AI-Driven Digital Twin for Visual Defect Inspection in Railway Prognostics and Health Management

by

Rahatara Ferdousi

A thesis
submitted to the University of Ottawa
in partial fulfillment of the
thesis requirement for the degree of
Doctoral in Philosophy Electrical and Computer Engineering

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those concerning consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

Prognostics and Health Management (PHM) ensures safety and reliability in railways by monitoring assets like tracks, bogies, and wheels. Visual defects, such as cracks and corrosion, are critical health indicators. Traditional PHM methods, relying on AI-driven Digital Twin (DT) frameworks, face challenges like data scarcity, integration complexity, and resource constraints. Large Language Models (LLMs) offer potential advantages with minimal data requirements, adaptability to unseen samples, and data generation capabilities. However, a well-defined approach to integrate LLM into DT ecosystems for railway PHM remains unexplored. Therefore, we introduce DefectTwin, a comprehensive LLM-based DT ecosystem designed to address the key challenges in PHM for railway defect inspection. Our approach overcomes data scarcity by proposing a customized synthetic data generation pipeline that enables the fine-tuning of a Multimodal and Multimodel (M^2) LLM component. The domain-adapted LLMs enhance the AI inference engine of DefectTwin, ensuring high accuracy and reliable performance for railway defect inspection applications. To further strengthen the ecosystem, we propose a pipeline that incorporates a core feature of the DT ecosystem: a Quality of Experience (QoE) feedback loop. This mechanism is implemented to enhance the performance of LLMs (e.g., the quality of generated outputs) based on user feedback. Moreover, the synthetic dataset generated by our pipeline reduces the resource-intensive processes typically associated with traditional PHM systems, such as extensive data processing and model training. We conducted set of experiments to evaluate our proposed methods. The accuracy of the generated output was evaluated using the Canadian Pacific Railway dataset and synthetic data from our proposed pipeline. Testing on 600 image-based cases achieved a precision of 0.92, outperforming GPT-4 at 0.68 and Gemini-Pro-Vision at 0.88, with an F1-score of 0.92. In zero-shot scenarios, DefectTwin achieved a precision of 0.60 compared to GPT-4 at 0.40

and Gemini-Pro-Vision at 0.48, with an F1-score of 0.62. For video data, zero-shot precision and F1-score reached 0.55. In text-to-text defect analysis, DefectTwin averaged 150 tokens per response with 1.5 seconds latency, surpassing GPT-4, which averaged 220 tokens and 2.7 seconds, and Gemini-Pro-Vision, which averaged 180 tokens and 2.3 seconds. We obtained comparatively decent answer relevance (0.79) and context relevance (0.97) in defect detection tasks, outperforming GPT-4 (0.43, 0.52) and Gemini-Pro-Vision (0.41, 0.51), highlighting its precision and contextual understanding in railway PHM. Usability tests of a prototype based on DefectTwin ecosystem, confirmed practical applicability, with a good SUS score at an acceptance range of 70%. DefectTwin stands as the first LLM-integrated DT, specifically designed for visual defect inspection in railway PHM. This research paves the way for broader integration of LLMs into DT ecosystems, offering researchers and practitioners advanced approaches for tackling PHM challenges in railway maintenance strategies.

Acknowledgements

I am deeply grateful to the Almighty for providing me with the strength and perseverance to bring this study to fruition. I extend my heartfelt gratitude to everyone who supported and encouraged me throughout this journey.

My sincerest thanks go to Professor Abdulmotaleb El Saddik for his exceptional supervision, always motivating me to pursue timely publication. His consistent guidance and insightful advice were invaluable in accomplishing this research.

I would also like to express my deep appreciation to Dr. Anwar Hossain, Dr. Chunsheng Yang, and Dr. Fedwa Laamarti for their continuous motivation and thoughtful suggestions, which greatly contributed to the successful completion of this thesis.

To my husband, Khan Tanvir Hossain, I extend my deepest gratitude for his resilience, understanding, and growing respect for my aspirations. His companionship has made the challenges of the last phase of my PhD journey bearable and meaningful.

Lastly, I dedicate this thesis to my first child, Eira Khan, who grew with me as I navigated this critical phase of completing my research. Carrying you while writing this thesis gave me extra strength and determination, and I look forward to sharing this achievement with you one day.

Table of Contents

List of Tables

List of Figures

Abbreviation

Table 1: List of Abbreviations

Abbreviation	**Full Form**
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CBM	Condition-Based Maintenance
CNN	Convolutional Neural Network
CPR	Canadian Pacific Railway
CPS	Cyber-Physical System
DLLMDS	Defect LLM Dataset
DT	Digital Twin
EBTW	Energy-Based Thresholding Wavelets
FEA	Finite Element Analysis
HSR	High-Speed Rail
IoT	Internet of Things
IVHM	Integrated Vehicle Health Management
LLM	Large Language Model
LSTM	Long Short-Term Memory
LSTM-RNN	Long Short-Term Memory Recurrent Neural Networks
M^2	Multimodal and Multimodel
ML	Machine Learning
PCA	Principal Component Analysis
PHM	Prognostics and Health Management
QoE	Quality of Experience
RNN	Recurrent Neural Network
RUL	Remaining Useful Life
SHM	Structural Health Monitoring
SM	System Message
SUS	Software Usability Score
SVM	Support Vector Machine
TTI	Text-to-Image
VPI	Virtual Prompt Injection

Chapter 1

Introduction

This chapter presents the background, research problem, application scenario, research objective and statement, thesis contribution, organization, and scholarly outputs.

1.1 Background

Prognostics and Health Management (PHM) in railways is an engineering approach focused on predicting the health and performance of railway assets, such as tracks, wheels, bogies, and signaling systems [22]. It enables proactive maintenance [50] strategies by monitoring infrastructure conditions [65], diagnosing faults, and forecasting potential failures [18]. The common focus of PHM systems include- enhancing safety, operational reliability, and efficiency in railway transportation systems [23]. The key application area includes predictive maintenance, condition-based monitoring, and lifecycle management of assets [57]. Predictive maintenance leverages advanced analytics to forecast failures before they occur [1], while condition-based monitoring provides real-time insights into asset performance and deterioration [21]. Both rely heavily on accurate detection of infrastructure issues, such as cracks, corrosion, and wear, which are critical for preventing catastrophic failures [7]. These applications aim to achieve high predictive accuracy, reliable decision making, and reduced operational disruptions to extend lifespan of assets for overall safety [37].

Digital Twin (DT) technology has emerged as a transformative tool in PHM, particularly in railway systems [90]. A DT is a virtual representation of physical assets, providing real-time data exchange and actionable insights to support predictive and condition-based maintenance [37]. By enabling real-time fault detection and predictive insights, DTs facilitate more efficient maintenance planning and improved system reliability [12]. Furthermore, DTs encompass three critical states—observational (insight), predictive (foresight), and actuation (oversight), that collectively address the needs of modern railway operations [33].

The integration of Artificial Intelligence (AI) into DT frameworks has further enhanced their potential [28] into PHM systems. AI-powered DTs can analyze vast amounts of data, such as images, text, and sensor readings, to identify and predict defects more effectively [38] [52]. However, traditional AI approaches, which often rely on machine learning algorithms [56], face limitations in handling the complexity and diversity of railway systems [73] [20]. They struggle with incomplete datasets, lack scalability, and fail to generalize to unseen defect types [4] [34].

Recent advancements in LLMs [84], such as ChatGPT [55] [62], present a promising solution to these challenges. LLMs excel in processing multimodal data [80], enabling them to handle the diverse inputs required for visual defect inspection. Their inherent generalizability, adaptability, and ability to work with limited data make them well-suited for addressing the shortcomings of traditional AI methods [47] [19]. By integrating LLMs into DT ecosystems [26], PHM in railways can benefit from improved defect detection accuracy, better generalization to unseen defect types, and more efficient maintenance strategies .

More details of AI-integrated DT and its evolution for visual defect inspection in railway PHM are elaborated in Chapter 2. In the next section, we discuss the research problems identified in this thesis.

1.2 Research Problem

The integration of AI with Digital Twins promises comprehensive monitoring and predictive maintenance for PHM applications in railway [1]. However, these systems often struggle to provide consistent accuracy and operational adaptability due to the complexity of railway environments and the dynamic nature of railway operations. As noted in [24], AI-enabled DT often encounters issues like data inconsistency and underutilization, which hamper predictive maintenance efforts. Effective PHM requires seamless integration of heterogeneous data sources including real-time sensor data, operational logs, and maintenance records [45]. Traditional AI models, such as those relying on standard machine learning techniques such as Support Vector Machine (SVM), Recurrent Neural Network (RNN) [74], or pretrained advanced deep learning models like You Only Look Once (YOLO) [44] often fail to generalize to new conditions or adapt to unseen defects without extensive retraining. Because, in railway systems, collecting such data is challenging due to the infrequency of failures and unique operational conditions [71]. Due to lack of data samples these ML models lead to inefficiencies in detecting and responding to emerging issues [86] [91].

The integration of LLMs into DT frameworks offers a scope to overcome the limitations inherent in traditional AI models, particularly in the context of visual inspections. LLMs, such as ChatGPT [79], LLaMA [58], and Gemini [80], have demonstrated the capability to learn patterns from minimal examples. This characteristics of LLM reduces the reliance on extensive labeled datasets, which are often scarce in railway PHM applications. Recent studies have explored the application of LLMs in PHM, highlighting their potential to enhance predictive maintenance strategies and improve the overall efficiency of railway systems [84] [62].

Motivated by the need to address data scarcity, enhance defect detection accuracy, and manage the complexity of multimodal data integration in railway PHM, this thesis investigates the following research questions:

1. What are the key gaps, challenge, and requirements in existing railway PHM research? (Addressed in Chapter 2)
2. How can data scarcity be effectively addressed in the development of AI-driven DT ecosystems for railway PHM applications? (Addressed in in Chapter 3)
3. How can data from diverse sources, including images, sensor data, and operational records, be efficiently integrated into a unified DT ecosystem for accurate and comprehensive asset health assesment? (Addressed in Chapter 3)
4. How can accuracy be improved in railway PHM applications, especially for predicting sudden and unpredictable changes in asset condition? (Addressed in Chapter 3)
5. How can AI-driven DT systems impact visual defect inspection and maintenance in railway PHM? (Analyzed in Chapter 4)
6. How can resource constraints, particularly the high costs associated with developing and deploying DT-based PHM systems, be effectively addressed? (Discussed in Chapter 4)

1.3 Application Scenario

In this section, we discuss some application scenarios to understand the need for a DT ecosytem for Railway PHM from an application point of view.

Imagine a railway operator responsible for maintaining the safety and efficiency of train operations across an extensive network. Traditional methods of defect detection in railway components, such as manual inspections, can be labor-intensive, costly, and prone to human error. These methods often fail to detect early-stage defects like minute cracks or subtle signs of wear, which can escalate into major issues if not addressed promptly.

In this context, the railway operator needs the DT system, which leverages the power of LLMs and multimodal AI to enhance defect detection and maintenance.The system continuously analyzes data from various sources, including high-resolution video streams from surveillance cameras and real-time sensor data installed along the tracks. This enables the early identification of potential defects, such as cracks, corrosion, or structural deformations, which are then highlighted in detailed reports and visual representations.

The design of a DT ecosystem for railway defect inspection has the following need:

- The DT system needs access to a comprehensive and high-quality dataset that captures various aspects of railway components and their potential defects.
- The system should be able to identify potential defects in their early stages, such as cracks, corrosion, or structural deformations.
- The DT system needs to be integrated with the existing railway infrastructure, including surveillance cameras, sensors, and data management systems.

- The system should have a user-friendly interface that allows railway operators to easily access information, monitor the health of railway components, and make informed decisions regarding maintenance.

In Chapter 2, we identify the key gaps and associated challenges to address these gaps in existing literature and to better understand the requirements that align with this need.

1.4 Research Objective and Statement

The primary objective of this thesis is to design and develop a DT ecosystem that leverages LLMs and multimodal AI to enhance the visual defect inspection in railway PHM system. Our research aims to achieve the following objectives:

1. To investigate the feasibility and effectiveness of using LLMs to address data scarcity in railway PHM for visual defect inspection.

2. To address the data integration challenges in developing AI-driven DT systems for railway PHM.

3. To develop and evaluate approaches for enhancing the accuracy of AI models used in railway PHM for reliable detection of defects in railway assets.

4. To explore and evaluate strategies for addressing resource constraints in the development and deployment of AI-driven DT systems for railway PHM.

5. To validate the performance and usability of the proposed LLM-integrated DT system for assessing its impact on maintenance efficiency.

Based on this objective we state our key research statement in this thesis as *To what extent can long-standing challenges in traditional PHM solutions for railway defect inspection be addressed?*

1.5 Thesis Contribution

This research focuses on designing and evaluating the proposed methodologies for visual defect inspection in railway systems using LLMs and multimodal AI. The contributions of this thesis include:

1. Designing and developing a synthetic data generation pipeline to address data scarcity in railway defect inspection, a key challenge in PHM systems that contributes to issues such as poor accuracy and resource constraints.

2. Implementing a M^2 LLM component for AI inference to overcome data integration challenges and improve accuracy in detecting previously unseen defects.

3. Creating a customized visual-instruct dataset focused on railway defects, enabling LLM fine-tuning to generate relevant outputs while eliminating the need for manual, costly, and time-consuming data collection and preprocessing.

4. Designing and developing an LLM-based instant user feedback mechanism that establishes a Quality of Experience (QoE) Feedback Loop, allowing the system to adapt to user needs and continuously enhance performance.

5. Introducing DefectTwin, the first LLM-based Digital Twin ecosystem for visual defect inspection in railway PHM.

1.6 Thesis Organization

The thesis is organized as follows:

- In Chapter 2, we present an overview of railway PHM. This chapter outlines the focus, application, technologies in the literature. We identify the key gaps , challenges to address those gaps and finally the requirements for designing a solution to address those gaps.

- In Chapter 3, we describe the proposed DefectTwin ecosystem as a solution the addressed gaps. This chapter a detailed explanation of the proposed methods, algorithms, and competitive advantages of our approach.

- In Chapter 4, we detail our evaluation process, including various experiments. This chapter begins with different use cases for the proposed ecosystem, followed by a comprehensive evaluation of the model's performance, and concludes by comparing our results with existing methods.

- In Chapter 5, we conclude with a discussion on the scope of future work. This chapter addresses the issues tackled, the limitations of the study, and potential improvements for future research.

1.7 Scholarly Output

The scholarly output published out of this thesis is given as followings.

1. Ferdousi, R., Yang, C., Hossain, M. A., Laamarti, F., Hossain, M. S., Saddik, A. E. (2024). Generative Model-Driven Synthetic Training Image Generation: An Ap-

proach to Cognition in Rail Defect Detection. *Cognitive Computation*.

2. Ghaboura, S., Ferdousi, R., Laamarti, F., Yang, C., El Saddik, A. E. (2023). Digital Twin for Railway: A Comprehensive Survey. *IEEE Access*.

3. Ferdousi, R., Laamarti, F., Yang, C., El Saddik, A. E. (2024). A Reusable AI-Enabled Defect Detection System for Railway Using Ensembled CNN. *Applied Intelligence*.

4. Ferdousi, R., Laamarti, F., Yang, C., El Saddik, A. E. (2022). RailTwin: A Digital Twin Framework For Railway. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*.

5. Yang, C., Ferdousi, R., El Saddik, A., Li, Y., Liu, Z., Liao, M. (2022). Lifetime Learning-enabled Modelling Framework for Digital Twin. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*.

6. Ferdousi, R., Hossain, M. A., El Saddik, A. E. (2024). TextureMeDefect: LLM-Based Synthetic Defect on Mobile Device. In *IEEE 43rd Conference on Consumer Electronics*.

7. Ferdousi, R., Hossain, M. A., Yang, C., El Saddik, A. E. (2024). DefectTwin: When LLM Meets Digital Twin for Defect Inspection. Under Review.

Chapter 2

Literature Review

This chapter first defines the PHM in railway, followed by synthesis of existing research from 2018-2024. After that, we summarize the trend of application, technology and gap based on the existing work. Finally, we outline the requirements to address the key gaps in current studies.

2.1 Definition of PHM in Railways

PHM is an engineering discipline focused on predicting the time at which a system or component will no longer perform its intended function. This prediction, known as the Remaining Useful Life (RUL), is crucial for decision-making in maintenance and operational planning [59].

In industrial applications, PHM involves monitoring equipment conditions, diagnosing faults [98], and forecasting failures [40] to enable condition-based maintenance strategies.

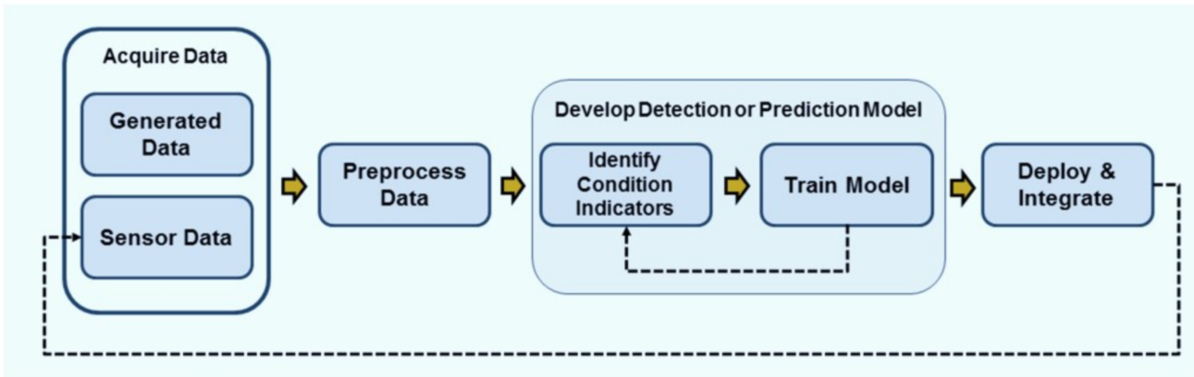


Figure 2.1: Basic Workflow of PHM Process [64]

This approach enhances reliability, reduces operational costs [11], and minimizes downtime by addressing issues before they lead to system failures [72].

The basic process of a PHM, as depicted in the Fig, 2.1, begins with the acquisition of data. This includes both generated data, which can be synthetic or simulated, and sensor data collected in real-time from railway infrastructure. Once the data is acquired, it undergoes a preprocessing stage where it is cleaned, normalized, and prepared for analysis. The next step is to develop a detection or prediction model, which involves identifying condition indicators that reflect the health of the system and training the model using this data to detect or predict faults or failures. Finally, the trained model is deployed and integrated into the operational system, where it continuously monitors the railway infrastructure. The system analyzes incoming data in real-time and provides valuable insights or alerts regarding the asset's condition, allowing for proactive maintenance actions. This feedback loop ensures that the PHM system continuously refines its predictions and contributes to the efficient management of the railway's health [41].

According to [22], PHM in the railway industry refers to the systematic process of monitoring, diagnosing, and predicting the health and performance of railway assets. Within the railway industry, PHM is applied to monitor and managthe health of critical components such as train bogies [88], tracks [75], and signaling systems [49]. By utilizing advanced sensors and data analytics, PHM systems detect anomalies, diagnose issues, and predict potential failures, facilitating timely maintenance interventions [51]. This proactive maintenance strategy improves safety, ensures operational efficiency, and optimizes maintenance resources [70]. Based on the current information, we can define PHM for Railway as followings.

PHM for railway is the systematic approach to monitoring, diagnosing, and predicting the health and remaining useful life of railway assets using advanced sensing, data analytics, and predictive models to enable proactive maintenance, enhance reliability, ensure safety, and optimize resource utilization.

2.2 State-of-the-art of PHM in Railways

In this section, we explore and summarize the purpose, methodology, technology, applications and challenges from existing work since 2018-2024, to understand the current trend of PHM research for railways.

Trend of railway PHM research in 2018

It can be observed from Table 2.1, that the trend in railway PHM research in 2018, is decisively toward the DT or CPS technologies to enable predictive maintenance.

Table 2.1: 2018: Summary of Research on Prognostics and Health Management for Railways

Reference	Focus	Method	Paradigm	Application Area	Limitations
[60]	Industrial AI for predictive maintenance in HSR	Signal processing, machine learning	CPS, and AI	Predictive maintenance, real-time monitoring	High maintenance costs, asset state snapshot collection
[61]	Digital twins for electric drive trains	Model-based simulation	DT	Fault prediction, operational optimization	Handling unmeasurable data
[82]	Determining the Equivalent Conicity for Railway Wheelset Maintenance with Deep Ensembles	Data Driven Neural network	AI	Wheelset maintenance scheduling	Instability prediction
[9]	A Case Study on the Health Assessment of the Point Machines	Data-Driven feature extraction using PCA	AI	Health assessment of point machines	Fault detection and diagnostics

For example, Liu et al. (2018) [60], proposed a CPS-enabled PHM framework, which utilizes advanced signal processing and machine learning to perform real-time performance monitoring and fault prediction. As discussed by Ludwig et al. (2018) [61], DT provides a model-based simulation approach that enhances operational support and fault prediction. Trilla and Cabré (2018) [82], developed a data-driven approach to estimate the equivalent conicity of railway wheelsets, which is indicative of dynamic instability. They employed deep ensemble neural networks to process related physical measurements, enabling the prediction of equivalent conicity values. Ardakani et al. (2012) [9], focused on the prognostics and health management of electro-mechanical point machines. They applied feature extraction techniques to time-stamped current and voltage data from point machines, followed by PCA to assess machine health. This approach enabled the detection of machine degradation and demonstrated the applicability of PHM techniques for fault diagnostics and prognostics in railway systems. Despite the benefits discussed in the current studies the integration of AI models into PHM system remained as a challenge due the hurdles of railway data collection. Because, to build a data-driven model or a real-time model for reliable prediction vast amount of data was needed.

