Asset Management Lifecycle

The lifecycle of railway asset management is a continuous, iterative process. It is designed to maintain infrastructure in a state of good repair while optimizing lifecycle costs. This cycle encompasses distinct stages, ranging from strategic planning to data-driven decision-making.

- **Planning:** The lifecycle begins with planning. This phase dictates strategic resource allocation and long-term maintenance scheduling. Planning is constrained by the need to perform work on active lines. This often requires possessions or maintenance windows where traffic is suspended. Planners must analyze historical traffic density and topological surveys. They must balance immediate maintenance costs against the long-term risk of asset failure (Tang et al., 2022).
- **Design and Construction (Building):** This phase encompasses the construction and renewal of infrastructure. Precision here is paramount. Initial construction quality is the primary determinant of an asset’s degradation curve. Modern engineering relies on Building Information Modeling (BIM) to create digital twins. These serve as virtual baselines for the physical asset (Thompson et al., 2025).
- **Inspection and Monitoring (Perception):** Perception forms the sensory foundation of management. It involves acquiring raw empirical data regarding the physical state of the track. Historically, this depended on visual observation. Today, it has evolved into a data-intensive process using multi-sensor arrays. Modern regimes employ diagnostic trains and autonomous vehicles. These platforms utilize laser profilometers, accelerometers, and ultrasonic sensors to generate the digital truth of the railway corridor (Bianchi et al., 2025)(Canadian National Railway, 2022).
- **Safety Assurance:** Safety is the non-negotiable constraint for all operations. It involves the continuous assessment of risk. Engineers must ensure that degradation mechanisms do not reach critical thresholds. For example, rail contact fatigue must be managed before it leads to derailment. Safety management relies on strict standards which define intervention limits for track parameters (Thaduri et al., 2015).
- **Change Detection:** Infrastructure degradation is typically cumulative rather than sudden. Change detection monitors these subtle evolutions over time. Engineers track the rate of change in parameters like longitudinal level and alignment. This process identifies deviations from a baseline golden state. Effective change detection identifies developing trends such as ballast shifting before they manifest as critical defects (Olivier et al., 2025).
- **Defect Detection:** While change detection monitors trends, defect detection isolates specific pathologies. This involves localizing and classifying physical anomalies. Common defects include squats, shelling, cracked sleepers, and missing fasteners. This stage requires high-resolution scrutiny. Even minor surface defects can propagate rapidly under dynamic loading. Accurate categorization is essential for selecting the appropriate remedial action (Oh et al., 2022).
- **Condition Assessment (Diagnosis):** Condition assessment bridges detection and decision-making. It synthesizes raw findings into a holistic view of asset health. This involves signal processing and statistical analysis. Raw inputs are converted into meaningful condition indices such as the Track Quality Index (TQI). This stage filters measurement noise to diagnose the root cause of anomalies (Vale and Simões, 2022).
- **Decision Making:** This stage represents the engineering equivalent of reasoning. Once a condition is diagnosed, engineers apply maintenance standards and economic constraints. They determine the necessary course of action. Decisions prioritize interventions based on criticality. Options include immediate emergency repair, scheduled maintenance, or deferred renewal. (De Donato et al., 2023).

2.2 Evolution of Maintenance Paradigms

The execution of the lifecycle described above is governed by the maintenance strategy in place. Historically, these strategies have evolved through three distinct generations.

2.2.1 First Generation: Corrective Maintenance

Corrective Maintenance (CM) or run-to-failure represents the reactive baseline. Actions are triggered only after a functional failure occurs. It is unsuitable for maintenance of safety-critical assets though CM minimizes upfront monitoring costs. For example, failure of a primary component such as a rail break can cause catastrophic derailments. Therefore CM is restricted to non-critical auxiliary components (Kljaić et al., 2023).

2.2.2 Second Generation: Preventive Maintenance

To mitigate the risks of CM, the industry adopted Preventive Maintenance (PvM). This approach relies on fixed cyclical intervals to replace components. It operates regardless of the actual physical condition. However, Bianchi et al. (2025) argue that PvM suffers from economic inefficiencies. It is influenced by localized variables like subgrade stiffness and micro-climates. PvM leads to two adverse outcomes: over-maintenance (wasting resources on healthy assets) and under-maintenance (failing to catch rapid degradation between intervals).