Trend of railway PHM research in 2019

In 2019, the focus of research on PHM in the railway industry was heavily directed toward utilizing AI methods as summarized in Table 2.2.

AI-driven methods in 2019, focused on practical, targeted applications for specific railway components. For instance, an AI-based condition monitoring system for rail infrastructure employed a machine learning pipeline to detect anomalies in sensor data, which helped identify early-stage rail defects [6]. Similarly, an experience-based health evaluation method was applied specifically to train wheels, using operational, and maintenance data to assess their condition and categorize them into health status levels [70].

Table 2.2: 2019: Summary of Research on Prognostics and Health Management for Railways

Reference	**Focus**	**Method**	**Pardigm**	**Application Area**	**Limitations**
[29]	Role of DT in vehicle health management (IVHM)	Analysis of DT applications and enabling technologies	DT	CBM, real-time health monitoring	Challenges in DT implementation for complex systems
[36]	Digital Shadow for optimized maintenance in rail freight	Data-driven operations optimization, prescriptive maintenance	DT	Operational efficiency, predictive maintenance for rail vehicles	Limited implementations of DTs, high costs
[70]	Experience-based health evaluation for train wheels	Health status evaluation, decision-making for wheel health	AI	CBM for train wheels	Limited data and need for broader evaluation methods

Reference	Focus	Method	Paradigm	Application Area	Limitations
[6]	AI-based condition monitoring for rail infrastructure	Machine learning pipeline, anomaly detection	AI	Predictive maintenance for rail infrastructure	Data challenges in connecting sensor data with rail conditions

DT were explored as Digital Shadows [36], to enable real-time monitoring and predictive maintenance, particularly in complex systems like rail vehicles and infrastructure. In [29], the authors emphasized DTs ability to simulate and monitor the health of a system. However, the challenges of subsystem integration in DT remained as a challenge. Therefore, DT implementation was limited at conceptual level for various case studies of railway PHM.

Trend of Railway PHM Research in 2020

In 2020, research in PHM for the railway industry focused on applying traditional machine learning approaches to enhance predictive maintenance and operational efficiency. In Table 2.3, we summarize the studies.

Table 2.3: 2020: Summary of Research on Prognostics and Health Management for Railways

Reference	**Focus**	**Method**	**Paradigm**	**Application Area**	**Limitations**
[49]	Diagnosis of rail corrugation in high-speed railways	Time-frequency analysis and feature extraction	AI	Defect detection in rail infrastructure	Complexity in modeling dynamic interactions
[86]	Predictive and proactive maintenance for railway power equipment	LSTM-RNN for failure prediction	AI	Maintenance optimization for power equipment	Data scarcity and model accuracy
[65]	Handling missing data in railway asset management	Data imputation and machine learning techniques	AI	Condition-based maintenance	Data quality and completeness

Reference	Focus	Method	Paradigm	Application Area	Limitations
[66]	Monitoring health status of train infrastructure assets	Positioning systems and data analysis	DT	Asset health monitoring	Accuracy and reliability of positioning data
[18]	Fault diagnosis of railway switch systems	EBTW and neural networks	AI	Fault detection in switch systems	Feature extraction and model training
[81]	Maintenance of railway infrastructure using CPS	Framework for CPS integration	CPS	Predictive maintenance and real-time monitoring for railway infrastructure	Complexity in integrating CPS with existing systems

For instance, Li and Shi (2020) [49], proposed a method for diagnosing rail corrugation in high-speed railways by analyzing dynamic responses of the vehicle, utilizing time-frequency analysis and feature extraction techniques. Wang et al. (2020) [86], introduced a predictive and proactive maintenance approach for high-speed railway power equipment using Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN), aiming to predict potential failures and optimize maintenance schedules .

The challenge of missing data in railway asset management was studied by McMahon et al. (2020) [65]. The authors explored various approaches to handle incomplete condition monitoring data, emphasizing the importance of data quality in implementing effective condition-based maintenance strategies. Chen et al. (2020) [18] improved fault diagnosis of railway switch systems by employing Energy-Based Thresholding Wavelets (EBTW) and neural networks, enhancing the accuracy of fault detection and contributing to the reliability of railway operations.

Few work focus on utilizing AI and sensing newtwork, as well as CPS to advance railway PHM. For example, Moradi et al. [66] focused on positioning systems for monitoring the health status of train infrastructure assets. Thaduri et al. (2020) [81] explored the use of CPS for the maintenance of railway infrastructure. They proposed a framework for implementing CPS in railway infrastructure and discussed its potential benefits, challenges, and future prospects. Their study emphasized the integration of CPS with railway maintenance systems to enable real-time monitoring, decision-making, and predictive maintenance.

Trends in Railway PHM research in 2021

The existing work in 2021 is summarized in Table 2.4. There was extensive application of DT for various predictive maintance tasks for railway operation.

Table 2.4: 2021: Summary of Research on Prognostics and Health Management for Railways

Reference	**Focus**	**Method**	**Paradigm**	**Application area**	**Limitations**
[97]	Fault diagnosis of railway point machines	Digital twin-assisted AI algorithms	DT, AI	Fault detection and diagnosis	Data accuracy and model validation
[46]	Maintenance technologies for railway assets	Analysis of DT and AI integration	DT, AI	Predictive maintenance	Development of supportive business models
[38]	Intelligent cyber-physical systems development	Integration of DT and AI	DT, AI	Real-time monitoring and decision-making	System complexity and data integration
[14]	Autonomous health monitoring of rails	FEA-ANN based approach	FEA, ANN	Defect detection in rails	Model accuracy and computational requirements

For example, Zhang et al. (2021) [98] introduced a DT-assisted fault diagnosis system for railway point machines, where AI-driven algorithms utilized DT data to identify and diagnose faults in real-time. Similarly, Kumar and Galar (2021) [46] explored transformative maintenance technologies. The authors emphasized DT's role in predictive maintenance and proposing business solutions to integrate such systems into railway management workflows. The convergence of DT with AI was further expanded upon by Groshev et al. (2021) [38], who developed a framework for intelligent cyber-physical system for future. Their study highlighted the integration of real-time monitoring and decision-making capabilities, emphasizing scalability and adaptability for complex railway networks. Brown et al. (2021) [14] focused on health monitoring by combining Finite Element Analysis (FEA) with Artificial Neural Networks (ANN) to monitor rail integrity. The authors proposed a more localized and defect-specific diagnostic approach, paving the way for smarter, autonomous railway systems. However, implementing DT and AI systems in practice remains a challenge. Integrating heterogeneous datasets, ensuring system interoperability, and managing computational requirements are significant barriers. Moreover, the development of an ecosystem to support these advanced technologies is crucial to facilitate adoption.

Trends in Railway PHM research in 2022

In 2022, research in PHM for railway systems continued to advance, with a notable emphasis on integrating ML algorithms into DT frameworks [33] to offer intelligent prediction of the status of railway assets [92]. As summarized in Table 2.5, technologies like IoT sensing, LSTM and Hybrid ML models have been proposed for monitoring and predictive applications.

Table 2.5: 2022: Summary of Research on Prognostics and Health Management for Railways

Reference	Focus	Method	Paradigm	Application Area	Limitations
[96]	Health status assessment of station equipment	G-LSTM neural network	DT, AI	Predictive maintenance	Data complexity, system interoperability
[52]	Cyber-physical intelligent transport system	Real-time monitoring and control	AI, CPS	Transportation monitoring	System integration, scalability
[59]	Exploration of digital twin technology in rail transit health management	Analysis of digital twin technology's origin and application framework	DT	Enhancing complex equipment operation	Integrating DT into existing rail systems
[83]	AI and predictive analytics for maintenance	Hybrid and cognitive DTs	DT, AI	Predictive maintenance	Data analytics, implementation cost

Reference	Focus	Method	Paradigm	Application Area	Limitations
[15]	Diagnostic system for passenger trains	Wireless monitoring system	IoT	Condition-based maintenance	Sensor deployment, data reliability
[54]	Lifecycle management of complex equipment	Machine learning integration	DT, AI	Predictive maintenance	Lifecycle data management, predictive accuracy

For example, Li et al. (2022) [53], explored machine-learning-driven DTs for lifecycle management of complex equipment, highlighting the potential of DTs combined with AI to predict equipment failures and optimize maintenance schedule. The authors in [52], introduced a cyber-physical intelligent transport system based on DT technology aiming for real-time monitoring.

Yao et al. (2022) [96], developed a health status assessment and prediction scheme for railway station equipment using a Graph Long Short-Term Memory (G-LSTM) neural network, demonstrating high accuracy in predicting equipment health. Additionally, Unal et al. (2022) [83], discussed data-driven AI and predictive analytics for the maintenance of industrial machinery. The authors emphasized on the role of hybrid and cognitive DTs [35], in predictive maintenance strategies. Castejón et al. (2022) [15] focused on developing a diagnostic system for condition-based maintenance of passenger trains. Their work presented a wireless monitoring system designed to retrofit passenger vehicles with cost-effective and easy-to-install instrumentation. Although the existing work shift focus towards integrating sensors and advanced AI algorithms for PHM system, the integration of sensors, and smart devices to existing railway limits practical implementation of the proposed approaches.

Trends in Railway PHM Research in 2023

Similar to the year 2022, a strong emphasis on integrating ML into DT to create comprehensive railway monitoring systems is observed in 2023. As summarized in Table 2.6, the application focus has shifted from traditional diagnostic methods to predictive and prescriptive maintenance strategies.

For instance, Wu et al. (2023) [89], proposed a DT-based framework for fault diagnosis of high-speed train bogies, utilizing a multi-layer convolutional neural network to improve diagnostic accuracy. Shimuz et al. (2023) [72], proposed a similar DT-based fault diagnosis framework for high-speed train bogies, demonstrating the growing importance of DT for component-level diagnostics. Additionally, Sresakoolchai and Kaewunruen [76], developed a DT integrated with deep reinforcement learning to enhance maintenance efficiency by analyzing track geometry and component defects. Marx et al. (2023) [63], conceptualized a DT for railway bridges, exemplified by the New Filstal Bridges, to monitor structural health and predict maintenance needs.

Chen et al. [17], highlighted the role of ML within DTs for predictive maintenance, emphasizing the importance of data-driven approaches in forecasting equipment failures. Ahmad et al. [5], proposed a DT-based approach for predicting rail surface damage in heavy haul operations, leveraging AI models to anticipate wear and tear. Kaewunruen et al. [43], applied AI technology to prognose and diagnose complex crack characteristics in railway concrete sleepers, significantly enhancing maintenance planning and safety strategies.

Furthermore, new methodologies involving ChatGPT-like large-scale foundation models were explored. Li et al. (2023) [55] provided a systematic review and roadmaps for integrating such models into PHM, emphasizing their potential for fault prediction and health management. Similarly, Wang et al. (2023) [85] investigated enhancing language models like ChatGPT by integrating local knowledge bases to improve the relevance and accuracy of maintenance recommendations in PHM applications.

Table 2.6: 2023: Summary of Research on Prognostics and Health Management for Railways

Reference	**Focus**	**Method**	**Paradigm**	**Application Area**	**Limitations**
[89]	DT-based fault diagnosis of high-speed train bogies	Multi-layer convolutional neural network	Digital Twin, AI	Fault diagnosis in train bogies	Real-time data integration and modeling
[76]	Maintenance efficiency improvement using DT and reinforcement learning	Deep reinforcement learning integrated with DT	DT, AI	Maintenance of railway track and components	Track defect analysis and data variability

Reference	Focus	Method	Paradigm	Application Area	Limitations
[63]	Concept for DT of railway bridges for monitoring	Structural health monitoring framework with DTs	DT	Maintenance of railway bridges	Structural complexity and data reliability
[17]	ML-enhanced predictive maintenance with DTs	Integration of ML with DTs	DT, AI	Predictive maintenance	Data fusion and system scalability
[72]	Real-time PHM without run-to-failure data for railway assets	Predictive algorithms for real-time PHM	AI	Maintenance of railway assets	Data insufficiency and scalability challenges
[5]	DT for predicting rail surface damage in heavy haul operations	Integration of DT with predictive models	DT, AI	Rail surface maintenance	Accurate wear prediction

Reference	Focus	Method	Paradigm	Application Area	Limitations
[43]	AI-based diagnosis of cracks in railway concrete sleepers	Prognostic AI models for crack detection	AI	Maintenance of concrete sleepers	Complexity of crack patterns
[55]	Foundation models like ChatGPT for PHM	Roadmap and surveys of large-scale AI models	AI	Fault prediction and condition monitoring	Lack of consensus and systematic reviews
[85]	Enhancing ChatGPT-like models for PHM with local knowledge bases	Integration of LLMs with local knowledge bases	AI, Knowledge Engineering	Industrial PHM applications	Domain-specific expertise and scalability

The research in 2023 demonstrates the convergence of AI, ML, and DT technologies to advance railway PHM. A notable trend is the application of large-scale foundation models like ChatGPT for PHM tasks [55]. Because, LLMs can process and interpret vast amounts of maintenance data, facilitating accurate fault diagnosis and enabling predictive maintenance strategies. By analyzing historical data, these models can identify patterns indicative of potential failures, allowing for timely interventions. As discussed in [85], combining LLMs with domain-specific local knowledge bases (LKBs) enhances their applicability in PHM. This integration improves the accuracy and relevance of the model's outputs, making them more insightful for specific industrial applications.

Trends in Railway PHM Research in 2024

The year 2024 signifies a continued advancement in the integration of AI and DT technologies. Emphasis has been placed on the generation of synthetic data and the application of ML algorithms across various PHM applications. As summarized in Table 2.7, the research focus predominantly lies in structural health monitoring and diagnostic challenges, including the detection of visible defects and fault diagnosis. Consequently, computer vision models have seen a significant increase in utilization in the PHM research conducted in 2024.

Table 2.7: 2024: Summary of Research on Prognostics and Health Management for Railways

Reference	**Focus**	**Method**	**Paradigm**	**Application Area**	**Limitations**
[44]	Explores AI to identify and classify cracks in railway sleepers	YOLOv5OBB model	AI	Maintenance and safety optimization for railway sleepers	Limited by the quality and diversity of training data; potential challenges in real-time deployment

Reference	Focus	Method	Paradigm	Application Area	Limitations
[71]	Proposes hybrid ML and physics-based solutions for fault diagnostics and health management	Hybrid ML models	AI	Fault diagnosis and predictive maintenance	High computational requirements; integration complexity with existing systems
[10]	Develops a digital twin-based SHM for railway bridges	ML-based vibration analysis	DT, ML	Structural health monitoring of railway bridges	Dependence on accurate sensor data; challenges in modeling complex bridge dynamics
[87]	Presents a digital twin for lifecycle management of railway infrastructure	Assessment and maintenance framework	DT	Lifecycle management for railway systems	Scalability issues; integration with diverse infrastructure components

Reference	**Focus**	**Method**	**Paradigm**	**Application Area**	**Limitations**
[8]	Proposes predictive-cognitive maintenance using embedded ML and DTs	Embedded ML and cloud-based digital twins for edge monitoring	DT, AI	Predictive maintenance for rail tracks	Data privacy concerns; latency in cloud-based processing
[74]	Introduces a cloud-based framework for railway vehicle dynamics	Cloud computing for DT-enabled simulation	DT	Predictive maintenance and vehicle dynamics	Connectivity issues in remote areas; potential data security risks
[39]	Develops a DT framework for condition monitoring	Use of physics-based models for synthetic data generation	DT, AI	Maintenance planning and monitoring	Accuracy of synthetic data; computational intensity of physics-based models

Reference	Focus	Method	Paradigm	Application Area	Limitations
[25]	Utilizes AI for predictive maintenance in railways	Advanced ML models like CNNs, RNNs, and SVMs	AI	Railway network monitoring and optimization	Requirement for large labeled datasets; interpretability of complex models

Kaewunruen et al. (2024) [44] utilized AI for identifying and classifying cracks in railway sleepers using the YOLOv5OBB model. Their work advances maintenance and safety strategies for railway infrastructure by providing early warnings of potential issues. Shen et al. (2024) [71] proposed a hybrid approach combining ML and physics-based models to enhance fault diagnostics and predictive maintenance. Revolutionizing Railways (2024) [25] explored advanced ML techniques, including CNNs, RNNs, and SVMs, for predictive maintenance and network optimization.

Armijo et al. (2024) [10] developed a DT-based structural health monitoring (SHM) system for railway bridges, incorporating low-cost sensors and ML-based vibration analysis to detect anomalies. Wilke et al. (2024) [87] focused on lifecycle management of railway infrastructure by presenting a DT framework for real-time assessment and maintenance prioritization. Arakistain et al. (2024) [8] introduced predictive-cognitive maintenance techniques leveraging embedded ML and DTs for edge monitoring of rail tracks. Smith et al. (2024) [74] developed a cloud-based framework for simulating railway vehicle dynamics, enabling DT-supported predictive maintenance and system optimization. Gupta et al.

(2024) [39] proposed a DT framework for condition monitoring and maintenance planning, using physics-based models for synthetic data generation.

The utilization of generative AI in Prognostics and Health Management (PHM) has continued to progress in 2024. For example, the authors in [78] study the integration of LLMs into PHM. They propose three progressive paradigms namely, PHM-LM 1.0, PHM-LM 2.0, and PHM-LM 3.0 for incorporating LLMs into PHM, along with discussions on technical approaches and challenges. They suggest improvements in data interpretation, reasoning, decision-making, and autonomous management within PHM systems across industries such as aerospace, manufacturing, maritime, rail, and energy. However, integrating LLMs into DT systems requires careful consideration to understand the challenges of deploying such PHM systems for industrial application

2.3 Application Trend of PHM in Railways

The timeline in Fig. 2.2, illustrates the key PHM applications in the railway sector, as identified for each year from 2018 to 2024. Each year introduces unique applications or builds upon earlier trends.

From 2018 to 2020, the focus was primarily on foundational applications like predictive maintenance, real-time monitoring, and operational optimization. These years laid the groundwork for advanced PHM systems by addressing core maintenance needs, detecting faults, and optimizing operations. Data-driven methods like anomaly detection and condition-based maintenance became prominent in this period.

Between 2021 and 2022, the focus shifted toward integration and scalability, with emerging technologies like DT and hybrid ML models playing a pivotal role. Applications evolved

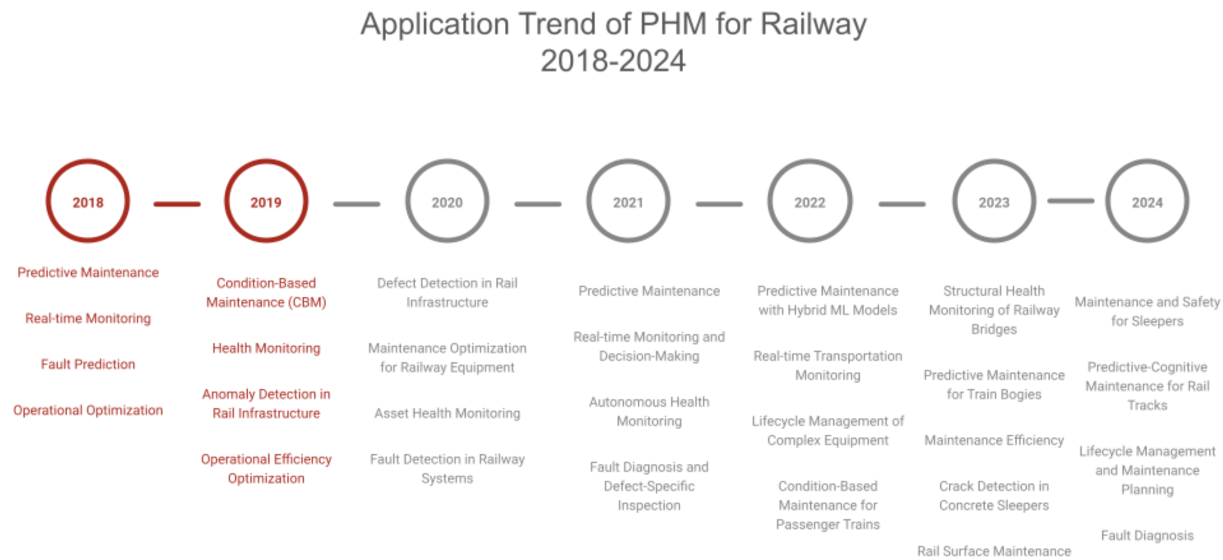


Figure 2.2: Application Trend of PHM in Railways (2018–2024)

to include lifecycle management of complex equipment, real-time decision-making, and infrastructure-level monitoring. By 2023 and 2024, PHM applications reached a phase of specialization and infrastructure emphasis, with notable advancements in structural health monitoring of railway bridges, predictive-cognitive maintenance for rail tracks, and fault diagnosis using AI-DT integration.

The trends illustrated in the graph demonstrate a clear transition from basic monitoring to advanced, AI-driven DT systems capable of and predictive diagnostics and condition based monitoring.

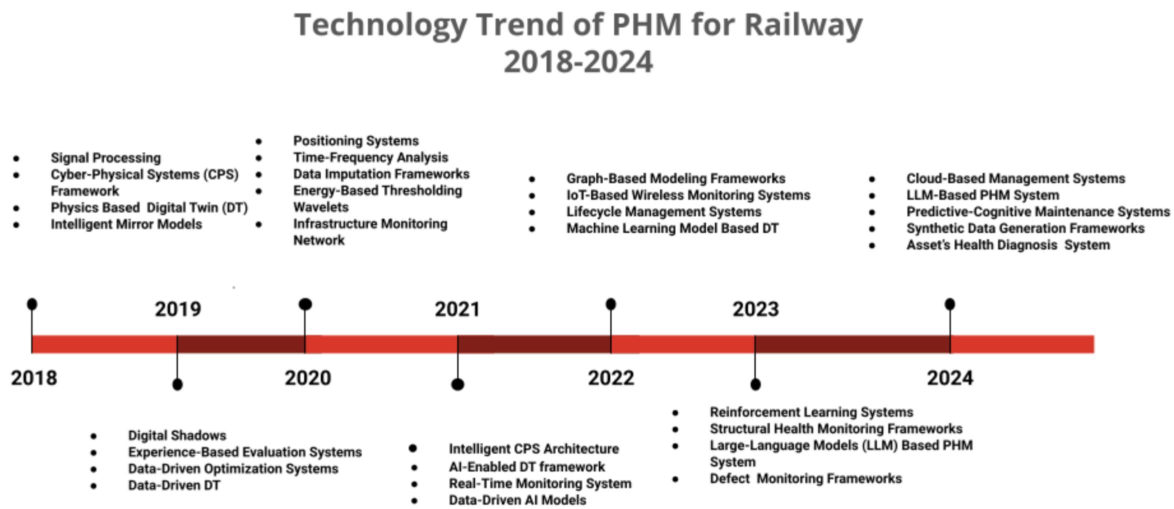


Figure 2.3: Technology Trend of PHM in Railways (2018–2024)

2.4 Technology Trend of PHM in Railways

The development of PHM technologies for railways from 2018 to 2024 in Fig. 2.3, showcases a dynamic shift in approaches, moving from physics-based DT to AI-driven DT as well as mathematical algorithm to LLM based PHM system.

.

The evolution of PHM technologies in the railway industry from 2018 to 2024 in Fig. 2.3, illustrates a significant shift towards data-driven, and AI-integrated solutions. Initially, in 2018, PHM systems in railways focused on basic methods, such as signal processing and CPS for fault detection and real-time monitoring. These early systems primarily used DTs as a physical model based conceptual framework to simulate and monitor railway components, though they were limited in scalability and adaptability for large railway networks.

In 2019, the development of technologies like digital shadows, data-driven optimization systems, and experience-based evaluation systems enhanced condition-based monitoring. These systems laid the foundation for more dynamic, data-driven approaches, but still lacked the flexibility to handle large-scale integration. As such, they were effective in localized monitoring but not yet ready for full railway system integration.

From 2020 onwards, there was a marked shift towards data fusion, advanced fault detection, and real-time monitoring systems. Key advancements in positioning systems and time-frequency analysis helped address data gaps, while energy-based thresholding and wavelet-based methods contributed to more accurate fault detection in dynamic systems. This transition reflected a growing demand for more comprehensive and accurate monitoring solutions.

By 2021, the integration of real-time data frameworks and AI-enhanced DTs allowed for improved decision-making capabilities. As seen in Fig. 2.4, from 2021 onwards, intelligent DTs began to take precedence over the traditional CPS-based DT.

The advancements in machine learning models and data-driven predictive analytics provided better fault detection and maintenance planning. However, the challenge remained to integrate modern technologies like IoT and DTs with existing, often outdated, infrastructure.

The trend continued in 2022, with the introduction of graph-based modeling frameworks, IoT-enabled wireless monitoring systems, and reinforcement learning models. These technologies enabled specialized solutions for critical areas such as structural health monitoring and crack detection. By 2023, LLMs emerged as a key tool for predictive main-

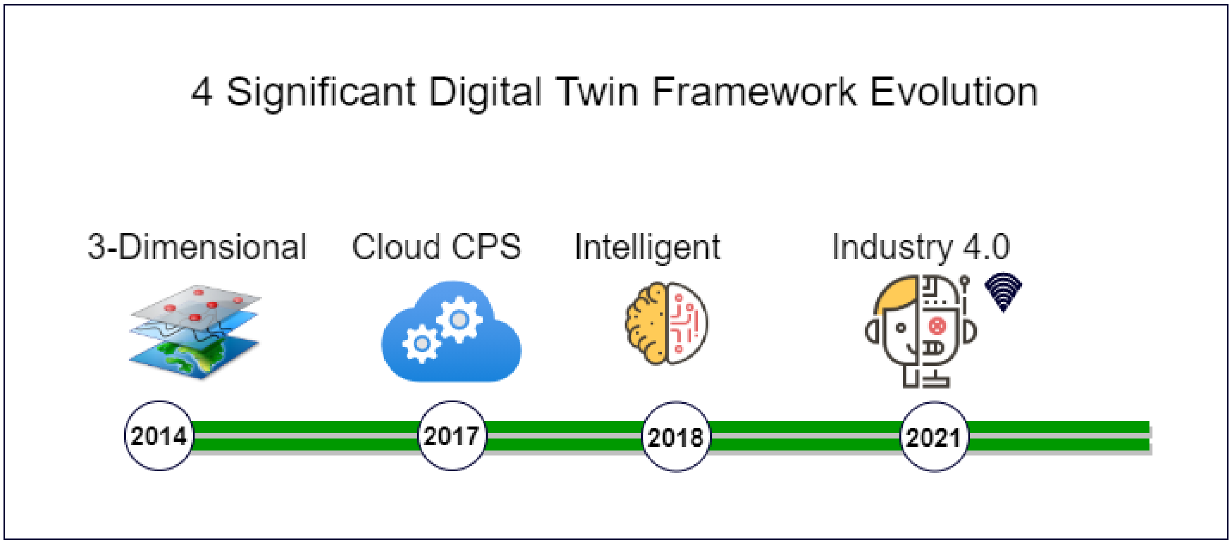


Figure 2.4: Evolution of DT Framework [32]

tenance and diagnostics, offering targeted monitoring for complex assets such as bridges, tracks, and switches. However, issues related to data quality, especially insufficient or low-quality data samples, continued to challenge the consistency of predictive maintenance applications.

Moving to 2023 and 2024, AI-driven DTs and predictive-cognitive systems became the forefront of PHM technologies, capable of providing real-time, scalable monitoring across vast railway networks. The integration of synthetic data generation frameworks which use physics-based models or AI to generate realistic data helped mitigate issues of data scarcity, enhancing model accuracy and robustness. Additionally, cloud-based management systems became integral to handling the massive data flows required for large-scale PHM applications, with LLMs offering powerful tools for dealing with a range of complex issues in railway PHM systems.

In summary, from 2018 to 2024, PHM in the railway sector advanced from basic monitoring systems to AI-powered solutions capable of handling complex, large-scale data and providing predictive insights to optimize maintenance. Despite significant advancements, challenges such as data quality, infrastructure integration, and real-time processing remain key areas for continued research and development.

2.5 Gaps and Challenges

In this section, we aim to understand the key gaps and the challenges to address those gaps based on the existing work during the period of 2018-2024. We illustrate the frequently reported gaps in Fig. 2.5, and detail them as follows.

- **Data Scarcity:** Data Scarcity is one of the root problem of achieving predictive accuracy to deploy reliable PHM system. The availability of comprehensive and high-quality data is the foundation for building accurate and reliable predictive models. Inaccurate data can lead to misinterpretations of asset health and unreliable predictions. Also the difficulty in benchmarking PHM models due to a lack of comprehensive dataset. The scarcity of data, particularly for specific failure modes and rare events, can hinder the development and validation of comprehensive prediction models.

  Gathering sufficient and appropriate data to represent the health parameters of railway components remains a critical challenge. Training robust AI models capable of handling the complexity of railway operations requires access to extensive and diverse datasets. Factors such as weather conditions, lighting, and the infrequent occurrence of defects or faults make it extremely challenging to capture the necessary data to extract the characteristics and patterns of an asset's health problems, including

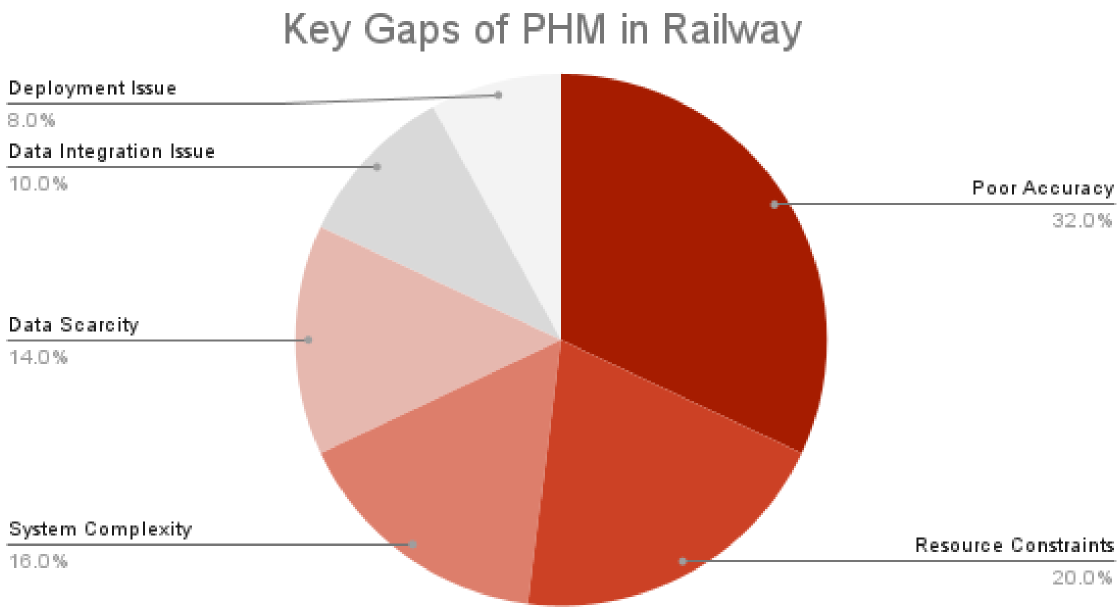


Figure 2.5: Key Gaps Identified in PHM Research from 2018-2024

cracks, wear, and tears. Clear images of such conditions are crucial. Additionally, integrating smart data-capturing devices into traditional railway infrastructure is complicated due to the outdated design of the components.

- **Data Integration Issue:** Data integration, centers the issue of combining data from diverse sources and formats within a PHM system. Ensuring the accuracy, consistency, and completeness of data from various sources is crucial for building reliable PHM systems. Both Existing work reported that the absence of labeled datasets limited the development of accurate predictive models.

  Addressing data integration becomes challenging due to the intricate nature of railway systems and the diverse sources of data that need to be combined for a holistic understanding of asset health. Accurately interpreting and correlating the raw data received from various sensors with the actual conditions of the rail infrastructure is a complex task. This requires advanced data processing techniques to filter noise, handle missing data, and translate sensor readings into meaningful insights about rail health. Inconsistent data formats, missing values, and errors in data collection can significantly impact the performance and reliability of PHM systems.

- **Poor Accuracy:** Poor accuracy stands as the mostly reported hurdle in PHM implementation in existing work. This involves the development of models capable of precisely forecasting the remaining of railway assets and anticipating potential failures with a high degree of confidence. Inaccurate predictions can have several detrimental consequences leading to wasted resources and increased operational costs. Poor accuracy also may cause increased downtime affecting service reliability and customer satisfaction. Ultimately it may risk safety of passengers and railway personnel.

  Addressing high predictive accuracy remain a concern for several challenges. The presence of unquantifiable factors affecting asset health poses a significant challenge for predictive models. Additionally, predicting sudden and unpredictable changes in asset condition accurately is not achieved when model is trained on limited samples. Variations in asset's conditions and operational parameters can significantly impact prediction accuracy.

- **Subsystem Integration Complexity:** Subsystem integration complexity underscores the difficulties in integrating various components and systems within the complex railway infrastructure. The integration of diverse technologies, including AI, and sensor networks into a unified DT ecosystem demands coordination to ensure compatibility and interoperability.

Data generated by different subsystems are often stored in isolated silos, making it challenging to access and integrate this information for comprehensive analysis. In addition, railway systems comprise a network of interconnected components like tracks, signaling systems, rolling stock, and power infrastructure. These subsystems often operate independently, utilizing different technologies and data formats. This inherent complexity poses a significant challenge when attempting to integrate them into a DT ecosystem.

- **Resource Constraints:** Resource constraints encompass the financial and computational limitations in developing and deploying PHM systems. This category highlights the need for cost-effective solutions that can be implemented within the budget constraints of railway operators. The existing high costs associated with traditional maintenance practices necessitate the development of PHM solutions that offer a positive return on investment. Although this challenge is mention in several research papers, there is hardly any direct discussion on the complexities and practicality of addressing resource constraints.

  In general, implementing advanced PHM systems, particularly those involving DT and AI, often requires a significant upfront investment in sensor technology, data infrastructure, software development, and skilled personnel. The increasing complexity of PHM models, particularly those utilizing data hungry ML models, demands substantial computational power for data processing, analysis, and model training. This necessitates investment in high-performance computing infrastructure, which can be expensive to acquire and maintain, especially for smaller railway operators.

- **Deployment Issue:** These deployment related issues present the practical aspects of integrating and sustaining these systems in real-world railway settings. Even though cloud-based deployment is a viable solution to address resource constraints such as high-performing computing infrastructure, ensuring cybersecurity, data privacy, connectivity, and latency in cloud-based deployments of railway PHM applications is extremely challenging due to several factors specific to the railway context.

For example, the delay in receiving and processing data from remote cloud servers to the train's control systems can impede the effectiveness of PHM applications, making it difficult to address issues promptly. The sensitive nature of data in the railway industry adds another layer of complexity to the deployment of PHM systems. Public perception and trust play a crucial role in the successful deployment of any technology, especially in critical infrastructure like railways. Data breaches or security incidents involving sensitive information can erode public trust and create resistance to the adoption of PHM systems, even if they offer significant safety and efficiency benefits.

2.6 Requirements

In this section we present the identified requirements to address key gaps for Railway PHM applications. The majority of the existing work focused on AI and DT paradigm in their methodologies. In the light of our previous literature synthesis, we summarize the following requirements and direction.

- **Data Availability:** Overcoming the logistical and technical challenges of gathering vast amount of data from heterogeneous sources across the railway network is needed. Data scarcity issue is required to be addrssed. Because, data scarcity is not merely a technical challenge; it's a fundamental prerequisite for unlocking the transformative potential of PHM in the railway industry. Without a clear understanding of asset behavior and degradation patterns, maintenance decisions may be suboptimal, leading to increased costs and potential disruption.

Synthetic data generation using LLM based approach can aid this gap [2]. Because, this approach has been gained attention in resolving data scarcity issues in various domains including healthcare, and manufacturing. The generative models like stable diffusion [42] can provide a way to generate rapid fault or defect images faster and flexibly to omit the need for costly physical modeling or real-world occurrences.

- **Data Integration:** Effective data integration is essential for creating a unified and comprehensive view of asset health, leading to improved prediction accuracy and better-informed decision-making. Cost-efficient data pre-processing pipelines, and advanced data fusion techniques are required to address inconsistencies. Handling multiple format of data in various mode like images of defect, lifespan time, operational records, defect inspection guidelines, even the video stream of railway operation and condition is crucial to address within the PHM system.

  Multimodal LLMs like LLAVA, ChatGPT-4, Gemini-1.5-Pro offers understanding image and text data without needing huge amount of samples or complex data standardization algorithm [80] [58] [69]. Application of multimodal LLMs may reduce the complexity of integration within a AI-driven DT system for railway PHM.

- **Predictive Accuracy:** Accurate predictions are crucial for minimizing disruptions to railway services. Inaccurate predictions can lead to unnecessary maintenance, wasting resources and potentially introducing new issues.

  Pre-trained models [19] that combine real-time and historical data effectively are needed. Incorporating domain-specific knowledge into AI models can enhance their accuracy. The capability of few-shot and zero-shot learning of customized domain specific LLMs [58] can be utilized to mitigate the predictive accuracy for newly evolving health issues of railway assets.

- **Integrated System:** Seamless integration is paramount for efficient data flow, real-time monitoring, and coordinated decision-making across the entire railway network. Ensuring effective communication and data exchange between various subsystems within the DT framework is crucial for accurate and comprehensive monitoring.

  To address these challenges, it is essential to design a modular DT ecosystem and that incorporate AI modules, and a QoE feedback loop [27] and [26]. Additionally, employing domain-specific LLMs that are fine-tuned or instruction-tuned can enhance their effectiveness. The multi-agent concept, which utilizes a single instance of an LLM for various tasks with different roles and capabilities, can also be strategically applied across various aspects of railway PHM.

- **Resource Constraints:** The existing work repeatedly emphasizes the crucial need to address resource constraints within the railway industry. High costs associated with developing, implementing, and maintaining advanced PHM technologies like DT need attention.

  Overcoming these constraints will require a multi-pronged approach, involving the development of supportive business models, prioritizing cost-effective solutions, and exploring emerging technologies like open-source Large Language Models (LLMs) and transfer learning. Synthetic data, LLM-based preprocessing, Quality of Experience (QoE)-based model adaptation, and faster, easier model customization may benefit the resource extensive steps of a PHM process. However, at a large scale, resource optimization algorithms need to be explored, and the impact of synthetic data on resource optimization must be assessed within a fully deployed PHM system.

- **Deployment Issue:** Addressing deployment issues is necessary for the long-term viability and effectiveness of PHM systems, ensuring their seamless integration into existing operational procedures. It requires careful planning, robust security measures, and the development of integration frameworks that facilitate seamless communication between cloud-based PHM systems and existing railway infrastructure.

  In ongoing research, technological advancements aim to overcome these obstacles, paving the way for more effective and practical deployment of cloud-based PHM solutions in the railway sector. Combining edge computing with cloud-based solutions could mitigate the challenges associated with cloud deployment. By processing data locally at the edge, latency can be reduced, and connectivity issues can be minimized. Hybrid approaches that leverage both real-time and historical data. for example, learning from near-realtime data (e.g., videoframes after certain intervals), Visual LLMs (VLMs) can identify patterns and detect anomalies in real-time. This enables immediate responses to potential issues without the need for extensive cloud processing.

2.7 Summary

AI-driven DT offers promising benefits to railway PHM process by facilitating its core steps including data acquisition, data processioning, model training and prediction within a closed feedback loop. Therefore, it is highly required to design and develop an AI-driven DT ecosystem for specific railway PHM application. Recent studies discusses the benefit of utilizing generative models and LLMs to resolve several persistent challenges in PHM applications for railway [84]. Because, these models hold potential for enhancing automation, decision-making, and intelligent maintenance strategies [93].

As the technology trend has been seen taking a leap towards using LLMs for railway PHM applications, the scope of LLM integration into AI-driven DT ecosystem is need to be designed. Also, the recent trends of applications focus on visual anomalies like crack or defect on railway assets. Therefore, we pick defect inspection of railway assets as an application area. In the following Chapter, we outline about the methodology to design and develop AI-driven DT ecosystem for visual defect inspection of railway assets.

Chapter 3

Methodology

This Chapter presents our proposed DT ecosystem (DefectTwin) for visual defect inspection in PHM for Railway. We detail the approach of incorporating LLM for the AI-inferencing into the the proposed ecosystem. We also explain how the AI-inferencing workflow works to maintain the QoE feedback loop in detail.

3.1 Overview of DefectTwin

We designed a DT ecosystem for railway defect inspection in light of the universal DT ecosystem proposed in [27]. The ecosystem comprises three key interconnected modules, as illustrated in Fig. 3.1. The user (system or human) collects defect data from various data sources, which is then processed and preprocessed to identify and understand railway defects through a multimodal interface. Users can also provide feedback based on their experience via the same multimodal interface. This feedback creates a continuous loop, ensuring improvements in the QoE of the railway defect inspection systems that follow

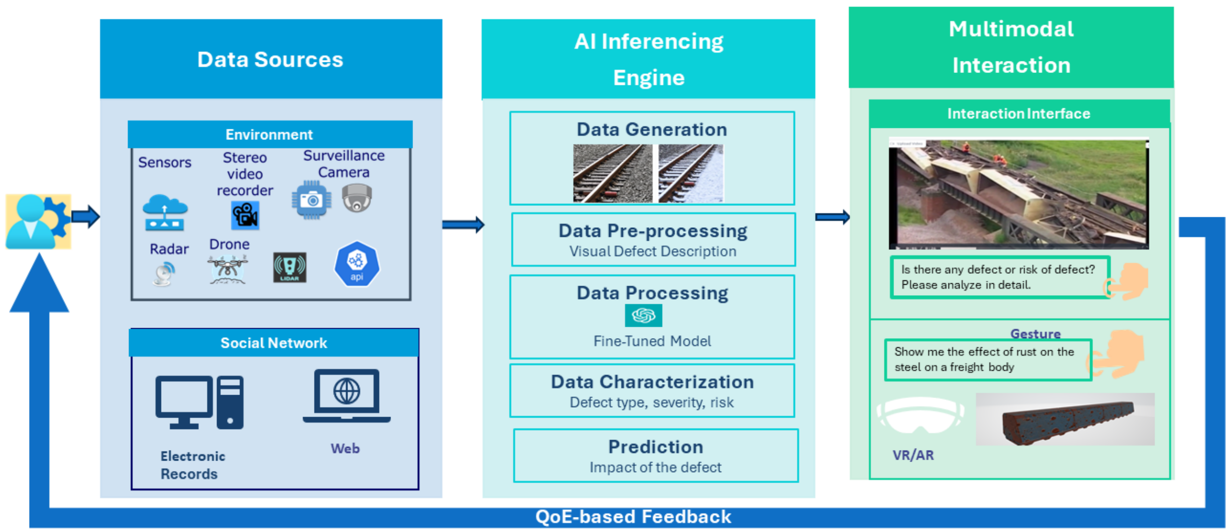


Figure 3.1: DefectTwin: A DT Ecosystem for Visual Defect Inspection in Railway PHM

the proposed DT ecosystem. The overall process in this ecosystem is aligned with the fundamental PHM process (See Fig. 2.1), except the training step. Instead of manual data pre-processing and training , the proposed DT ecosystem relies on LLM-based AI inference. More details of the DefectTwin ecosystem are outlined as follows.