2.2.3 Third Generation: Predictive Maintenance (PdM)

The current state-of-the-art is Predictive Maintenance (PdM). Unlike PvM, PdM uses condition data to forecast the future state of the asset. It relies on modeling the P-F Interval, the time between detecting a potential failure (P) and a functional failure (F). By estimating the Remaining Useful Life (RUL), managers can schedule interventions at the optimal economic moment (Thaduri et al., 2015). However, PdM requires a holistic understanding of the visual environment. This necessitates advanced automated detection systems to process the massive volumes of visual data generated by modern inspection platforms (Tang et al., 2022).

2.3 The Role of Digital Twins

To synthesize the massive influx of heterogeneous data ranging from 1D sensor telemetry and 2D inspection imagery to 3D LiDAR point clouds, the railway domain is increasingly adopting **Digital Twin (DT)** technology. A DT is defined as a dynamic virtual replica of a physical system that updates in real-time to reflect the asset's current state throughout its lifecycle (Thompson et al., 2025). In the railway context, a DT acts as a central semantic repository which integrating static design data (e.g., BIM models) with dynamic condition data (e.g., sensor readings and maintenance logs).

Recent reviews highlight a critical bottleneck in current DT frameworks (De Donato et al., 2023). Traditional DTs excel at ingesting structured time-series telemetry. However, they struggle with unstructured visual data. Bridging the gap between raw imagery and semantic digital records currently requires rigid Computer Vision models. These models are limited by their inability to recognize novel degradation patterns without extensive retraining. This limitation underscores the need for the Vision-Language integration discussed in this paper.

3. The Evolution of Inspection: From Manual Surveys to Foundation Models

The historical progression of railway inspection methodologies can be categorized into four distinct evolutionary phases as shown in Figure 1 characterized by increasing levels of automation, data abstraction, and semantic reasoning capabilities.

3.1 Phase 1: Manual and Heuristic Surveys

Historically, inspection relied entirely on human expertise. Track-walkers would visually assess the physical condition of the track through a process inherently subjective, labor-intensive, and fraught with safety risks (Thaduri et al., 2015). While expert human judgment allows for sophisticated reasoning (e.g., inferring that a leaning tree poses a specific risk during a storm due to soil saturation), the lack of scalability and digital traceability renders this approach insufficient for modern high-speed networks. The data generated often handwritten logs or basic digital forms lacked the granularity required for predictive modeling.

3.2 Phase 2: Traditional Computer Vision (CV)

The early 2000s saw the introduction of automated optical inspection systems utilizing traditional Computer Vision techniques. These systems relied on hand-crafted features such as Gabor filters for texture analysis, Canny edge detection for crack identification or Hough transforms for detecting the parallel geometry of rails (Olivier et al., 2025).

While effective for measuring rigid geometric parameters (e.g., track gauge or rail profile wear), these rule-based algorithms proved brittle in unstructured environments. As noted by (Tang et al., 2022), traditional CV systems struggle with the high variability of outdoor railway environments. CV algorithms fails when it subjected to variable lighting conditions, shadows, surface contaminants (oil, rust), or weather-induced visual noise.

3.3 Phase 3: Deep Learning and the Closed-World Ceiling

The advent of Deep Learning marked a significant leap in performance. Supervised Convolutional Neural Networks (CNNs) specifically architectures like YOLO, ResNet, and Mask R-CNN provided robust feature extraction that enable the reliable detection of specific components like clips, bolts, switches, ties, insulators etc even in noisy environments (Oh et al., 2022). Furthermore, specialized deep neural networks have been successfully developed for complex geometric tasks such as continuous track point extraction and instance tracing for rail path proposals (Ghorbanalivakili et al., 2023; Ghorbanalivakili and Sohn, 2025)

However, this era use the closed-world assumption as a fundamental architectural ceiling. A supervised CNN operates on a fixed taxonomy defined during training. For instance, a model trained to detect concrete sleepers will fail catastrophically if deployed on a line with wooden sleepers or slab track unless it is explicitly retrained with new annotated data. The cost of manually annotating pixel-perfect masks for every possible asset variant affects the scalability of AI solutions.