3.1.1 Data Sources

The data sources, comprising various sensors and data acquisition systems, is responsible for continuously and periodically monitoring the physical railway infrastructure for signs of defects. This unit ensures comprehensive data collection, capturing essential parameters that could indicate potential issues.

Example

For instance, in a practical setting, surveillance cameras located at railway stations continuously send video streams to the DefectTwin system. These cameras capture high-resolution images and videos of the railway tracks, which are then analyzed for defects such as cracks or deformations. Additionally, environmental data is collected through weather APIs, providing context such as temperature, humidity, and precipitation, which can affect railway infrastructure.

3.1.2 AI Inferencing Engine

The AI-inferencing engine is the core of the DefectTwin ecosystem responsible for bringing intelligence to automate the defect inspection task. To ensure the collected data is suitable for AI analysis, it undergoes several preprocessing steps:

1. Data Generation: In cases where specific types of data are lacking, the system can generate additional synthetic datasets to fill the gaps. For instance, if there is insufficient data on a specific type of crack, synthetic images of this crack type can be generated using LLM APIs and added to the dataset. This ensures that the AI models are trained on a comprehensive and diverse dataset, improving their ability to detect and analyze defects accurately. This Data generation is proposed in DefectTwin to address the most challenging issue in automating defect inspection due to lack of data scarcity. We have detailed this step in Section 3.2.1.

2. Data Pre-processing: The raw video data is cleaned to remove any noise or irrelevant information, such as background frames including crowds in the platform, which do not pertain to the defects of railway tracks. The cleaned data is standardized to

ensure uniformity in format and scale, making it compatible with the AI models used in DefectTwin. For example, if we use a video-to-text generation model that captures defect characteristics well if frames are passed as images, the Data standardization unit aids that step. The standardized data is labeled with relevant tags, such as the type and location of defects detected. To enhance the dataset, various orientations of images and videos of defects are generated and added to the dataset to bring variety and better capture defect characteristics. This augmentation helps improve the accuracy of the AI models used for prediction and classification.

3. Data Processing: The pre-processed data is used to fine-tune a suitable LLM, customized specifically to capture the fine-grained details of defects that are often overlooked or hard to detect in real-life conditions. Traditional machine learning models typically rely on a fixed set of examples, which limits their ability to generalize when they encounter new or uncommon defects. By fine-tuning the LLM, we ensure that the AI model is not only learning from the data but also adapting to capture subtle details, even those that may not be present in the initial dataset.

4. Prediction: The AI inferencing engine is the central component of the DefectTwin ecosystem. It integrates multiple advanced AI models to perform high-precision defect detection and predictive maintenance. For instance, the types of defects and levels of defect risk are predicted.

3.1.3 Multimodal Interaction

The Multimodal Interaction unit empowers both human users and automated systems to engage with the AI inference engine's outputs in an intuitive and interactive manner. Using gesture-based controls within a user-friendly interface, users can seamlessly explore and analyze the AI-generated results. For enhanced visualization, the AI's outputs can be projected into immersive VR/AR environments, allowing users to explore and assess defect impacts in a highly interactive and realistic setting.

Examples

For example, users can inquire about potential defects, including their classifications and associated risk levels, derived from video streams. Furthermore, if users wish to visualize the progression of rust on a steel freight body, they can interact with a 3D model in a VR/AR environment. The multimodal processor integration (see Section 3.2.4) dynamically maps the AI-generated textures, enabling users to experience the impact of defects in a simulated, real-world context. This immersive approach enhances both understanding and decision-making for defect analysis and maintenance planning.

3.1.4 QoE Feedback

The QoE feedback loop in DefectTwin is crucial for maintaining high user satisfaction and system performance. It allows users to provide feedback on the quality of the generated media (e.g., images, and videos) and interact with the system through multimodal interfaces. This feedback is used to refine the AI models, ensuring that DefectTwin delivers the desired results. Specifically, users can interact with DefectTwin in the virtual space, evaluate the outcomes of AI inferencing, and send feedback on their experience through the multimodal interface. The AI inferencing unit accommodates this feedback, updates necessary parameters, and sends updated feedback to the users. This process runs in a loop to continuously enhance the QoE for DefectTwin users. By enabling continuous interaction and feedback from users, the system dynamically adjusts to meet user expectations and requirements.

Example

If a user wants to see rust on steel but the AI-inferencing engine outputs cracks instead, the user can send feedback indicating the discrepancy. This situation is not uncommon due to the involvement of generative models. In response, DefectTwin adjusts LLM parameters such as temperature or system messages based on the user's feedback. While many LLM chatbots have this feature, leveraging it for DT in DefectTwin presents unique challenges.

In our case, addressing QoE is not straightforward because we lack established policies or feedback to train a reinforcement human/AI feedback model. Therefore, we have handled it using our innovative solution. InstaUF ensures that user feedback is promptly integrated, allowing the system to dynamically adjust and improve the accuracy and relevance of the generated outputs.

In this section, we provided a high-level overview of the DefectTwin ecosystem. The workflow of the key enabler- M^2 LLM-based AI inferencing is described in the next section in more detail.

3.2 M^2 LLM-based AI Inferencing Pipeline

The proposed M^2 LLM-based AI Inferencing Pipeline aims to generate a high-fidelity in-domain synthetic dataset for fine-tuning a base LLM to improve the performance of multimodal decoders used for various purposes, such as text-to-image, video-to-text, and image-to-text. The workflow components in the proposed AI inferencing pipeline are presented in Fig. 3.2.

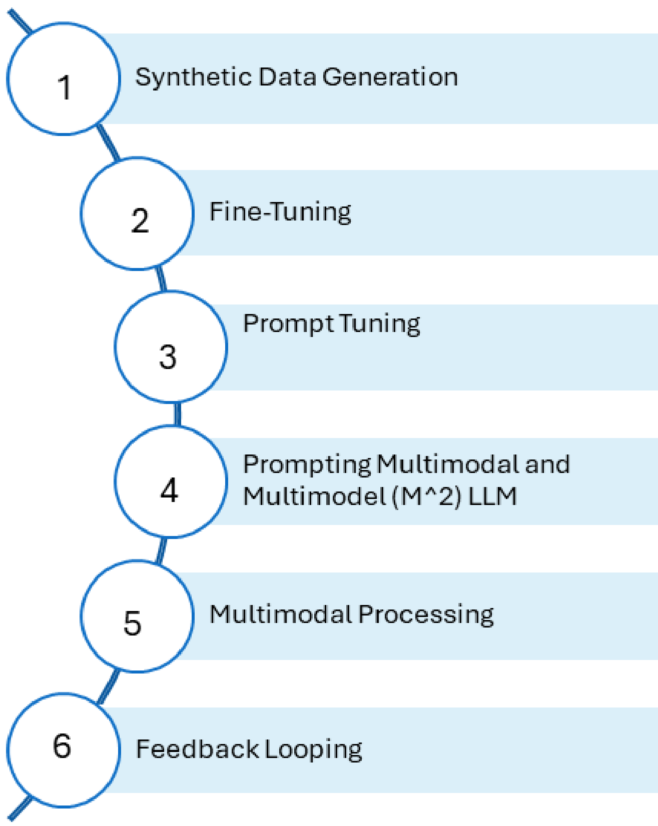


Figure 3.2: Workflow of the AI-Inferencing Engine Pipeline for Railway Defect Inspection

3.2.1 Synthetic Data Generation

In our proposed synthetic defect generation pipeline shown in Fig. 3.3, we leverage an LLM with visual captioning capabilities to create synthetic images with defects. The different tasks in this step are as follows.

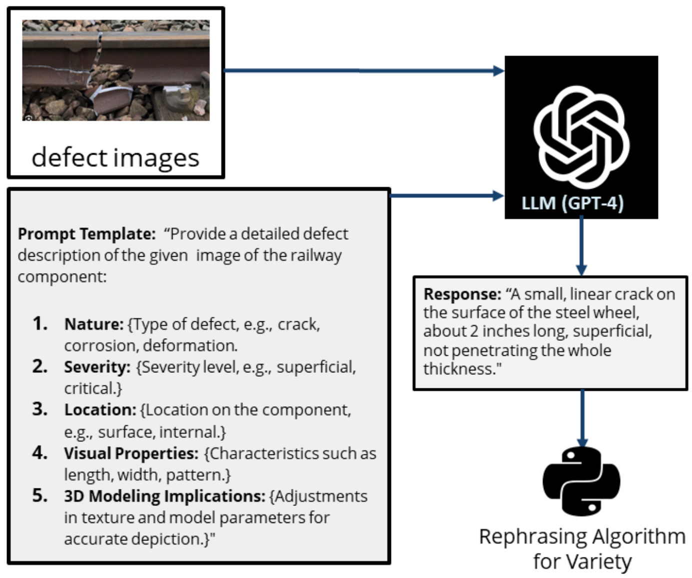


Figure 3.3: Synthetic Dataset Generation Pipeline

Template-based Caption Generation

We chose GPT-4, which has been utilized for creating prominent visual-instruction datasets such as LLAVA, MIMIC IT, Objaverse, and Sceneverse [48]. Furthermore, from the literature, we found that LLM fine-tuned using rephrased samples achieved high accuracy. As demonstrated in Fig. 3.3, the process starts by taking a raw image as input and passing it through the LLM to generate a descriptive caption using a popular visual captioning technique known as the template-based caption generation approach. For our specific application, we developed a prompt template in collaboration with domain experts to capture essential visual defect characteristics, particularly those challenging to capture in real-life scenarios.

Rephrasing Algorithm for Variety

The template-based caption serves as input for rephrasing, as given in Algorithm 3.1. This technique transforms the prompt into new, intricate illustrations that accurately depict the specified elements missing in the original caption. By utilizing this procedure, we produce a significant number of synthetic examples, providing rich training examples for tuning the custom LLM. The Defect LLM dataset (DLLMDS) Generation Pipeline aims to solve the challenge of data scarcity in deploying LLM-based DT solutions.

The proposed algorithm for generating DLLMDS, aims to maximize the variety of the samples (D) and minimize the reconstruction loss (L). This can be represented as:

$$\underset{D,L}{\text{maximize}} \quad D - \lambda L \tag{3.1}$$

where:

- D is the variety of the samples,
- L is the reconstruction loss,
- λ is a trade-off parameter that balances the two objectives.

Algorithm 3.1 Rephrasing Algorithm for Variety

Input: A list of template-based captions from defect images: `captions` $= [c_1, c_2, ..., c_n]$

Output: A dataset with multiple diverse and complex samples per caption, each accompanied by a system message.
Step-1: Create an empty list: `DS` $= []$
Step-2: For each caption (c_i) in `captions`, create an empty set for unique samples: `Samples` $= \{\}$
Step-3:
while the number of samples is less than K **do**
  Generate a new sample (s_{new}) using a language model for c_i.
  Add s_{new} to `Samples`.
end while
Step-4: For each generated sample (s_{new}) in `Samples`, formulate a system message. Combine s_{new} with the system message to form a structured data entry.
Step-5: Append all structured entries from `Samples` to `DS`. Ensure `DS` does not contain duplicates. =0

Each sample must be unique and more complex than the previous ones. The total number of samples must be less than or equal to K. These constraints can be represented as:

$$s_{\text{new}} \neq s_{\text{old}}, \forall s_{\text{old}} \in \text{Samples} \quad \text{and} \quad |\text{Samples}| \leq K \tag{3.2}$$

The algorithm achieves this by generating unique and complex samples until it reaches the maximum number of samples (K), while ensuring that the reconstruction loss is minimized.

The Algorithm 3.1, leverages template-based captions to generate a diverse and complex dataset. Each caption is expanded into multiple detailed samples using a language model, and each sample is paired with a system message to guide the fine-tuned DefectTwin LLM. Let us consider the following example:

Given a list of template-based captions: *captions = "A crack on the rail", "Corrosion at the joint", "A missing bolt"*

The algorithm generates multiple unique and complex samples for each caption. For example, for the caption "A crack on the rail", the generated samples could be:

Samples = s1, s2, s3. where, s1 = "A crack 3 inches long on the rail surface, perpendicular to the track direction."; s2 = "A diagonal crack on the rail with a depth of 2mm, located near the joint."; s3 = "A longitudinal crack running along the rail track, extending 5 inches."

Each sample is paired with a system message: *System Message = "Given the defect description provided, identify potential risks and recommend preventive measures."* The variety and complexity of samples are ensured by iterating until the number of unique samples, S_{new}, for each caption reaches a predefined limit K:

$$|S_{\text{new}}| < K \tag{3.3}$$

Each new sample s_{new} is generated using a language model prompt: *"You are generating data to train an LLM. Based on the initial description: ci, create a prompt/response pair ensuring the response is more complex and diverse than previous ones."*

Unique and complex samples s_{new} are added to the set Samples:

$$\text{if } (s_{\text{new}} \text{ is unique and complex}) \Rightarrow \text{add } s_{\text{new}} \text{ to Samples} \tag{3.4}$$

The dataset DS is compiled by appending all structured entries from Samples, ensuring no duplicates in DS:

$$DS = DS \cup \text{Samples} \tag{3.5}$$

3.2.2 Fine-Tuning

Once the synthetic dataset is generated using the proposed algorithm, it consists of both visual descriptions of defects and corresponding instructional data. We use this customer dataset to fine-tune a base model. We selected GPT-3 as the base model due to its proven ability to understand and generate human-like text, particularly when handling complex, domain-specific information [79].

The fine-tuning process involves customizing the GPT-3 to the specific domain of railway defect inspection. By training the model on the generated dataset, we enable the detailed visual cues associated with defects. In addition, we provide relevant instructions on how to address or analyze them to the model. Fine-tuning is performed iteratively, with multiple passes over the dataset to ensure that the model adapts to the intricacies of defect detection. Throughout this process, we continuously monitor performance metrics, such as the model's ability to generalize to unseen defects and its accuracy.

3.2.3 Prompt Tuning and Prompting the M^2 LLM

The initial input to the fine-tuned LLM for DefectTwin consists of three elements: system messages (SM), user prompts, and Virtual Prompt injections (VPI). We describe each of these elements in detail.

The fine-tuned LLM integrates these inputs to create a comprehensive multimodal input. This is then passed to a Text-to-Image (TTI) model, generating a visual representation of the defect. The diffusion model produces an image showing the steel wheel with a visible radial crack, enhancing the realism and accuracy of the defect depiction (Fig. 3.4). The final output is more informative and precise, aiding in better visualization and understanding of the defect.

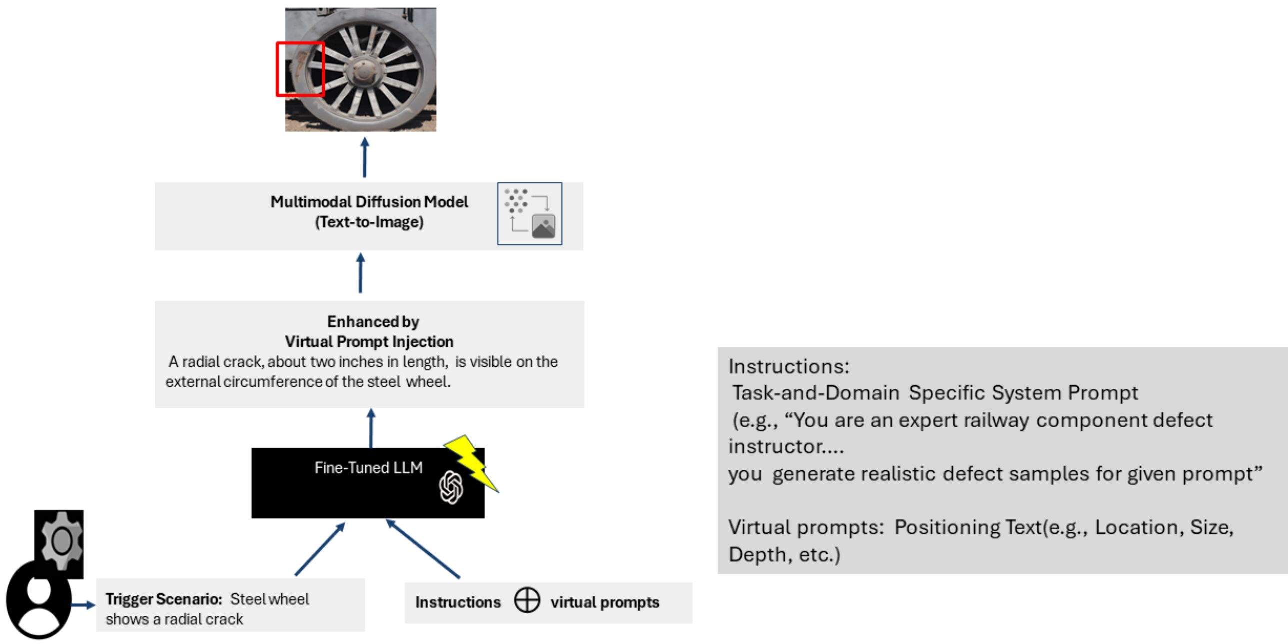
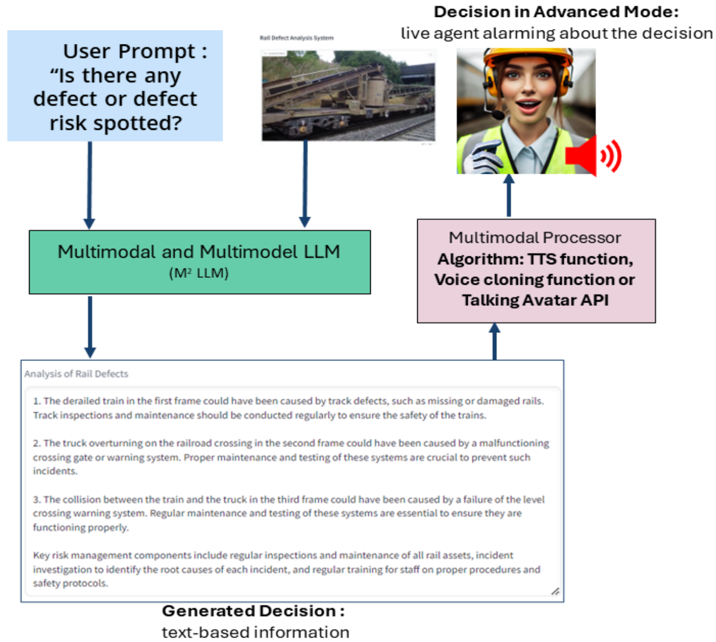
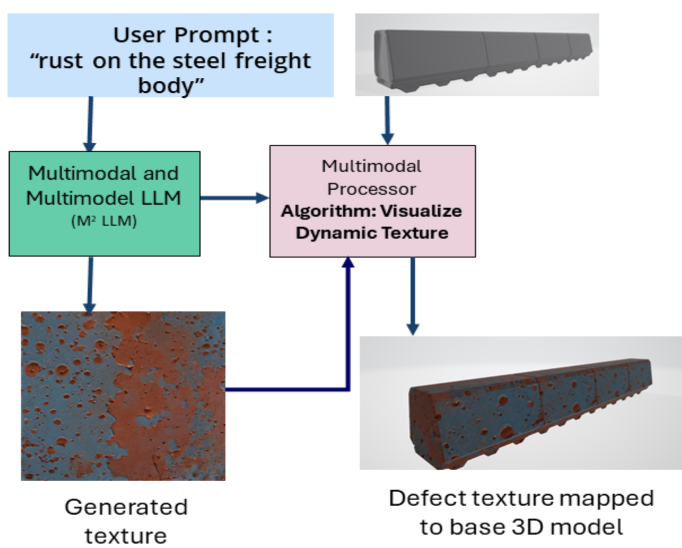


Figure 3.4: The Fine-tuned Defect LLM Integrates System Messages and VPI to Generate a Realistic Depiction of a Radial Crack on a Steel Wheel.

As illustrated in Fig. 3.4, the process begins with a user providing a simple trigger scenario: "Steel wheel shows a radial crack." This scenario is processed by the fine-tuned LLM using system messages, user prompts, and the Visual Prompting Interface (VPI). System messages, such as "You are an expert railway component defect instructor," provide context. The user prompt describes the defect scenario. VPI adds details like location, size, and depth, e.g., "A radial crack, about two inches in length, is visible on the external circumference of the steel wheel".



(a)



(b)

Figure 3.5: Use of Multimodal Processor in DefectTwin (a) Example I: Defect Analysis. (b) Example II: Predictive Visualization of Defect Characteristics.

3.2.4 Multimodal Processing

The tuned prompt is passed to multimodal diffusion models, such as text-to-image, image-to-text, 2D image-to-3D, and video-to-text. These models generate images, videos, or 3D models that accurately depict the defects based on the tuned prompt. The output of these models is passed to the Multimodal Processor. The primary task of this unit is to take the output from the M^2 LLMs and transform it into a format that is consumable for the end-user. This involves interpreting various types of inputs, processing and transforming the generated data, and finally, outputting the results in a user-friendly manner. We discuss the general workflow of the Multimodal Processor concerning the two examples illustrated in Fig. 3.5 in the context of a DefectTwin system.

Example 1 - Twining defect analysis process

Let us consider an application to replicate the defect analysis of the railway defect. As illustrated in Fig. 3.5a, DefectTwin acts like an information twin by automating the decision-making process by analyzing the video stream. In this example, the AI inferencing engine receives multimodal inputs, including a video stream and a user prompt. Based on this input M^2 LLM in the DefectTwin framework make a decision. However, the decision might not be in a format that's easy for the user to understand. The Multimodal Processor converts this decision into a talking avatar, effectively communicating complex information in a user-friendly manner.

Example 2- Texture Mapping and Visualization

In this example, the M^2 LLM generates a defect texture based on the user prompt. However, this raw texture might not be directly usable. This is where the Multimodal Processor comes into play. As you can observe in Fig. 3.5b uses a texture mapping algorithm to map the generated texture onto a base 3D model. This allows for dynamic visualization of defects in a simulated environment, enhancing the realism and usability of the data.

3.2.5 Feedback Looping

Users interact with the fine-tuned LLM through a multimodal interface and provide feedback on the system's performance. This feedback can vary as follows. 1) Positive or Negative Feedback. For example, positive feedback: "The LLM accurately identified the cracks in the railway track image." and negative feedback: "The LLM failed to identify the rust on the railway bolts." 2) Score-based Feedback. e.g., on a scale of 1 to 10: 6 is scored. Because it was able to identify major defects but missed out on minor ones. 3) Open-ended Feedback. e.g., Dissatisfaction: "You gave unrealistic defect". Refinement Need: "You should be able to differentiate between different types of defects such as cracks, rust, and mechanical wear." 4) Mixed Feedback - A score of 7 out of 10. "While it generally identifies major defects, it struggles with minor defects and often misses rust and small cracks". The user-interaction mechanism within DefectTwin is broken down as follows.

Feedback Processing

The feedback is input into an InstaHF pipeline that incorporates an instruct-tuned LLM designed to handle feedback. The instruct-tuned feedback LLM processes the feedback instantly. *Let (F) represent the feedback, where (F) can be a score or textual feedback.*

Based on the feedback, the system updates the message of the employed M^2 LLM. *Let (SM) represent the system message output by the LLM, and* SM_{new}*) represent the updated message.*

The update function can be represented as:

$$SM_{new} = Update(SM, F) \tag{3.6}$$

For example, if the LLM incorrectly identifies a crack in the railway track, the engineer might provide feedback as: a score of -1 and a comment "Missed the small cracks" Based on this feedback the system message might be instructed to pay more attention to the size of the defect.

Fine-Tuning Cycle

The variety of fine-tined LLM capabilities is required when the M² LLM cannot handle a specific type of defect, new samples are generated based on user feedback.

For example, if the LLM is fine-tuned on rust-based defects and incapable of handling mechanical defects like cracks or breaks, analyzing the user feedback new synthetic dataset is generated, and the current fine-tuned LLM is re-fine-tuned with new capability.

Each update builds on top of the previous model, retaining past improvements while incorporating new refinements. Let t represent the periodic update interval.

$$LLM_{t+1} = update(LLM_t, SM_{new}) \tag{3.7}$$

InstaHF for Optimization

The InstaHF pipeline is broken into two main components: the main function (see Algoritm 3.2) and the fine-tuning function (See Algorithm 3.3).

The main function handles the main loop of the algorithm, which collects user feedback, updates system parameters, and decides when to call the fine-tuning function. On the other hand, the fine-tuning function generates synthetic data using the DLLMDS pipeline and performs fine-tuning using the generated synthetic data to add more capabilities to the current fine-tuned LLM.

We aim to maximize user satisfaction score S subject to the constraint that number of fine tuning step FT is minimized. This can be formulated as:

$$\begin{aligned} \underset{S,FT}{\text{maximize}} \quad & S \\ \text{subject to} \quad & FT \leq T, \end{aligned} \tag{3.8}$$

where T is the total number of iterations.

So, the main function can be represented by the following iterative equation:

$$S_t = \begin{cases} 100 & \text{if } t\%\alpha = 0 \text{ or } S_{t-1} < \beta \\ f(S_{t-1}, F_t) & \text{otherwise} \end{cases} \tag{3.9}$$

where:

- S_t is the user satisfaction score at iteration t,
- F_t is the feedback at iteration t,
- $f(S_{t-1}, F_t)$ is the feedback processing function,
- $\alpha =$ ft_interval is the fine-tuning interval,
- $\beta =$ satisfaction_threshold is the satisfaction threshold.

Algorithm 3.2 Algorithm InstaUF - Main Function

Input:

- Fine-Tuned LLM (LLM_i), System Message (SM), Instruction (instruction)
- LLM settings parameters (LSP): Top-p (p), Top-k (k)
- Termination Criteria (tc), Fine-Tuning Interval (ft_interval)
- User Satisfaction Threshold (satisfaction_threshold)

Output:

- Fine-Tuned LLM (LLM_{i+1}) (only if fine-tuning occurs)
- Updated System Message (SM), Updated Instruction (instruction), Updated LSP

Initialize feedback vector (feedbacks) as an empty list.
Initialize iteration counter (counter) to 0.
Initialize user satisfaction score (satisfaction) to 100%.
while true **do**
  Collect user feedback (F).
  Append F to feedback vector (feedbacks).
  Process feedback using the Feedback Processing Function: (SM, instruction, p', k', satisfaction) = Update(SM, F, instruction).
  Update system parameters (SM, instruction, p', k').
  if satisfaction ¡ satisfaction_threshold **or** counter reaches ft_interval **then**
    Call Fine-Tune Function (LLM, D_s, p, k, feedbacks)
    Reset feedback vector (feedbacks).
    Reset satisfaction to 100% if it was below threshold.
    Reset counter to 0 if interval was reached.
  end if
  Increment counter by 1.
  if tc is met **then**
    break
  end if
end while
Return Updated LLM (only if fine-tuning occurred), SM, instruction, and LSP (p', k')
=0

Algorithm 3.3 Algorithm InstaUF - Fine-Tuning Function

Input:

- LLM, Synthetic Dataset (D_s), Top-p (p), Top-k (k), Feedback Vector (feedbacks)

Output:

- Fine-Tuned LLM (LLM_{i+1})

Generate synthetic dataset (D_s) using DLLMDS pipeline.
Fine-tune the LLM using (D_s): FineTune(LLM, D_s, p, k).
Set LLM_{i+1} to the fine-tuned LLM.
Return (LLM_{i+1}) =0

The fine-tuning function in Algorithm 3.3 can be represented by the following equation:

$$LLM_{i+1} = \text{FineTune}(LLM_i, D_s, p, k) \tag{3.10}$$

where:

- LLM_i is the fine-tuned LLM at iteration i,
- D_s is the synthetic dataset,
- p and k are the top-p and top-k parameters,
- $\text{FineTune}(LLM_i, D_s, p, k)$ is the fine-tuning function.

Fine-tuning is a resource-exhaustive procedure. When pre-trained LLM is used on cloud-platform there is a cost associated for the time spent for the fine-tuning. Also, with local models high-computing environment is required. The algorithm balances maximizing user satisfaction and minimizing fine-tuning, resetting satisfaction to 100% after fine-tuning and maintaining it above a threshold. So that, it limits fine-tuning to necessary times and regular intervals.

Chapter 4

Experiment

In this chapter, we detail the data and experimental analysis as a proof-of-concept of the proposed methodologies in this research. Additionally, the ablation study presented in this chapter discusses the role of our proposed methodologies in addressing the target research questions in light of the experimental results.

4.1 Data

In this research, we have employed both original [1] and synthetic data[2] to evaluate the performance and usefulness of DefectTwin. The details of the data used in this research are described as follows.

[1] Please download the kaggle dataset from here
[2] Please download our open-source dataset from here

4.1.1 Raw Data

In this research, we utilized two primary visual defect image datasets as a raw data source for the AI inferencing pipeline.

- **Canadian Pacific Railway (CPR) Defect Dataset** The first dataset is the 1000-defect Canadian Pacific Railway (CPR) defect dataset [34] [92]. This dataset includes images of freight components with various defects along with their descriptive labels. Specifically, it consists of 26 unique descriptive labels for the defect samples. This dataset is based on real-world data. However, its limited sample diversity as well as its highly-preprocessed format necessitate the inclusion of additional data for comprehensive evaluation.

- **Expanded General Category Dataset** We incorporated a more general category of samples to address the scarcity of samples in the CPR dataset. This expanded dataset includes images of various damaged railway components, such as freight components, rail tracks, and railway surfaces. We collected open-source images using keywords like- defect rail, defect damage, railway components, and freight components. Cumulatively, we selected 150 samples to form this expanded dataset.

4.1.2 Synthetic Datasets for Fine-Tuning

To enhance our datasets, we implemented a synthetic data generation pipeline. The following datasets are generated using the DLLMDS pipeline.

- **Defect visual-instruct dataset (D)** We first created a dataset that includes visual instructions and responses related to defects and maintenance.

- **Texture visual-instruct dataset (T)** To accommodate the need for understanding fine visual defect characteristics, our synthetic data generation pipeline generated defect texture visual response data during a fine-tuning cycle when fine-tuned LLM could not handle defect texture-specific responses.

The data description of the visual-instruct datasets are (T) is presented in Table 4.1

Table 4.1: Visual Instruct DS

Metric	**Value**
Total Instructions	500 X2
Unique Instructions	430(D) - 480(T)
Unique Responses	425(D) - 490(T)
Average Instruction Length	50 characters
Average Response Length	60 characters

Total 1000 examples are used for fine-tuning to customize the LLM specifically on visual description of the railway defect characteristics.

4.1.3 Test Data for Accuracy Evaluation

To evaluate the efficacy of our proposed defect twin framework, we required different types of defect datasets. We assessed the accuracy of image-based defect detection using 100 images from the Canadian Pacific Railway defect dataset. In total 600 test cases were used for evaluating the LLM's accuracy for image data. 100 images from the open-source expanded dataset. Each of the test image were paired with three different types of prompt including: i) defect detection: "Is there any defect, what type of defect is this?" ii) risk identification: "Is there any potential risk or incident?" iii) maintenance: "Is there any maintenance related issue?”. The examples of prompts used for defect inspection can be found by clicking on this link. The maintenance related prompt examples are also available.

To evaluate the accuracy of defect identification and defect risk identification from video inputs, we employed data from both the CPR defect dataset and open-source rail track defect data. Additionally, we included 15 minutes of video content, split into segments of 10 to 30 seconds each. Similar to the images, 50 video clips were also paired with the three different kind of prompts. In total we used 150 test cases for the video.

4.1.4 Obtaining the Ground Truth

As we required ground-truth data for evaluating the accuracy of our models, we utilized samples for which labels were available. For the image dataset, we used the CPR dataset and the rail defect dataset from Kaggle, both of which were labeled by domain experts. For the video dataset, we obtained the ground-truth data for frames containing defect risks or defects, as well as the types of defects, from the video captions. We followed a customized systematic approach to ensure label reliability as described below:

- **Ground Truth for Image-Based Evaluation:** We used labeled datasets, including the CPR dataset and the Kaggle rail defect dataset, both annotated by domain experts. To ensure consistency, we wrote a script that cross-referenced predicted labels with the dataset’s available labels. We selected a subset of defect categories from the dataset’s label sheet and matched model predictions against these predefined labels. **For example,** if a defect in an image was labeled as “Crack” by domain experts, our script compared the output from the model to this label. If the model predicted “Crack,” it was considered correct; otherwise, it was marked as an incorrect classification.

- **Ground Truth for Video-Based Evaluation:** Unlike static images, video frames required additional processing to extract ground-truth labels. We derived the ground-truth labels from the provided video captions, which described defects and defect risks in the corresponding video frames. The output for each video segment was compared to these ground-truth labels to assess accuracy. **For example,** a video caption might state, "A visible crack is detected on the rail surface at 00:10 seconds." Our script then checked if the class prediction for that frame matched the defect type (e.g., "Crack"). If the prediction aligned with the ground-truth label from the caption, it was considered correct; otherwise, it was incorrect.

Following the approach described above, we derived accuracy metrics by computing the number of correctly matched labels against the total number of evaluated cases, similar to traditional CNN-based models. However, ground truth in our study was derived from pre-labeled datasets or video captions may not always precisely align with defect occurrences in video frames. Even expert-labeled datasets can contain annotation inconsistencies or errors, leading to misleading ground truth labels. Therefore, some defects may be difficult to classify, causing variations in labeling across different experts. More details of the accuracy metrics are described in the next section.

4.2 Evaluation Parameters

We utilized various evaluation metrics to assess the M^2 LLM-based AI inferencing components of DefectTwin. In this section, we detail the rationale behind selecting various evaluation parameters in this study as well as their interpretation.

4.2.1 Performance Metrics for Accuracy

For evaluating the defect detection accuracy, we used the following metrics.

- **Precision** $= \frac{\mathbf{TP}}{\mathbf{TP+FP}}$**:**. A precision of 1 indicates perfect detection accuracy with no false positives. Precision measures the proportion of true positive defect detections among all positive detections made by the model. In our case, the precision metric ensures that the detected defects are indeed true defects, minimizing false positives.

- **Recall** $= \frac{\mathbf{TP}}{\mathbf{TP+FN}}$**:** The Recall metric is crucial for understanding how well the model identifies all possible defects, minimizing false negatives. A recall of 1 signifies all actual defects are detected with no false negatives.

- **F1-score** $= 2 \cdot \frac{\mathbf{Precision \cdot Recall}}{\mathbf{Precision+Recall}}$**:**. A higher F1-score indicates effective defect detection with a good balance between precision and recall. The F1-score ensures that both the true defects are accurately identified and the number of false positives is minimized.

- **AUC (Area Under the Curve):** The AUC score measures the ability of the model to distinguish between classes (defect vs. no defect) across various threshold settings. A higher AUC value indicates better distinction capabilities. An AUC of 1 indicates perfect distinction between defect and non-defect classes, while 0.5 suggests no distinction.

4.2.2 Performance Metrics for Quality of Generated Output

To evaluate the quality of the responses, we used several quantitative metrics on both text-to-text and video-to-text generated responses. For text, we considered token latency, while for video-to-text, we assessed frame vs. latency. The following metrics were employed to provide a comprehensive evaluation:

- **Answer Relevance:** Measures how relevant the generated response is to the query. In our defect detection context, a higher score (closer to 1) indicates that the responses are more relevant to the specific defects being queried.

- **Context Relevance:** Evaluates how well the response fits within the broader context of the conversation. For defect detection, this metric ensures that the responses are not only relevant but also contextually appropriate, enhancing the overall usefulness of the information provided.

- **ROUGE-L Score:** Measures the longest common subsequence between the generated response and the reference text. In our context, a higher ROUGE-L score indicates that the generated responses closely match the expected answers, reflecting the model's ability to generate accurate and precise defect descriptions and solutions.

- **Count of Tokens Generated:** Token is the atomic unit of text in LLM. This metric evaluates the number of tokens generated in the responses. A higher count indicates more detailed response. However, token generation has significant cost when used on cloud-platform. Also, higher number of token may cause hallucination when out-of-context information is produced. Therefore, limited number of token generation is required in this case.

- **Usefulness:** We followed the LLM-as-a-judge evaluation approach from another study [58], and adopted the following metrics to evaluate the model's performance on a scale of 1 to 10: helpfulness, relevance, accuracy, and level of detail. Helpfulness checks how well the response answers the question. Relevance ensures the response is related to the question. Accuracy verifies that the information is correct. The level of detail looks at how thorough and detailed the response is. A higher score indicates better performance.

- **Latency:** The latency is a crucial metric for understanding the efficiency and scalability of DefectTwin in the context of real-time applications. Lower latency indicates faster response times. Also, the time taken to generate the response charge costs for cloud-based LLM deployment. Therefore, lower latency is required for timely defect detection and maintenance actions.

These metrics provide a comprehensive understanding of the LLM-based AI inferencing performance in generating relevant, contextually appropriate, accurate, and timely responses, ensuring its usefulness in real-time and multimodal CE applications.

4.2.3 Candidate for Comparison

There are numerous LLM models currently available that perform a wide range of tasks, both as single models and multimodal models. However, for our specific evaluation, we selected instruction-following multimodal LLMs, such as Instruct-BLIP, GPT-4, LLAVA, and Gemini-Pro-Vision. Table 4.2, outlines the models selected, the criteria for their selection, and the identified contexts and generalizability aspects considered in our evaluation.

To understand the relevance of the generated responses, we focused on in-domain contexts such as defect description, defect identification, and defect maintenance as the reference contexts. Test prompts were specifically crafted to align with these domains to ensure accurate evaluation.

Table 4.2: Model Selection Criteria and Contexts

Items	**Selection Criteria**	**Identified**
Models	Instruct Multimodal LLM / Multiagent LLM	Instruct-BLIP, GPT-4, LLAVA, Gemini-Pro-Vision
Context	Test prompts to domain	Defect description, case-based, Maintenance
Generalizability	In-domain Visuals, Zero-shot	Infrastructure (Bridge, Station), Assets (Wheel, Gate, Surface), Track

We considered railway infrastructures like bridges and stations as out-of-domain railway components for defect analysis to evaluate zero-shot generalizability. Additionally, we selected specific defect categories such as defects in freight bodies, including wheels, gates, doors, rail surfaces, and tracks. By including in-domain visuals and zero-shot scenarios, we aimed to comprehensively assess the models' capabilities in familiar and novel contexts.

4.3 Ablation Study

We conducted several ablations to evaluate the impact of DefectTwin for railway defect inspection incorporating multimodal data. This section answers the target research questions we set for this research as followings.

4.3.1 Does DefectTwin Achieve Variety in Synthetic Examples?

The proposed synthetic dataset generation approach for fine-tuning significantly enhances the diversity in defect-specific characteristics compared to the basic visual captioning method. This enhancement is achieved through a rephrasing algorithm that diversifies the template-based captions, leading to richer and more varied descriptions within the dataset.

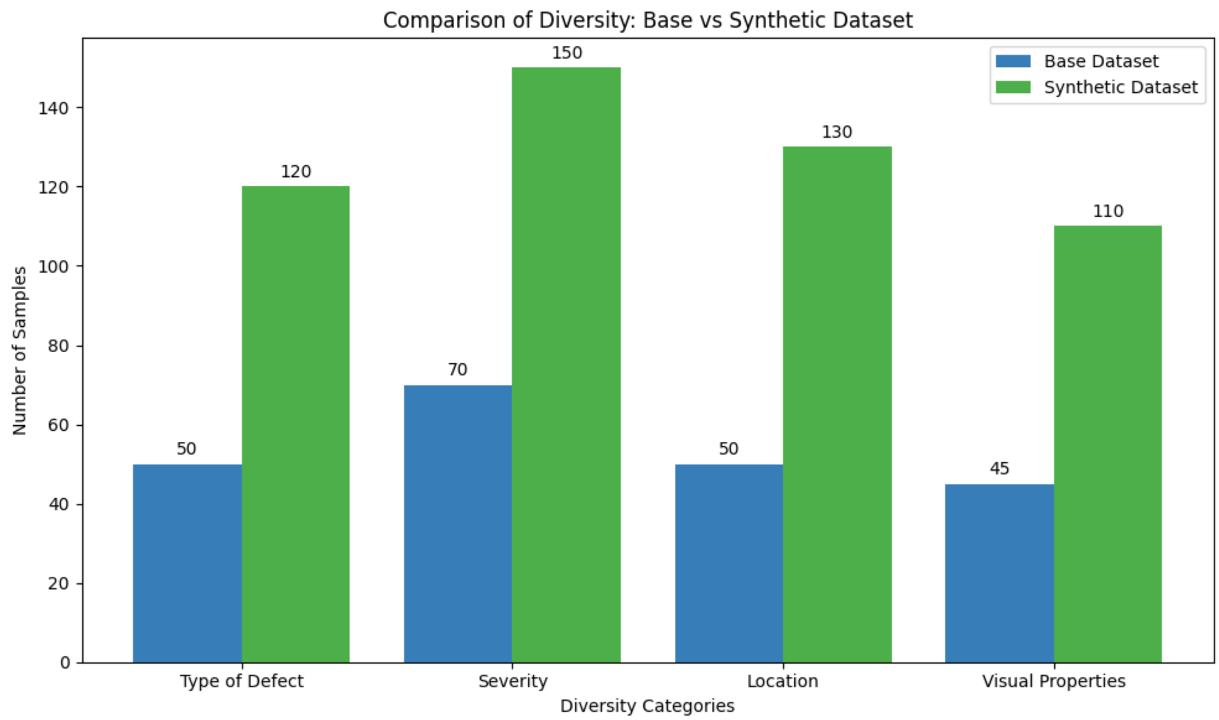
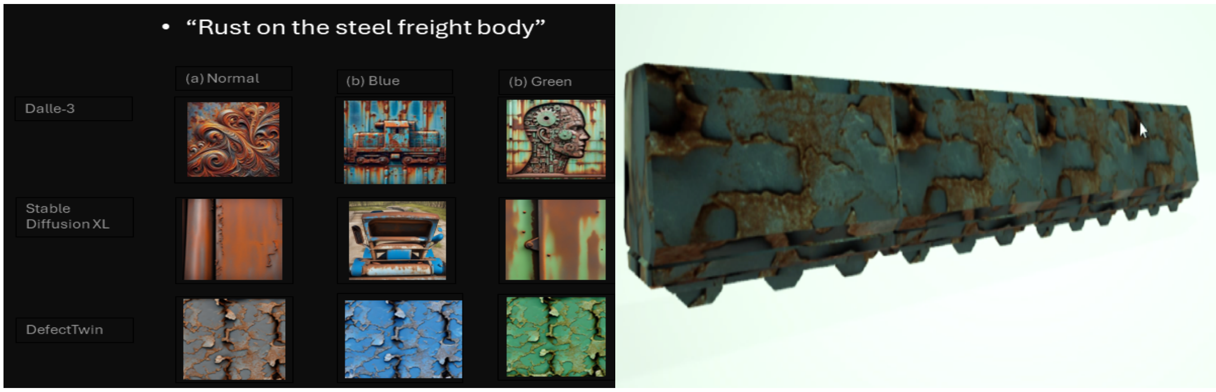


Figure 4.1: Sample Variety Achieved by Employing DLLMDS pipeline.