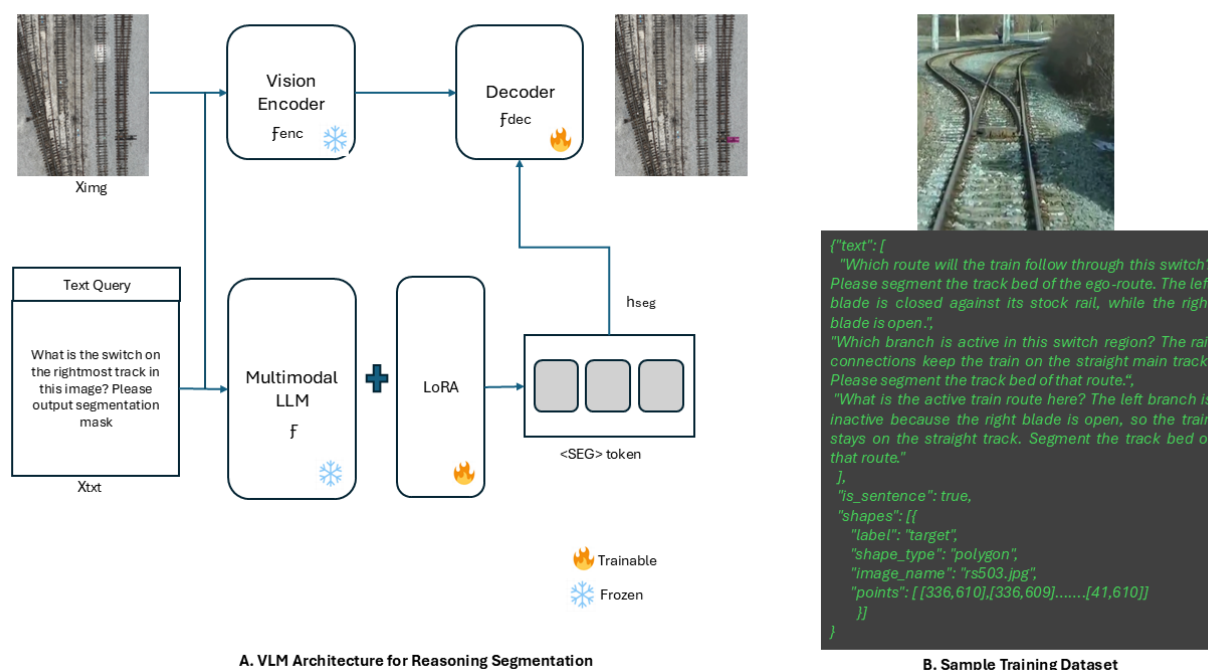


Figure 2. **VLM Architecture and Fine-Tuning Data for Reasoning Segmentation.** (A) Reproduced from the LISA framework (Lai et al., 2024). This diagram illustrates a theoretical application of MLLMs to railway diagnostics. The model processes a track image (x_{img}) and a spatially-aware text query (x_{txt}) using frozen backbones and trainable LoRA adapters to generate a hidden segmentation token (h_{seg}), which a decoder translates into a precise pixel-level mask. (B) A sample JSON dataset structure required to fine-tune such models for the railway domain mapping reasoning (e.g., determining the active train route through a switch) directly to polygon segmentation coordinates (Vakili, 2024).

4. Fundamentals of Vision-Language Models

The paradigm shift from closed-world Convolutional Neural Networks (CNNs) to open-world Vision-Language Models (VLMs) fundamentally alters visual representation learning. Rather than mapping high-dimensional image tensors to discrete pre-defined classes, VLMs map visual inputs and natural language into a shared semantic embedding space. Following the taxonomy of Lin (2025), the VLM landscape comprises three primary architectural paradigms:

- **Dual-Encoder Architectures:** Pioneered by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), this approach utilizes independent Vision Transformers (ViT) and text Transformers. Trained via symmetric contrastive loss (e.g., InfoNCE) on massive web-crawled datasets, they excel at zero-shot classification but lack generative capabilities and struggle with dense pixel-level prediction.
- **Encoder-Decoder with Cross-Attention:** Models like BLIP-2 (Li et al., 2023a) act as a bridge by introducing an alignment module (e.g., Q-Former) to extract visual tokens and fuse them with textual inputs via cross-attention, enabling complex visual question answering (VQA).
- **Decoder-Only Multimodal Large Language Models (MLLMs):** The current state-of-the-art (e.g., LLaVA) leverages pre-trained Large Language Models (LLMs) as a cognitive backbone. Visual tokens are projected directly into the LLM's embedding space via a Multi-Layer Perceptron (MLP) (Liu et al., 2023), offering autoregressive, deep contextual reasoning (Li et al., 2025).

To bridge global image-text alignment and dense localization, phrase grounding architectures like GLIP (Li et al., 2022) and Grounding DINO (Liu, 2024) uses Transformer-based detectors with grounded pre-training for open-set detection. For pixel-perfect masking, the Segment Anything Model (SAM) (Kirillov et al., 2023) and X-Decoder (Zou et al., 2023) enable zero-shot segmentation driven by arbitrary prompts that bypasses traditional segmentation bottlenecks.

Deploying these massive models on edge devices (e.g., inspection drones) remains challenging due to compute constraints. Token sparsification methodologies like FastVLM (Vasu et al., 2025) and SparseVLM (Zhang et al., 2024b) address this by dynamically pruning redundant visual tokens (e.g., uniform ballast or sky) to reduce inference latency by up to 40% without significant accuracy loss.

5. The Data Engine: Datasets and Benchmarking

Unlike traditional computer vision models constrained by human-annotated closed-set bounding boxes, VLMs utilize weakly supervised image-text pairs to capture unbounded open-vocabulary representations.

The VLM training pipeline is bifurcated. Phase one, massive-scale pre-training aligns vision and text encoders using billions of noisy image-text pairs from datasets like LAION-5B (Schuhmann et al., 2022). Phase two, visual instruction tuning utilizes pure-text LLMs to synthesize complex multi-turn JSON dialogues (e.g., ShareGPT4V (Chen et al., 2024)) to teach cognitive reasoning.

Evaluating MLLMs requires moving beyond standard mAP or Intersection over Union (IoU) metrics. Evaluation now incorporates generative metrics (ROUGE, BLEU, BERTScore) and comprehensive diagnostic suites like MMMU (Yue et al., 2024) and MMBench (Liu et al., 2024) to test spatial relationship reasoning and logical deduction. The critical vulnerability for safety-critical domains is *hallucination* confidently describing absent objects. Benchmarks like POPE (Li et al., 2023b) and HallusionBench (Guan et al., 2024) quantify this failure mode making rigorous evaluation imperative before deployment.

6. Emerging Railway Applications of VLMs

While the integration of VLMs in the railway industry is nascent, recent literature demonstrates profound applications in cognitive perception and automated safety management.

6.1 Cognitive Intrusion Perception

Traditional railway intrusion detection relies on geometric heuristics or pixel-level segmentation which cannot assess the latent intent of an object. Zhang et al. (2026) recently introduced **CogRail**, the first comprehensive benchmark designed for cognitive intrusion perception in intelligent railway systems. Evaluating state-of-the-art VLMs on this dataset reveals their capacity for complex spatio-temporal reasoning. Rather than merely detecting a pedestrian, fine-tuned VLMs can assess spatial positioning (e.g., “on the ballast” vs. “outside the track”), motion state, and dynamically assign threat levels (Safe, Potential Threat, Serious Threat).

6.2 Intelligent Safety Guidance Agents

Moving beyond perception, VLMs are being actively utilized for role-specific operational decision-making. Kang et al. (2026) proposed **RAISA**, a vision-based intelligent safety guidance agent. RAISA integrates a vision-language model (BLIP-2) with a large language model (Mistral-7B) utilizing a Modular Prompt-Orchestrated Reasoning framework. By mimicking “Chain-of-Thought” (CoT) reasoning, the system ingests front-view railway imagery, abstracts the situational context, and generates highly specific semantically grounded guidance. Crucially, the system can output differentiated commands providing immediate braking instructions for train operators while simultaneously generating coordination protocols for traffic dispatchers. This demonstrates the viability of VLMs as holistic human-in-the-loop Digital Twin collaborators.

7. Mapping General VLM Capabilities to Future Rail Applications

While the native railway applications discussed above are promising the broader computer vision and remote sensing literature contains advanced general-purpose VLMs that have yet to be fully utilized in the rail sector. This section maps these state-of-the-art domain-agnostic architectures to outstanding railway asset management challenges.