As illustrated in the comparative diagram Fig. 4.1, the DLLMDS pipeline effectively captures a comprehensive range of defect characteristics in the synthetic dataset, which is minimally covered by the base dataset generated through simple visual captioning.

The impact of fine-tuning with a diverse set of examples on texture generation for defect visualization is demonstrated in the subsequent output. This ablation involves generating the texture 'rust on the steel' on a freight body in three distinct colors—Normal, Blue, and Green—as shown in the accompanying Fig. 4.2.

Each model was tasked with the same objective—to generate realistic rust textures—but



(a) (b)

Figure 4.2: Comparative Analysis of Rust Texture Simulation on Steel Freight Bodies Across Different Generative Models. (a)Model Comparisons for Simulating Rust Textures in Varied Colors (b)3D Visualization of Rust Impact on Steel Freight Body

the results varied significantly. Dalle-3 and Stable DiffusionXL outputs, while visually interesting, tend to abstract or stylize the rust effect, which may detract from practical applications where precise defect portrayal is necessary. Fig. 4.2a, displays the outputs from three different generative models: Dalle-3, Stable DiffusionXL, and DefectTwin. DefectTwin, on the other hand, produces outputs that closely resemble real-world rust patterns, capturing the intricate details and variegated coloration characteristic of actual rust. The fine-tuning process and advanced preprocessing functions have significantly enhanced the quality and realism of the output textures. These textures exhibit a vivid representation of rust, providing a granular and lifelike appearance that greatly aids in the realism of defect visualization as illustrated in Fig. 4.2b. This improvement is particularly beneficial for applications requiring high-fidelity visual simulations, such as in maintenance planning and predictive diagnostics.

Also depicted in Fig. 4.3, for three different scenarios, our proposed tool TextureMeDefect captured the fine-grained details to address the user's request more appropriately than other models. For scenario-1, the base model SDXL generated texture on a small freight, while DALLE-3 generated creative art on the texture. Because none of these models are

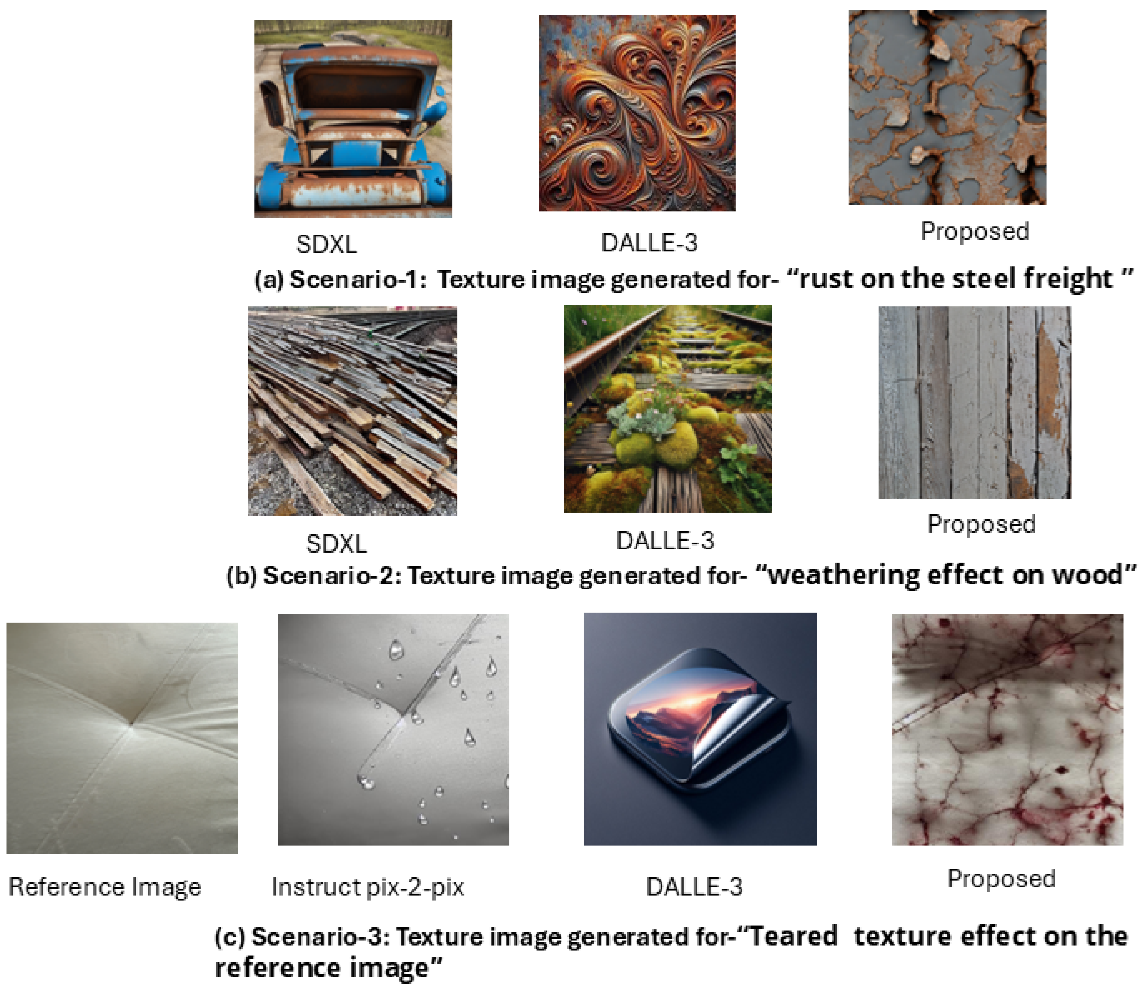


Figure 4.3: Comparison of Synthetic Textures Using Proposed Vs Existing Image Generation Tool

customized specifically for defect texture details, which makes these generated textures unrealistic to apply directly to a 3D model. A similar result is seen for the custom prompt-based texture generation. Interestingly, for scenario-3, the instruct pix-2-pix base model generated an image with a tear (in the context of crying) due to hallucination. Additionally, for such advanced features to generate in-paint texture, DALLE-3 completely fails to produce relevant output. This is where our proposed AI-inferencing Engine stands out compared to the existing solutions in generating realistic textures.

4.3.2 Does DefectTwin Identify Unseen Data and Classes with High precision and Consistency?

To specifically address the generalizability of the DefectTwin AI inferencing models based on the zero-shot performance. This ablation study aims to understand the capability to adapt and perform accurately when presented with new scenarios outside their training scope.

- **For image data:** The Proposed Model exhibits a zero-shot precision of 0.60 and an F1-score of 0.62, markedly higher than other models (See Table 4.3a). This performance suggests necessary feature extraction and generalization capabilities that are crucial for practical applications where unknown defect types may appear. In contrast, other models like Instruct-BLIP and LLAVA-Instruct show significantly lower precision and F1-scores (e.g., Instruct-BLIP at 0.35 precision and 0.37 F1-score), indicating a struggle to generalize to new defect types and scenarios.
- **For video data:** Despite the inherently more challenging nature of video due to its dynamic content, the Proposed Model still leads with a precision of 0.55 and

Table 4.3: Performance Metrics of Models on Different Media

a.	**In-Domain (Track, Assets)**				**Zero-Shot (Infrastructure)**			
Model	**Precision**	**Recall**	**F1-score**	**AUC**	**Precision**	**Recall**	**F1-score**	**AUC**
Instruct-BLIP	0.55	0.58	0.57	0.58	0.35	0.40	0.37	0.38
LLAVA-Instruct	0.85	0.86	0.85	0.86	0.45	0.45	0.47	0.48
GPT-4o	0.68	0.64	0.62	0.83	0.40	0.45	0.42	0.43
Gemini-Pro-Vision	0.88	0.88	0.88	0.89	0.48	0.50	0.49	0.50
Proposed	0.92	0.93	0.92	0.93	0.60	0.65	0.62	0.63

b.	**In-Domain (Track, Assets)**				**Zero-Shot(Infrastructure)**			
Model	**Precision**	**Recall**	**F1-score**	**AUC**	**Precision**	**Recall**	**F1-score**	**AUC**
Instruct-BLIP	0.30	0.35	0.32	0.34	0.20	0.25	0.22	0.24
GPT-4o	0.65	0.68	0.65	0.67	0.52	0.52	0.51	0.53
LLAVA	0.35	0.40	0.37	0.39	0.25	0.30	0.27	0.29
Gemini-Pro-Vision	0.71	0.72	0.71	0.73	0.45	0.48	0.45	0.47
Proposed	0.76	0.74	0.77	0.77	0.55	0.58	0.55	0.57

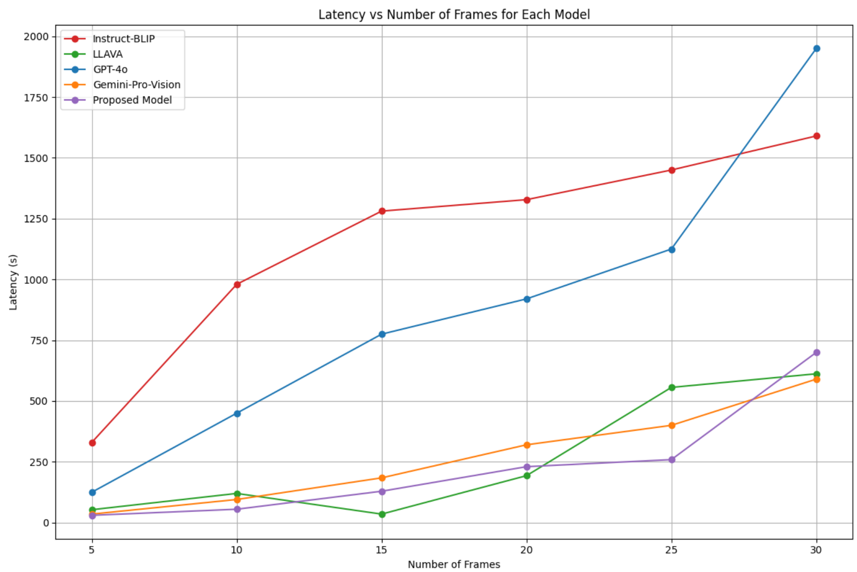
an F1 score of 0.55 (See Table 4.3b). These figures, while lower than those for image data, still outperform other models like GPT-4o and Gemini-Pro-Vision, which demonstrate lesser abilities to handle unseen video data effectively.

The generalizability analysis based on zero-shot performance highlights the Proposed Model's capabilities and areas for enhancement.

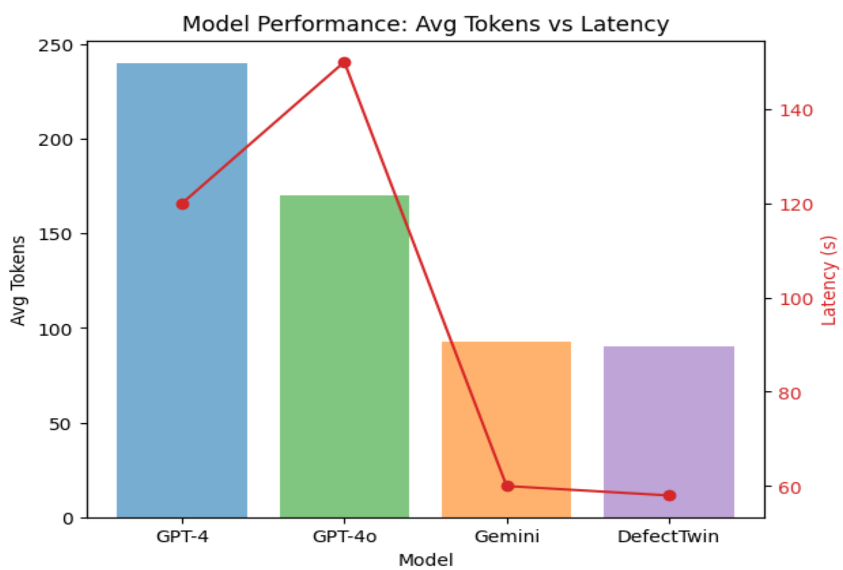
4.3.3 Does DefectTwin Adapt Based on User Needs?

We conducted this ablation based on several examples in the context of railway defect inspection. The outcomes of this study are presented below.

- **Surface Defect Visualization:** Initially, DefectTwin's visualization of a railway tunnel with surface defect lacked detail. After fine-tuning, it produced detailed and realistic textures accurately depicting surface defects (Fig.4.5a). In contrast, Dalle-3 and SDXL-Light provided visually appealing but less accurate results (Fig. 4.5b).



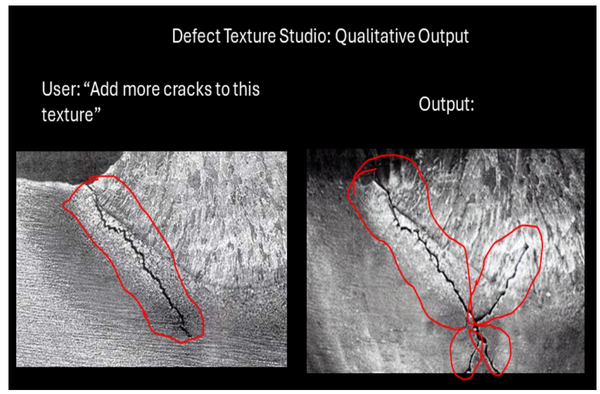
(a)



(b)

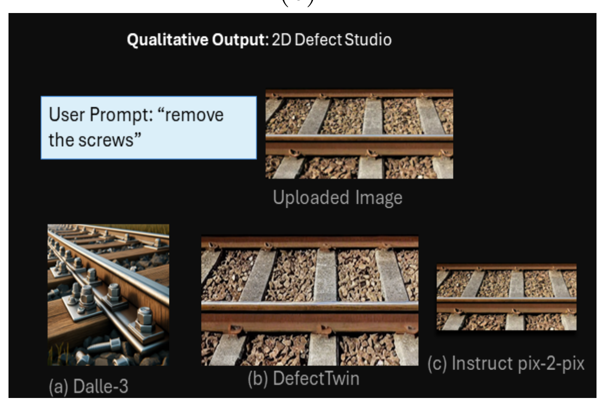
Figure 4.4: (a) Latency and Token Generation: Text-To-Text (b) Latency VS Number of Video Frames Processing

(a)



(b)

(c)



(d)

Figure 4.5: (a) Surface Defect Visualization by DefectTwin (b) Surface Defect Visualization by Existing Model (c) Screw Removal Prompt (d) Crack Enhancement Prompt

- **Crack Enhancement Prompt:** Before fine-tuning, DefectTwin's output for adding cracks was not detailed enough (Fig.4.5c). Post-fine-tuning, it accurately added pronounced and detailed crack textures, significantly enhancing defect visibility. Compared to this, other models struggled to achieve the same level of detail.

- **Screw Removal Prompt:** DefectTwin effectively removed screws while maintaining the integrity of the underlying image (Fig.4.5d), demonstrating superior performance. Dalle-3 provided an aesthetically pleasing but less accurate response, while Instruct pix-2-pix offered more practical outputs but was not as precise as DefectTwin.

The study on the qualitative outputs clearly demonstrates DefectTwin's ability to adapt its responses based on specific user inputs, especially after undergoing fine-tuning focused on visual texture descriptions. Such adaptability is required for applications requiring high fidelity in defect depiction and ensures that the tool can be reliably used in diverse operational settings. The improvement post-fine-tuning underscores the importance of specialized training datasets in enhancing the model's ability to interpret and respond to complex defect-related prompts accurately.

4.3.4 Is DefectTwin Resource-efficient?

To address this question, we focus solely on token usage and latency resulting from the involvement of multiple LLMs in our system. However, to fully assess resource efficiency, a more comprehensive evaluation should be conducted in a practical setting, where the system is deployed over a connected railway network. Here, we need to measure the latency and token conciseness to evaluate the suitability of the proposed AI inferencing approach for integrating seamlessly into real-world DT applications. We outline our findings as followings.

- **Performance on Token Generation and Latency:** The graph depicted in Fig. 4.4b, illustrates the relationship between the average number of tokens generated by each model and their corresponding latency. DefectTwin shows an optimal balance with a moderate number of tokens and comparatively lower latency than GPT-4O, which exhibits a peak in latency despite a decrease in token count. This suggests that DefectTwin manages to maintain efficiency in processing, crucial for real-time applications, where delay can critically impact user experience.

  DefectTwin exhibits a consistent but lower increase in latency with more frames than Instruct-BLIP [13] and LLAVA (see Fig. 4.4a), demonstrating better scalability for continuous data streams like video monitoring.DefectTwin's efficient token management reduces computational load and minimizes irrelevant responses, crucial for accurate and fast processing in systems with limited resources.

- **Latency Across Multiple Frames:** The second graph in Fig. 4.4a, extends this analysis to video data, depicting latency trends as the number of frames increases. Here, DefectTwin shows a consistent increase in latency with more frames but remains on a lower trajectory compared to models like Instruct-BLIP and LLAVA, which exhibit significantly higher latency spikes. This indicates that while DefectTwin's latency increases with more complex inputs, it scales more gracefully than other models, which is vital for applications involving continuous data streams, such as video monitoring through surveillance cameras.

- **Token Conciseness and System Efficiency:** The concise handling of tokens by DefectTwin is crucial, especially when integrated with systems like text-to-image models where verbose responses could mislead the decoders, leading to inaccurate or out-of-context outputs. The efficient token management by DefectTwin ensures that only necessary information is processed, reducing the computational load and minimizing the risk of generating irrelevant responses. This efficiency is paramount when system resources are often limited, and processing speed is critical.

- **Estimated Cost Analysis:** We estimate a cost analysis based on both token generation and processing time for each scenario using the fine-tuned model. As illustrated in the graphs in Fig. 4.4a , the latency varies with increased number of frames. Also form Fig. 4.4b, we can observe that the latency changes with the number of token generation. Additionally Based on these factors the cost can be estimated using the following formula:

$$
\begin{aligned}
C = & \sum_{i=1}^{n} (\text{Tokens}_i \times \text{Cost per Token}_i) \\
& + \sum_{j=1}^{m} \left(\text{Processing Time}_j \times \text{Cost per Second}_j\right)
\end{aligned}
\tag{4.1}
$$

Where Tokens_i is the number of tokens generated in the ith task; Cost per Token_i is the cost per token in the ith task; Processing Time_j is the processing time in seconds for the jth model; and Cost per Second_j is the processing cost per second in the jth model.

Let us consider for processing a video clip of 15 s token generation is low upto 150 tokens) with a short processing time, resulting in the lowest overall cost, while for a 30s clip, token generation increases to upto 260 tokens due to more raising the overall cost, despite manageable processing times. This analysis indicates that while token generation is a key factor, processing time plays a critical role in overall cost, particularly for scenarios involving more computationally demanding tasks.

4.3.5 Does DefectTwin Generate Useful and Relevant Responses?

Analyzing the performance across various models in text-only defect analysis, we observe distinct patterns and levels of effectiveness in handling specific defect-related queries. A detailed comparison and analysis based on the metrics of Answer Relevance, Context Relevance, Helpfulness, Relevance, Accuracy, and Level of Detail is presented as follows.

- **Relevance:** Table 4.4, and Table 4.5 both demonstrate that comparatively DefectTwin performed better than related models in most of the tasks.

  However, relevance decreased significantly for measurement or maintenance-related tasks. Because the fine fine-tuning focused more on defect description rather than monitoring. Benchmarks like GPT-4o and Gemini show moderate to low performance in these metrics, which could be due to less effective training on domain-specific data or limitations in their understanding of the context within technical settings.

  By contrast, DefectTwin excels notably in both Answer Relevance and Context Relevance across different categories (Defect Description, Case-Based, Maintenance, and Measurement), indicating a robust understanding of the query context and delivering relevant responses. For instance, in defect description, DefectTwin achieves an Answer Relevance of 0.88 and a Context Relevance of 0.97, substantially higher than its counterparts.