7.1 Open-Vocabulary and Open-Ended Defect Detection

Unlike traditional closed-set detectors (e.g., YOLO), generalist frameworks like Grounding DINO (Liu, 2024) enable open-set detection via text queries. Going further, VL-SAM (Lin et al., 2024) achieves training-free open-ended segmentation by

prompting the Segment Anything Model (SAM) using VLM attention maps. In a railway context, this allows for the detection of amorphous and variable hazards. For example, Weng et al. (2025) demonstrated that foundation models can effectively identify flooded regions or landslide debris in remote sensing imagery via zero-shot transfer. This capability allows Infrastructure Managers to rapidly deploy detectors for seasonal hazards (e.g., “snow drifts on track” or “fallen autumn leaves”) simply by modifying the text prompt which effectively bypassing the data annotation bottleneck (Lin, 2025).

7.2 Reasoning Segmentation for Structural Analysis

Railway engineering requires precise geometric understanding. The model outputs a segmentation mask based on a complex linguistic query rather than a predefined class index. Reasoning Segmentation frameworks Figure 2 such as LISA (Lai et al., 2024) and LLM-Seg (Wang and Ke, 2024) resolve complex spatial queries. A rail inspector could hypothetically prompt the model to “*Segment the left rail showing surface rust*” forcing the AI to semantically identify and spatially intersect both concepts. Sun et al. (2024) demonstrated this can be achieved training-free via LLM supervision offering a scalable solution for fine-grained component masking without pixel-level training.

7.3 Visual Question Answering (VQA) for Asset Diagnostics

Current inspection workflows often involve manually reviewing gigabytes of video footage. VLMs enable a shift towards Visual Question Answering (VQA) where the inspection system functions as a conversational agent. By feeding inspection imagery into a Multimodal Large Language Model (MLLM), operators can interrogate the data directly. For example, an engineer could ask a model “*Is the pantograph horn showing signs of abnormal wear*”. The VLM jointly reasons over the visual input and the textual query to extract task-relevant visual cues and return an answer tailored to the query.

7.4 Temporal Reasoning and Change Detection

Railway assets exist in a dynamic environment. Therefore, understanding change over time is as critical as assessing current condition. Traditional CV models struggle to align temporal sequences without complex registration. However, recent Earth Observation VLMs such as TEOChat (Irvin et al., 2024) introduce the ability to reason across temporal image sequences. If applied to rail corridors, a temporal VLM could ingest months of drone imagery to evaluate progressive ballast fouling or vegetation encroachment. Utilizing in-context value learning (Ma et al., 2024), these architectures could transition monitoring from reactive defect logging to true Predictive Maintenance (PdM) (Bianchi et al., 2025).

8. Outstanding Challenges and Future Directions

While Vision-Language Models offer a theoretical solution to the closed-world bottleneck, their practical deployment in railway asset management faces significant engineering and safety hurdles. This section outlines the critical barriers that future research must address to bridge the gap between academic benchmarks and operational reality.

- **Computational Overhead and Edge Deployment** The computational cost of models like LLaVA prohibits edge deployment without advancements in visual token sparsification (Vasu et al., 2025). Developing rail-specific pruning algorithms that preserve fine-grained defect details while discarding irrelevant background context is a critical open research question.
- **Reliability and Hallucination in Safety-Critical Systems** Railways are safety-critical systems. A false negative (missing a broken rail) can be catastrophic, while excessive false positives disrupt operations. VLMs, being generative by nature are susceptible to hallucination. The VLM model can confidently describes objects or defects that are not present in the image. Unlike traditional CNNs which output a calibrated confidence score, the textual output of a VLM is difficult to validate probabilistically. (Weng et al., 2025) highlights the need for Explanation-Driven Model Reliability, where models not only detect a defect but provide visual evidence (e.g., pointing to specific pixels) and reasoning traces to justify their output. Furthermore, incorporating Generative Error Correction mechanisms where a secondary verifier model critiques the initial output could significantly enhance robustness in safety-critical workflows (Ghosh et al., 2024).
- **The Domain Gap and Data Scarcity** Although VLMs are trained on billions of image-text pairs, the representation of specific railway engineering concepts (e.g., insulated rail joints, thermite welds, or pantograph horns) in general web data is sparse. This creates a domain gap where models may misclassify specialized assets. While zero-shot transfer is powerful, it has limits. A coordinated effort by the railway research community to curate open-source multi-modal datasets (linking imagery, LiDAR, and expert maintenance logs) is essential to fine-tune the next generation of Rail-VLMs (Tang et al., 2022).
- **Bridging the 2D-3D Divide** Finally, railway asset management is inherently spatial. Decisions are made based on precise 3D geometry (e.g., clearance gauging), yet most VLMs operate natively in 2D pixel space. Integrating 2D semantic understanding with 3D spatial precision remains a frontier challenge. Future work must focus on 3D-Aware Vision-Language Models that can directly ingest point clouds or align 2D semantic masks with 3D Digital Twins (Cheng et al., 2025).