- **Usefulness:** The Fig. 4.6, illustrates that DefectTwin outperformed existing models in terms of helpfulness, relevance, accuracy, and level of detail scoring above 8 out of

Table 4.4: Answer Relevance Across Different Task: Multimodal

(a) Defect Detection

Model	**Answer Relevance**	**Context Relevance**	**ROUGE-L Score**
Instruct-BLIP	0.65	0.37	0.31
GPT-4o	0.43	0.52	0.54
LLAVA	0.45	0.58	0.36
Gemini-Pro-vision	0.41	0.51	0.21
Proposed	0.79	0.97	0.94

(b) Defect Risk Identification

Model	**Answer Relevance**	**Context Relevance**	**ROUGE-L Score**
Instruct-BLIP	0.17	0.20	0.22
GPT-4o	0.36	0.49	0.32
LLAVA	0.59	0.52	0.21
Gemini-Pro-Vision	0.31	0.67	0.25
Proposed	0.78	0.95	0.92

(c) Maintenance Recommendation

Model	**Answer Relevance**	**Context Relevance**	**ROUGE-L Score**
Instruct-BLIP	0.35	0.22	0.476
GPT-4o	0.51	0.35	0.62
LLAVA	0.27	0.43	0.33
Gemini-Pro-Vision	0.35	0.53	0.28
Proposed	0.46	0.86	0.84

Table 4.5: Answer Relevance Across Different Task: Text

(a) Defect Description

Model	Answer Relevance	Context Relevance
GPT-4	0.65	0.87
GPT-4o	0.54	0.62
Gemini	0.41	0.51
Proposed	0.88	0.97

(b) Case-Based

Model	Answer Relevance	Context Relevance
GPT-4	0.62	0.89
GPT-4o	0.65	0.55
Gemini	0.88	0.63
Proposed	0.91	0.95

(c) Maintenance

Model	Answer Relevance	Context Relevance
GPT-4	0.61	0.44
GPT-4o	0.57	0.52
Gemini	0.39	0.54
Proposed	0.82	0.89

(d) Measurement

Model	Answer Relevance	Context Relevance
GPT-4	0.63	0.58
GPT-4o	0.52	0.48
Gemini	0.41	0.56
Proposed	0.64	0.59

most of the criteria.

Figure 4.6: Response Utility for Rail Defect Inspection.

ChatGPT-4 and GPT-4o, while scoring reasonably in some categories, do not exhibit the same level of performance as DefectTwin, especially in categories requiring deep domain expertise like Maintenance and Measurement. The Gemini model, similarly, lags in more nuanced categories, which highlights potential gaps in its training or model architecture that may not fully capture the complexities of defect-related data.

The strong performance of DefectTwin, especially in providing contextually relevant and detailed responses, suggests that its training on a diverse and comprehensive defect dataset, possibly enhanced with domain-specific optimizations, significantly impacts its effectiveness. However, to fully evaluate the utility of the response human(expert)-evaluation is crucial to be addressed in further improvements of DefectTwin.

While DefectTwin performs well overall, there is notable variability in its performance in Measurement-related queries compared to other categories. Although still scoring relatively high, this suggests that DefectTwin may benefit from further training or refinement to enhance its understanding and response generation in highly technical or quantitatively driven contexts.

4.3.6 Is a Deployed Tool on DefecTwin Ecosystem Usable?

To address this question we deployed a tool TextureMeDefect, based on the DefectTwin ecosystem. This web-based tool allows users to create realistic defect textures interactively on images of railway components taken with smartphones or tablets. We analyzed the software usability score (SUS) across three scenarios as depicted in Fig. 4.7.

In **Scenario-1** (Library-Based Selection), users select predefined materials and defects from a library to generate textures using pre-trained models; in **Scenario-2** (Creative Prompt-based Generation), users write prompts to describe desired textures, which the system refines and accordingly generates the texture; and in **Scenario-3** (Image-Based Generation), users upload images and specify defects to visualize them on specific material, enabling a simulative understanding of defects on real-life components. Following we detail the participants, procedure , and score interpretation for our evaluation.

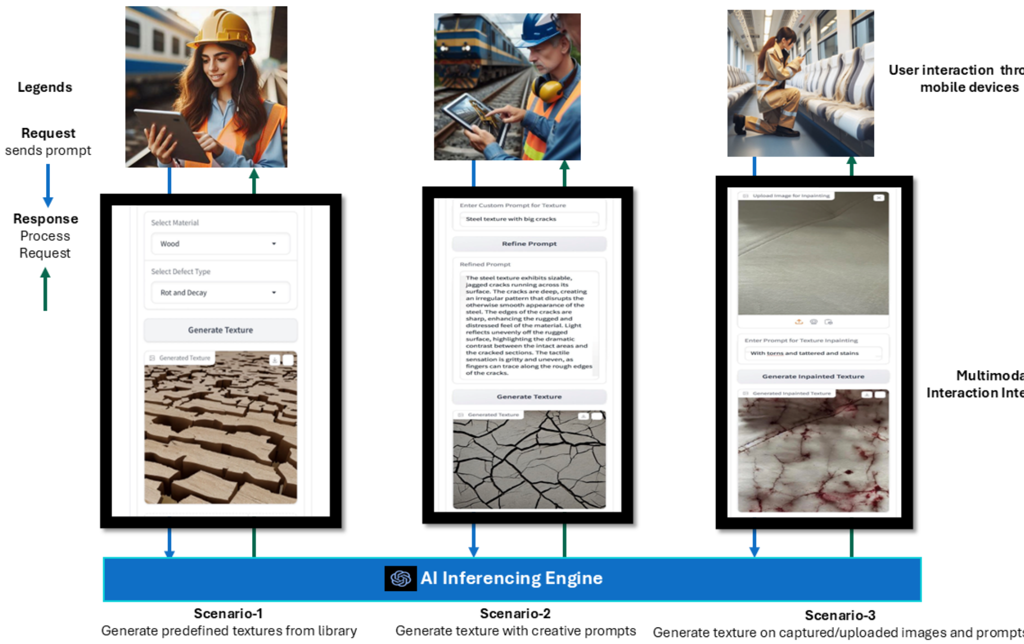


Figure 4.7: TextureMeDefect Tool for Defect Texture Generation on Mobile Devices [31]

Participants

A total of 15 voluntary participants took part in the testing with the following distribution: 46.67% Android users and 53.33% iOS (iPad) users. Among them, 87.10% of the users are expert users who are familiar with generating image using AI-tools, while 12.90% are non-expert users who have never performed AI-based image generation.

Procedure

The objective of this usability study is to evaluate how users interact with TextureMeDefect through three scenarios stated earlier. Here are the modified 10 SUS questions to evalaute the TextureMeDefect tool, focusing on the defect texture generation process:

- **Q1:** I found the process of generating defect textures using the input (dropdown/textbox/image upload button) options to be straightforward.

- **Q2:** I felt confident generating defect textures by using this user interface.
- **Q3:** The system's instructions for generating textures through different modes (dropdown, prompt, image upload) were clear and easy to understand.
- **Q4:** I believe that most users would learn to use this tool quickly.
- **Q5:** I felt that the time taken to generate defect textures was reasonable.
- **Q6:** I found the system unnecessarily complex when using custom prompts. (Reverse scored)
- **Q7:** I would need technical assistance to use the tool for generating defect textures. (Reverse scored)
- **Q8:** The tool was smooth and responsive when generating textures from images and descriptions.
- **Q9:** I found the generated defect textures to be realistic and aligned with my input.
- **Q10:** I would use this tool again for defect texture generation in the future.

The SUS Scale provides a global view of subjective assessments of usability, scored on a scale from 0 to 100. Questions 6 and 7 are reverse scored (i.e., a higher agreement is negative for usability). Questions 1, 2, and 9 focus on specific modes (dropdown, prompt-based, and image-based) in the tool.

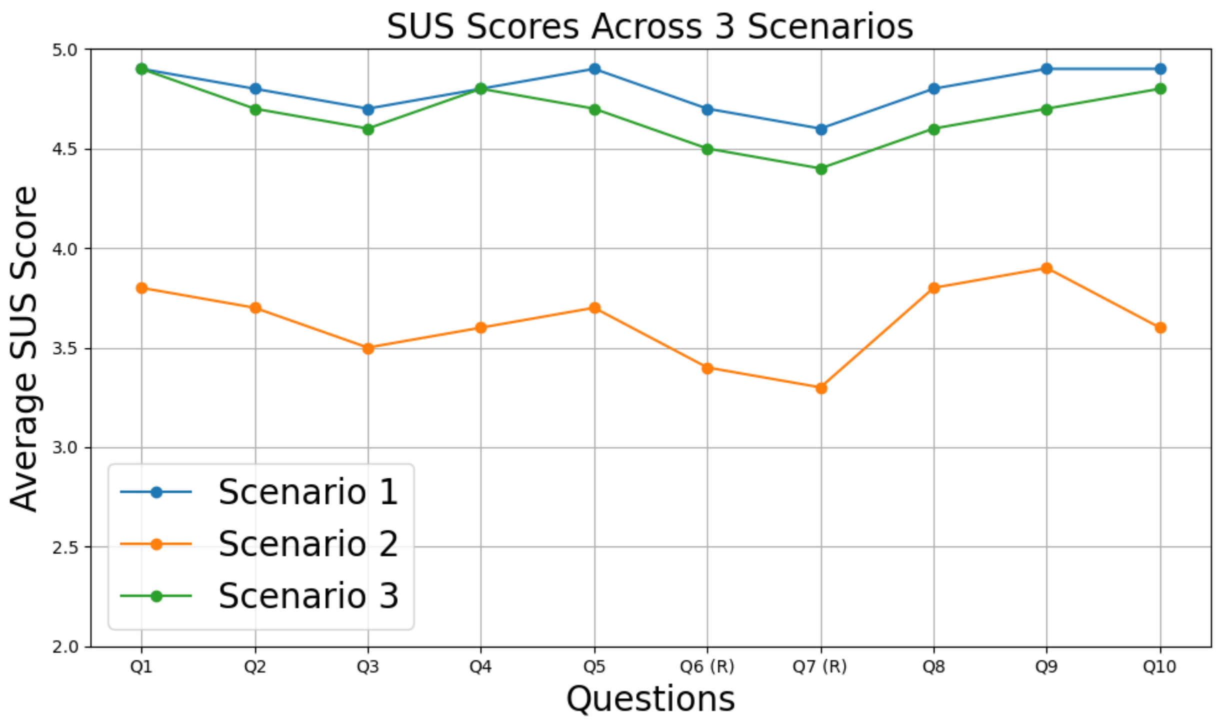


Figure 4.8: Average SUS Score per Question Across three Scenarios

Score interpretation

Scenario-1 demonstrates good usability across all questions, with an overall average close to 70%, indicating the tool was easy to use, and participants found it relatively intuitive and efficient. Ease of generating textures (Q1) and time taken (Q5) were rated the highest, with a score of 4.9. Users reported high confidence (Q2) and were satisfied with the realism of the textures (Q9).Overall, Scenario-1 reflects a usable tool with few issues that require attention. Scenario-2 had a lower average SUS score of 55%, reflecting average usability with several usability issues. Scenario 2 highlights the need for significant improvements, particularly in instructions and tool complexity, as reflected in the clarity of instructions (Q3) score of 3.5.Similar to Scenario 1, Scenario 3 also had an average SUS score of 70%, indicating good usability but with some room for improvement. Ease of use (Q1), time taken (Q5), and confidence (Q2) were rated relatively high, around 4.9-4.7. Users found the tool responsive (Q8) and appreciated the realism of the textures (Q9) generated in this scenario. However, The complexity of the tool (Q6, reverse scored) and need for assistance (Q7, reverse scored) were slightly lower compared to Scenario 1, reflecting some additional complexity in using custom prompts or uploading images.

Based on the score interpretation, it could be concluded that DefectTwin can be useful when deployed as a good tool. However, the complexity of interface may reduce the user satisfaction, therefore the design of the multimodal interface need to be considered and further heuristics can be designed for standard design of the interface of LLM-enabled DefectTwin tools.

Chapter 5

Conclusion and Future Work

This chapter concludes the research by summarizing the key findings, discussing the limitations of the study, and proposing directions for future research. It underscores the contributions made to the field of AI-driven DT systems for visual defect inspection in railway PHM, while also reflecting on the broader implications and potential enhancements for this work.

5.1 Summary of Contribution

This thesis aims to design and develop DefectTwin, an innovative LLM-based Digital Twin ecosystem for visual defect inspection in Railway PHM. DefectTwin addresses the limitations of traditional AI approaches by integrating LLMs to enhance data utilization and improve the accuracy of defect detection. This research addressed critical challenges in existing approaches, including data scarcity, multimodal data integration, predictive accuracy, and operational efficiency, by leveraging the capabilities of generative models, multimodal AI, and QoE-driven adaptation mechanisms. The proposed system demonstrated its effectiveness in tackling persistent problems in railway maintenance, achieving high precision, adaptability, and usability.

At the heart of the research lies the synthetic data generation pipeline, which mitigates the challenge of insufficient data by creating diverse, realistic defect datasets. This pipeline enabled the fine-tuning of an M^2 LLM component, which seamlessly integrates multimodal inputs such as text, images, and video to improve data fusion and prediction accuracy. The incorporation of a QoE feedback loop further allowed continuous performance refinement based on real-world user feedback, enhancing both the system's adaptability and user satisfaction.

Our experimental results demonstrate that DefectTwin outperforms existing models, achieving high precision for both image and video data. Notably, DefectTwin exhibits better zero-shot performance for unseen examples, highlighting its adaptability to new scenarios. Furthermore, the system demonstrates low latency for image and text generation, crucial for real-world deployment. User evaluations confirm DefectTwin's usability, with a decent SUS score at acceptance range.

5.2 Limitations

Although this thesis address crucial gaps in advancing PHM railway by addressing persistent challenges, there are few remaining challenges still to be addressed. We discuss the following limitations and possible extensions of this research based on our findings.

1. **Computational and infrastructural overhead:** The integration of LLM components significantly increases computational requirements. This should be taken into

consideration for resource-constrained environments. For example, edge computing devices or real-time systems, where maintaining efficiency while ensuring high prediction accuracy can be challenging. Additionally, the dependence on cloud-based processing introduces concerns related to operational costs, data privacy, and security. Although the study analyzed token usage and latency, a complete evaluation of resource efficiency requires deployment in a real-world setting over a connected railway network. Analyzing computational load, communication overhead, and energy consumption in a practical scenario is essential for optimizing resource utilization.

Beyond computational demands, infrastructural overhead also plays a critical role. Many existing railway monitoring systems rely on outdated or legacy infrastructure that lacks the necessary computational capacity to support LLM-driven inferencing. These older systems often have limited processing power, insufficient memory, and incompatible communication protocols, making integration with modern AI pipelines inefficient.

2. **Deployment related issue:** In this study, we utilized static datasets, such as images and pre-recorded videos, to evaluate the performance and effectiveness of the proposed DefectTwin ecosystem. While this approach allowed for controlled experimentation and consistent validation of the system's capabilities, it does not fully reflect the dynamic nature of real-world railway operations. Further research is needed to explore the integration of real-time streaming data from diverse sources within the DefectTwin ecosystem, addressing challenges related to data synchronization, processing speed, and system latency.

3. **Improving user satisfaction and reliability:** The usability study of the deployed tool-TextureMeDefect [31], revealed challenges related to interface complexity, par-

ticularly in scenarios involving custom prompts and image uploads. Refining the multimodal interface design and incorporating user-friendly features is essential to enhance overall usability and user satisfaction. Additionally, the generated outputs may not always be equally helpful for addressing maintenance-related queries, as they require more contextually grounded responses based on the specific policies and guidelines of defect inspection for various railway assets. These guidelines often differ across countries, devices, and organizational standards. Furthermore, access to such documents is frequently restricted, adding another layer of complexity to ensuring the tool's effectiveness in real-world applications.

4. **Inherited issues of LLMs:** While LLMs provide significant advancements in defect detection and railway PHM, they inherit certain fundamental limitations. One of them is domain-specific hallucination. LLMs fine-tuned on general defect datasets may generate hallucinated outputs when applied to highly specialized domains like railway defect analysis. This can lead to incorrect or overly confident predictions. Another issue is that the field of LLMs is continuously evolving, with new versions frequently improving on prior limitations. This introduces a challenge in maintaining system consistency and adaptability over time.

5.3 Future Direction

In this section we present possible improvements to the remaining challenges. The following improvements are recommended to be applied in further extension of this thesis. Also, we illustrate the examples of few ongoing improvements from our future projects.

5.3.1 Improvements to the Resource-efficiency

Deploying LLMs on hybrid platforms can effectively address resource constraints by optimizing performance across diverse environments. For example Groq's Language Processing Unit (LPU) offers a compelling solution in this context [95]. LPU is designed for AI inference tasks to deliver exceptional compute speed and energy efficiency, achieving up to 10 times better performance per watt compared to traditional GPUs [68]. This setup allows organizations to leverage Groq's LPU technology on-site, providing enhanced AI inference capabilities within their own facilities. Additionally, Groq offers the GroqCloud™ service, enabling access to LPU technology via the cloud. This flexibility allows enterprises to choose between cloud-based and on-premises deployments based on their specific needs and preferences [3]. Developing lightweight variants of the Multimodal LLM components can be adopted for edge computing devices [94].

Additionally, techniques like model pruning, quantization, and knowledge distillation can reduce computational demands while maintaining prediction accuracy [77]. However, deploying the DefectTwin system on a connected railway network to analyze real-world computational loads, communication overhead, and energy consumption is required to address the resource-constraint issue more comprehensively.

5.3.2 Improvements to the Deployment

Enabling data preprocessing at the edge to reduce latency and the volume of data transmitted to central servers. This approach can improve system responsiveness in dynamic railway environments. Pilot deployments can be conducted to evaluate the system's ability to handle live streaming data in diverse railway scenarios [16]. Improvement of deployment issues also required to consider data variability, incomplete streams, and hardware constraints to refine real-time capabilities. Combining edge computing for local data pre-processing with cloud-based processing for complex inference tasks can be beneficial for both resource constraints and deployment related issues. Therefore, we recommend a hybrid model which can balance computational load and reduce communication overhead while addressing concerns about latency and energy consumption.

5.3.3 Improvements to the Reliability and Usability

To improve the reliability of the deployed tools require developing mechanisms to dynamically integrate and reference country-specific and device-specific maintenance guidelines.

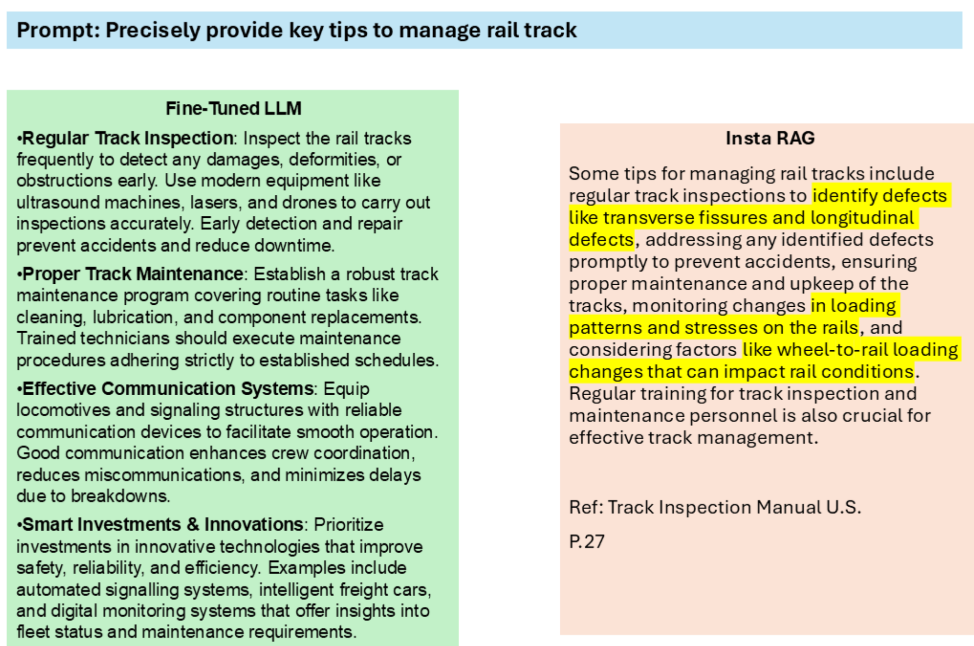


Figure 5.1: Response by Fine-Tuned LLM and Insta-RAG based Chatbot From Ongoing Work

We recommend incorporating Multi-modal-RAG [67] [30] along with the fine-tuned LLMs to enables the system. This hybrid models can provide responses grounded in specific policies and guidelines pertinent to various railway assets. From our ongoing work in this regard, we illustrate an example in Fig. 5.1. It can be clearly observed from the output that RAG-based approach offers relevant, explainable response. For example, the Insta RAG response, clearly highlights specific terms and concepts like "transverse fissures," "longitudinal defects," and "loading patterns."

While the fine-tuned LLM provides a broad overview of key practices in rail track management. Additionally, the explicit reference to the source (e.g., "Ref: Track Inspection Manual U.S.") enhances explainability, as users can directly access the original documents for verification, thereby improving trust in the response.