9. Conclusion

The railway industry stands at a critical juncture where the rigidity of traditional Deep Learning limits the management of increasingly complex networks. The closed-world assumption creates blind spots for rare anomalies and incurs unsustainable retraining costs. Vision-Language Models emerge as a transformative solution to these limitations by enabling zero-shot anomaly detection and automated semantic reporting. This review synthesizes these capabilities into a cohesive vision for next-generation railway asset management that prioritizes on-demand natural language asset identification. We conclude that VLMs serve as a vital catalyst for proactive Predictive Maintenance by bridging the gap between raw sensor data and true semantic understanding.

References

- Abdelhamed, A., Afifi, M., Go, A., 2024. What do you see? Enhancing Zero-Shot Image Classification with Multimodal Large Language Models. *arXiv preprint arXiv:2405.15668*.
- Bianchi, G., Fanelli, C., Freddi, F., Giuliani, F., La Placa, A., 2025. Systematic review railway infrastructure monitoring: From classic techniques to predictive maintenance. *Advances in Mechanical Engineering*, 17(1), 1–26.
- Canadian National Railway, 2022. Cn's autonomous track inspection program. Accessed: 2024-05-20.
- Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D., 2024. ShareGPT4V: Improving large multi-modal models with better captions. 370–387.
- Cheng, Y. et al., 2025. 3d aware region prompted vlm.
- De Donato, L., Dirnfeld, R., Somma, A., De Benedictis, A., Flammini, F., Marrone, S., Azari, M. S., Vittorini, V., 2023. Towards AI-assisted digital twins for smart railways: preliminary guideline and reference architecture. *Journal of Reliable Intelligent Environments*, 9, 303–317.
- Ghorbanalivakili, M., Kang, J., Sohn, G., Beach, D., Marin, V., 2023. TPE-Net: Track point extraction and association network for rail path proposal generation. *Proceedings of the IEEE 19th International Conference on Automation Science and Engineering (CASE)*.
- Ghorbanalivakili, M., Sohn, G., 2025. TRIT-Net: Triplet-based railway instance tracing network using attraction field representation. *Proceedings of the 22nd Conference on Robots and Vision (CRV)*.
- Ghosh, S., Kumar, S., Seth, A., Chiniya, P., Tyagi, U., Duraiswami, R., Manocha, D., 2024. LipGER: Visually-Conditioned Generative Error Correction for Robust Automatic Speech Recognition. *arXiv preprint arXiv:2406.04432*.
- Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y. et al., 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14375–14385.
- Hidirov, S., Guler, H., 2019. Reliability, availability and maintainability analyses for railway infrastructure management. *Structure and infrastructure engineering*, 15(9), 1221–1233.
- Irvin, J. A., Liu, E. R., Chen, J. C., Dormoy, I., Kim, J., Khanna, S., Zheng, Z., Ermon, S., 2024. TeoChat: A large vision-language assistant for temporal earth observation data. *arXiv preprint arXiv:2410.06234*.
- Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *International Conference on Machine Learning (ICML)*, PMLR, 4904–4916.
- Kang, K. H., Yang, K. O., Choi, J. W., 2026. RAISA: A Vision-Based Intelligent Safety Guidance Agent for Railway Operations via Modular Prompt-Orchestrated Reasoning. *IEEE Open Journal of Intelligent Transportation Systems*.

- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al., 2023. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.
- Kljaić, Z., Pavković, D., Cipek, M., Trstenjak, M., Mlinarić, T. J., Nikšić, M., 2023. An Overview of Current Challenges and Emerging Technologies to Facilitate Increased Energy Efficiency, Safety, and Sustainability of Railway Transport. *Future Internet*, 15(11), 347.
- Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J., 2024. Lisa: Reasoning segmentation via large language model. 9579–9589.
- Li, J., Li, D., Savarese, S., Hoi, S., 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning (ICML)*, PMLR, 19730–19742.
- Li, L. H., Zhang, P., Zhang, H., Yang, J., Chunyuan, L., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N. et al., 2022. Grounded language-image pre-training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10965–10975.
- Li, Y., Du, Y., Lin, K., Chen, H., Mishra, S., Zhai, X., 2023b. Evaluating object hallucination in large vision-language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 292–305.
- Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., Shi, G., 2025. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1587–1606.
- Lin, F., 2025. Vision Language Models: A Survey of 26K Papers (CVPR, ICLR, NeurIPS 2023–2025). *arXiv preprint arXiv:2510.09586*.
- Lin, Z., Wang, Y., Tang, Z., 2024. Training-free open-ended object detection and segmentation via attention as prompts. *Advances in Neural Information Processing Systems (NeurIPS)*, 69588–69606.
- Liu, H., Li, C., Wu, Q., Lee, Y. J., 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 34892–34916.
- Liu, Shilong, Z. Z. T. R. F. L. H. Z. J. Y. Q. J. e. a., 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. 38–55.
- Liu, Y., Duan, H., Zhang, Y., Li, J., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z. et al., 2024. Mmbench: Is your multi-modal model an all-around player? *European conference on computer vision*, 216–233.
- Lourenço, A., Osório, F., Risca, D., Marreiros, G., 2026. Axle sensor fusion for online continual wheel fault detection in wayside railway monitoring.
- Ma, Y. et al., 2024. Vision Language Models are In-Context Value Learners.
- Oh, K., Yoo, M., Jin, N., Ko, J., Seo, J., Joo, H., Ko, M., 2022. A Review of Deep Learning Applications for Railway Safety. *Applied Sciences*, 12(21), 10572.
- Olivier, B., Guo, F., Qian, Y., Connolly, D. P., 2025. A review of computer vision for railways. *IEEE Transactions on Intelligent Transportation Systems*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, PMLR, 8748–8763.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 25278–25294.
- Sun, W., Du, Y., Liu, G., Kompella, R., Snoek, C. G. M., 2024. Training-Free Semantic Segmentation via LLM-Supervision. *arXiv preprint arXiv:2404.00701*.
- Tang, R., De Donato, L., Besinović, N., Flammini, F., Goverde, R. M. P., Lin, Z., Liu, R., Tang, T., Vittorini, V., Wang, Z., 2022. A literature review of Artificial Intelligence applications in railway systems. *Transportation Research Part C: Emerging Technologies*, 140, 103679.
- Thaduri, A., Galar, D., Kumar, U., 2015. Railway assets: A potential domain for big data analytics. *Procedia Computer Science*, 53, 457–467.
- Thompson, E. A., Lu, P., Alimo, P. K., Atuobi, H. B., Akoto, E. T., Abbew, C. K., 2025. Revolutionizing railway systems: A systematic review of digital twin technologies. *High-speed Railway*.
- Vakili, M., 2024. Railway perception foundation model dataset. https://github.com/mvakili96/Railway_Perception_FoundationModel.
- Vale, C., Simões, M. L., 2022. Prediction of railway track condition for preventive maintenance by using a data-driven approach. *Infrastructures*, 7(3), 34.
- Vasu, P. et al., 2025. Fastvlm: Efficient vision-language modeling. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19769–19780.
- Wang, J., Ke, L., 2024. LLM-Seg: Bridging Image Segmentation and Large Language Model Reasoning. 1765–1774.
- Wang, R., Hong, T., 2025. Few-shot defect detection in industrial scenarios: A comprehensive review of challenges, advances, and frontier trends. *MATEC Web of Conferences*, 413, 04005.
- Weng, X., Pang, C., Xia, G.-S., 2025. Vision-Language Modeling Meets Remote Sensing: Models, Datasets and Perspectives. *IEEE Geoscience and Remote Sensing Magazine*.
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Ge, G., Meng, T., Wang, J., Xia, Y. et al., 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9556–9567.

Zhang, J., Huang, J., Jin, S., Lu, S., 2024a. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5625–5644.

Zhang, Y. et al., 2024b. Sparsevlm: Visual token sparsification for efficient vlm inference.

Zhang, Y., Zhang, Q., Xu, W., Wang, Y., Wu, Y., Li, X., Dai, X., Zhang, H., Cui, Z., Guo, B. et al., 2026. CogRail: Benchmarking VLMs in Cognitive Intrusion Perception for Intelligent Railway Transportation Systems. *arXiv preprint arXiv:2601.09613*.

Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L. et al., 2023. Generalized decoding for pixel, image, and language. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15116–15127.