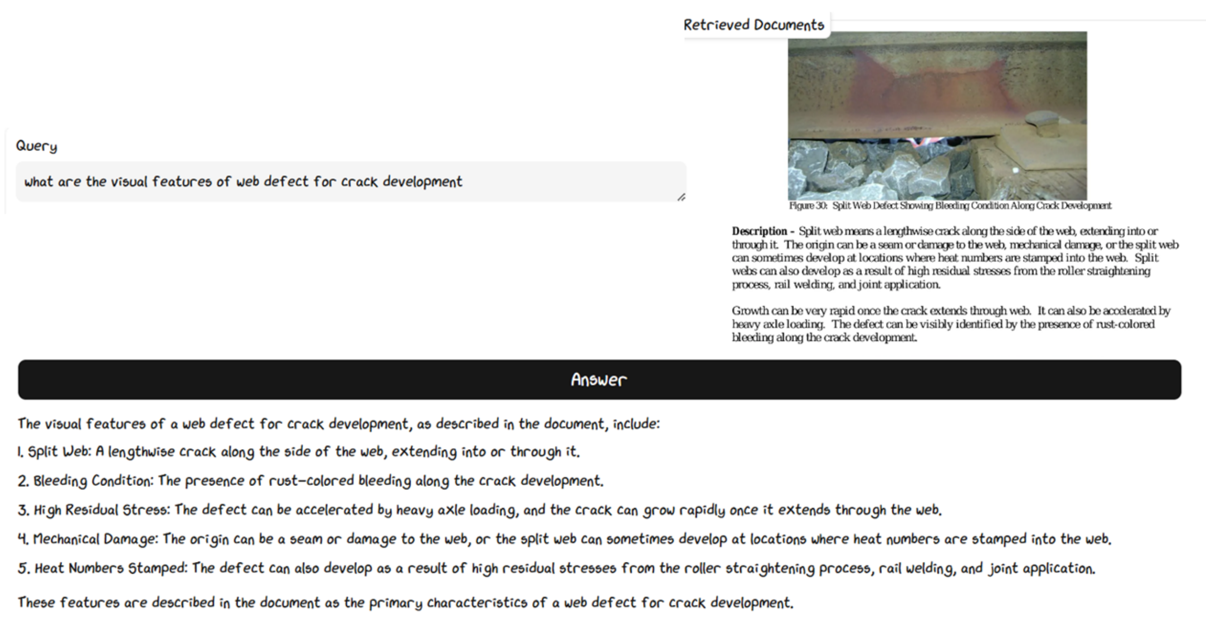


Figure 5.2: Response by Multimodal RAG Tool From Ongoing Work

Also the output from our multimodal RAG-based tool in Fig. 5.2, showcases a query about the visual features of a web defect for crack development, processed using a retrieval-based method. The retrieved documents provide specific visual evidence (e.g., images of split web defects and descriptions of rust-colored bleeding, heavy axle loading impacts, and residual stresses) and a corresponding text-based explanation. By combining image and text retrieval, the tool provides comprehensive responses, improving understanding for defect inspection scenarios.

Maintenance personnel can query the system for both visual illustrations and textual guidelines, aiding decision-making. In our future work, we aim to extend approach can extend to various industries where multimodal information is vital (e.g., transportation, manufacturing).

While accessing and integrating relevant documents, the system can deliver more accurate and trustworthy information, collaboration with railway organizations to access restricted documents is essential to enhance response relevance without compromising privacy.

5.3.4 Improvements to the LLM Posed Issues

The AI-inferencing engine of DefectTwin is designed to aggregate LLM components M^2 LLM according to the user's need. Therefore adapting newly evolved LLM for AI-inferencing is seemless. However, to offer more dynamic control, customizable system instructions can help steer model behavior according to domain-specific constraints. By dynamically adjusting prompts, safety filters, or response structures, domain experts can guide the model toward more relevant outputs while avoiding misleading or hallucinated information. Additionally, instead of requiring manual reconfiguration for every new LLM version, a self-adaptive prompt engineering module can be developed. This module would analyze model behavior and dynamically adjust query structures to maintain consistent and accurate outputs.

Synthetic data is on its way to addressing data scarcity and fueling precision toward reliability in industrial applications.

References

[1] Asad Abbas. Ai for predictive maintenance in industrial systems. *International Journal of Advanced Engineering Technologies and Innovations*, 1(1):31–51, 2024.

[2] Sayed Abdel-Khalek, Abeer D. Algarni, Ghada Amoudi, Salem Alkhalaf, Fahad Mohammed Alhomayani, and Shankar Kathiresan. Leveraging ai-generated content for synthetic electronic health record generation with deep learning-based diagnosis model. *IEEE Transactions on Consumer Electronics*, pages 1–1, 2024.

[3] Dennis Abts, Garrin Kimmell, Andrew Ling, John Kim, Matt Boyd, Andrew Bitar, Sahil Parmar, Ibrahim Ahmed, Roberto DiCecco, David Han, John Thompson, Michael Bye, Jennifer Hwang, Jeremy Fowers, Peter Lillian, Ashwin Murthy, Elyas Mehtabuddin, Chetan Tekur, Thomas Sohmers, Kris Kang, Stephen Maresh, and Jonathan Ross. The groq software-defined scale-out tensor streaming multiprocessor: From chips-to-systems architectural overview. In *Proceedings of the IEEE 49th Annual International Symposium on Computer Architecture (ISCA)*, pages 1029–1041, 2022.

[4] Sajad Saraygord Afshari, Fatemeh Enayatollahi, Xiangyang Xu, and Xihui Liang. Machine learning-based methods in structural reliability analysis: A review. *Reliability Engineering & System Safety*, 219:108223, 2022.

[5] Sanjar Ahmad, Maksym Spiryagin, Qing Wu, Esteban Bernal, Yan Sun, Colin Cole, and Bruce Makin. Development of a digital twin for prediction of rail surface damage in heavy haul railway operations. *Vehicle System Dynamics*, 62(10):1–22, 2023.

[6] Wasim Ahmad. *Artificial Intelligence-Based Condition Monitoring of Rail Infrastructure*. PhD thesis, University of Twente, 2019.

[7] Moayad Aloqaily, Ouns Bouachir, Fakhri Karray, Ismaeel Al Ridhawi, and Abdulmotaleb El Saddik. Integrating digital twin and advanced intelligent technologies to realize the metaverse. *IEEE Consumer Electronics Magazine*, 12(6):47–55, 2022.

[8] Ivan Arakistain, David García, Diego Zamora, Alberto Armijo, Ana Fernández-Navamuel, Jose Carlos Jimenez, and Unai Beristain. Predictive-cognitive maintenance for advanced integrated railway management. In *Proceedings of the 10th European Workshop on Structural Health Monitoring (EWSHM 2024)*, pages 1–8, 2024.

[9] Hossein Davari Ardakani, Diego Robles, Hervé Borrion, and Clive Roberts. Phm for railway system—a case study on the health assessment of the point machines. In *Annual Conference of the PHM Society*, volume 10, 2018.

[10] Alberto Armijo and Diego Zamora-Sánchez. Integration of railway bridge structural health monitoring into the internet of things with a digital twin: A case study. *Sensors*, 24(7):2115, 2024.

[11] Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen-tau Yih. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*, 2024.

[12] Esteban Bernal Arango, Qing Wu, Maksym Spiryagin, and Colin Cole. Augmented digital twin for railway systems. *Vehicle System Dynamics*, 62(1):67–83, 2024.

[13] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[14] Luke Brown, Shukri Afazov, and Daniele Scrimieri. Towards autonomous health monitoring of rails using a fea-ann based approach. In *Advances in Computational Intelligence Systems*, pages 569–576. Springer, 2021.

[15] Cristina Castejón, Enrique Soriano-Heras, Juan Carlos García-Prada, and Alejandro Bustos. On the development of a diagnostic system for condition based maintenance of passenger trains. In *2022 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–6. IEEE, 2022.

[16] Chenglin Chen, Fei Wang, Min Yang, Yong Qin, and Yun Bai. Edge-enabled real-time railway track segmentation. *arXiv preprint arXiv:2401.11492*, 2024.

[17] Chong Chen, Huibin Fu, Yu Zheng, Fei Tao, and Ying Liu. The advance of digital twin for predictive maintenance: The role and function of machine learning. *Journal of Manufacturing Systems*, 71:581–594, 2023.

[18] Qianyu Chen, Gemma Nicholson, Clive Roberts, Jiaqi Ye, and Yihong Zhao. Improved fault diagnosis of railway switch system using energy-based thresholding wavelets (ebtw) and neural networks. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021.

[19] Hyunseok Chung, Sunyoung Hyun, and Young-Guk Ha. Battlefield situation awareness using pretrained generative llm. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 397–398, 2024.

[20] Alice Consilvio, José Solís-Hernández, Noemi Jiménez-Redondo, Paolo Sanetti, Federico Papa, and Iñigo Mingolarra-Garaizar. On applying machine learning and simulative approaches to railway asset management: The earthworks and track circuits case studies. *Sustainability*, 12(6):2544, 2020.

[21] Adolfo Crespo del Castillo, Marco Macchi, and Laura Cattaneo. Digital twin-driven condition-based maintenance. In *Research Anthology on BIM and Digital Twins in Smart Cities*, pages 421–449. IGI Global, 2023.

[22] Pierre Dersin, Allegra Alessi, Olga Fink, Benjamin Lamoureux, and Mehdi Brahimi. Prognostics and health management in railways. In *Handbook of RAMS in Railway Systems*, pages 83–102. CRC Press, 2018.

[23] Ma Didi, Hu Lili, and Liu Kai. Safety evaluation of high speed railway lte-r communication system based on ahp and fuzzy comprehensive evaluation. In *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pages 211–214. IEEE, 2019.

[24] Ruth Dirnfeld. Digital twins in railways: State of the art, opportunities, and guidelines. 2022.

[25] John Doe and Jane Smith. Revolutionizing railways: An ai-powered approach for enhanced monitoring and optimization. *International Journal of Railway Technology*, 10(2):123–145, 2024.

[26] Abdulmotaleb El Saddik and Sara Ghaboura. The integration of chatgpt with the metaverse for medical consultations. *IEEE Consumer Electronics Magazine*, 2023.

[27] Abdulmotaleb El Saddik, Fedwa Laamarti, and Mohammad Alja'Afreh. The potential of digital twins. *IEEE Instrumentation & Measurement Magazine*, 24(3):36–41, 2021.

[28] Itxaro Errandonea, Jon Goya, Unai Alvarado, Sergio Beltrán, and Saioa Arrizabalaga. Iot approach for intelligent data acquisition for enabling digital twins in the railway sector. In *2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC)*, pages 164–168. IEEE, 2021.

[29] Cordelia M. Ezhilarasu, Zakwan Skaf, and Ian K. Jennions. Understanding the role of a digital twin in integrated vehicle health management (ivhm). In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 1053–1060. IEEE, 2019.

[30] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2024.

[31] Rahatara Ferdousi, M. Anwar Hossain, and Abdulmotaleb El Saddik. Texturemedefect: Llm-based defect texture generation for railway components on mobile devices. *arXiv preprint arXiv:2410.18085*, 2024.

[32] Rahatara Ferdousi, Fedwa Laamarti, M Anwar Hossain, Chunsheng Yang, and Abdulmotaleb El Saddik. Digital twins for well-being: an overview. *Digital Twin*, 1(7):7, 2022.

[33] Rahatara Ferdousi, Fedwa Laamarti, Chunsheng Yang, and Abdulmotaleb El Saddik. Railtwin: a digital twin framework for railway. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 1767–1772. IEEE, 2022.

[34] Rahatara Ferdousi, Fedwa Laamarti, Chunsheng Yang, and Abdulmotaleb El Saddik. A reusable ai-enabled defect detection system for railway using ensembled cnn. *arXiv preprint arXiv:2311.14824*, 2023.

[35] Rahatara Ferdousi, Chunsheng Yang, M Anwar Hossain, Fedwa Laamarti, M Shamim Hossain, and Abdulmotaleb El Saddik. Generative model-driven synthetic training image generation: An approach to cognition in railway defect detection. *Cognitive Computation*, pages 1–16, 2024.

[36] Julian Franzen, Jannis Stecken, Raphael Pfaff, and Bernd Kuhlenkötter. Using the digital shadow for a prescriptive optimization of maintenance and operation. In *Advances in Production, Logistics and Traffic*, pages 265–276. Springer, 2019.

[37] Sara Ghaboura, Rahatara Ferdousi, Fedwa Laamarti, Chunsheng Yang, and Abdulmotaleb El Saddik. Digital twin for railway: A comprehensive survey. *IEEE Access*, 11:120237–120257, 2023.

[38] Mikhail Groshev, Cláudio Guimarães, Jorge Martín-Pérez, and Antonio de la Oliva. Toward intelligent cyber-physical systems: Digital twin meets artificial intelligence. *IEEE Communications Magazine*, 59(8):14–20, 2021.

[39] Himanshu Gupta and Pradeep Kundu. Digital twin development for feed drive systems condition monitoring and maintenance planning. In *Proceedings of the European Conference of the PHM Society 2024 Doctoral Symposium*, page 4, 2024.

[40] Saeed H-Nia, Jesper Flodin, Carlos Casanueva, Mathias Asplund, and Sebastian Stichel. Predictive maintenance in railway systems: Mbs-based wheel and rail life prediction exemplified for the swedish iron-ore line. *Vehicle System Dynamics*, pages 1–18, 2022.

[41] M. Hicham, J. Smith, and A. Doe. Toward an intelligent diagnosis and prognostic health management system for autonomous electric vehicle powertrains: A novel distributed intelligent digital twin-based architecture. *IEEE Access*, 12:12345–12356, 2024.

[42] Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, and Shijian Lu. Visual instruction tuning towards general-purpose multimodal model: A survey. *arXiv*, 2023.

[43] Sakdirat Kaewunruen, Mohannad AbdelHadi, Manwika Kongpuang, Withit Pansuk, and Alex M. Remennikov. Digital twins for managing railway bridge maintenance, resilience, and climate change adaptation. *Sensors*, 23(1):252, 2023.

[44] Sakdirat Kaewunruen, Abdullah Abimbola Adesope, Junhui Huang, Ruilin You, and Dan Li. Ai-based technology to prognose and diagnose complex crack characteristics of railway concrete sleepers. *Discover Applied Sciences*, 6(5):1–16, 2024.

[45] Sakdirat Kaewunruen and Qiang Lian. Digital twin aided sustainability-based lifecycle management for railway turnout systems. *Journal of Cleaner Production*, 228:1537–1551, 2019.

[46] Uday Kumar and Diego Galar. Transformative maintenance technologies and business solutions for the railway assets. In *Handbook of Advanced Performability Engineering*, pages 565–595. Springer, 2021.

[47] Vimal Kumar, Priyam Srivastava, Ashay Dwivedi, Ishan Budhiraja, Debjani Ghosh, Vikas Goyal, and Ruchika Arora. Large-language-models (llm)-based ai chatbots: Architecture, in-depth analysis and their performance evaluation. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 237–249. Springer, 2023.

[48] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[49] Jianbo Li and Hongmei Shi. Rail corrugation diagnosis of high-speed railway based on dynamic responses of the vehicle. In *2020 Prognostics and Health Management Conference (PHM-Besançon)*, pages 148–152, 2020.

[50] Wei Li, Yifan Zhang, Yifan Wang, Yifan Chen, and Yifan Liu. Railway equipment health management system based on fault prediction technology. *Journal of Rail Transport Planning & Management*, 9(2):100–110, 2019.

[51] Wei Li, Yifan Zhang, Yifan Wang, Yifan Chen, and Yifan Liu. Application exploration of digital twin in rail transit health management. In *2022 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–6. IEEE, 2022.

[52] Wei Li, Yifan Zhang, Yifan Wang, Yifan Chen, and Yifan Liu. Cyber-physical intelligent transport system based on digital-twin technology. *IEEE Transactions on Intelligent Transportation Systems*, 23(5):5246–5255, 2022.

[53] Wei Li, Yifan Zhang, Yifan Wang, Yifan Chen, and Yifan Liu. Digital twin enabled industry 4.0 predictive maintenance under reliability-centred strategy. In *2022 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–6. IEEE, 2022.

[54] Wei Li, Yifan Zhang, Yifan Wang, Yifan Chen, and Yifan Liu. Machine-learning-driven digital twin for lifecycle management of complex equipment. *IEEE Transactions on Industrial Informatics*, 18(8):5246–5255, 2022.

[55] Yan-Fu Li, Huan Wang, and Muxia Sun. Chatgpt-like large-scale foundation models for prognostics and health management: A survey and roadmaps. *Reliability Engineering & System Safety*, page 109850, 2023.

[56] Chang Liu, Baoqing Zheng, Ye Li, and Xianpeng Yang. Digital twin-based detection of rail surface defects using machine learning algorithms. *Transportation Research Part C: Emerging Technologies*, 118:102745, 2020.

[57] Donghui Liu, Kan Liu, Xiaoming Xu, Yunheng Cai, and Quanyu Wang. Research and application on prognostics and health management technology for high-speed train traction system. In *Proceedings of the 6th International Conference on Electrical Engineering and Information Technologies for Rail Transportation (EITRT) 2023*, pages 316–321. Springer, 2024.

[58] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[59] Wei Liu, Xiaolong Xu, Lianyong Qi, Xiaokang Zhou, Hanzhi Yan, Xiaoyu Xia, and Wanchun Dou. Digital twin-assisted edge service caching for consumer electronics manufacturing. *IEEE Transactions on Consumer Electronics*, 2024.

[60] Zongchang Liu, Chao Jin, Wenjing Jin, Jay Lee, Zhiqiang Zhang, Chang Peng, and Guanji Xu. Industrial ai enabled prognostics for high-speed railway systems. In *2018 IEEE international conference on prognostics and health management (ICPHM)*, pages 1–8. IEEE, 2018.

[61] Christian Ludwig, Eleni Tsouchnika, and Ulrich Wever. Digital twins for large electric drive trains. In *2018 Petroleum and Chemical Industry Conference Europe (PCIC Europe)*, pages 1–7. IEEE, 2018.

[62] Sarah Lukens and Asma Ali. Evaluating the performance of chatgpt in the automation of maintenance recommendations for prognostics and health management. In *Annual Conference of the PHM Society*, volume 15, 2023.

[63] Stefan Marx, Ilya Zaidman, Holger Naraniecki, and Alper Lazoglu. Concept for a digital twin of railway bridges on the example of the new filstal bridges. *ce/papers*, 5(4):1005–1012, 2023.

[64] MathWorks. Data preprocessing for condition monitoring and predictive maintenance, 2023. Accessed: 2024-11-25.

[65] Paul McMahon, Tieling Zhang, and Richard A. Dwight. Approaches to dealing with missing data in railway asset management. *IEEE Access*, 8:48177–48194, 2020.

[66] Ramin Moradi, Yuheng Zheng, Michael Hutchinson, Michael Roth, Kanwal Jahan, Jon Goya, and Unai Alvarado. Positioning for train-infrastructure asset health status monitoring within the sia-project. In *Proceedings of the 33rd International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2020)*, pages 2948–2959, 2020.

[67] Ahmad Rageh, Saeed Eftekhar Azam, Qusai Alomari, Daniel Linzell, and Richard Wood. Model updating and parameter identification for developing digital twins for riveted steel railway bridges. In *Recent Developments in Structural Health Monitoring and Assessment–Opportunities and Challenges: Bridges, Buildings and Other Infrastructures*, pages 285–318. World Scientific, 2022.

[68] Santosh Raghavan, Igor Arsovski, and Dinesh Maheshwari. The groq® lpu™ inference engine: 10x more energy efficient than gpus. here's why. Technical report, Groq, Inc., 2024.

[69] Yingjia Shang, Zhijun Liu, Jiawen Kang, M. Shamim Hossain, and Yi Wu. Adversarial attacks on vision-language model-empowered chatbots in consumer electronics. *IEEE Transactions on Consumer Electronics*, pages 1–1, 2024.

[70] Ruiyuan Shen, Tiantian Wang, and Qizhang Luo. Towards prognostic and health management of train wheels in the chinese railway industry. In *2019 Prognostics and System Health Management Conference (PHM-Qingdao)*, pages 1–6. IEEE, 2019.

[71] Yanyan Shen, Liying Ma, and Khashayar Khorasani. Hybrid machine learning-based methodologies for fault diagnosis, prognosis and health management. In *2024 36th Chinese Control and Decision Conference (CCDC)*, pages 4318–4325. IEEE, 2024.

[72] Minoru Shimizu, Suresh Perinpanayagam, Bernadin Namoano, and Andrew Starr. Real-time prognostics and health management without run-to-failure data on railway assets. *IEEE Access*, 11:28724–28734, 2023.

[73] Gauransh Singh, Praveen Kumar, Rishabh Kumar Mishra, Sanskriti Sharma, and Kushagra Singh. Security system for railway crossings using machine learning. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 135–139. IEEE, 2020.

[74] John Smith, Emily Johnson, and Michael Brown. Towards digital twin trains: Implementing a cloud-based framework for railway vehicle dynamics simulation. *International Journal of Rail Transportation*, 12(3):245–260, 2024.

[75] Jessada Sresakoolchai and Sakdirat Kaewunruen. Track geometry prediction using three-dimensional recurrent neural network-based models cross-functionally co-simulated with bim. *Sensors*, 23(1):391, 2022.

[76] Jessada Sresakoolchai and Sakdirat Kaewunruen. Railway infrastructure maintenance efficiency improvement using deep reinforcement learning integrated with digital twin based on track geometry and component defects. *Scientific Reports*, 13(1):2439, 2023.

[77] Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. A simple and effective pruning approach for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[78] Laifa Tao, Shangyu Li, Haifei Liu, Qixuan Huang, Liang Ma, Guoao Ning, Yiling Chen, Yunlong Wu, Bin Li, Weiwei Zhang, Zhengduo Zhao, Wenchao Zhan, Wenyan Cao, Chao Wang, Hongmei Liu, Jian Ma, Mingliang Suo, Yujie Cheng, Yu Ding, Dengwei Song, and Chen Lu. An outline of prognostics and health management large model: Concepts, paradigms, and challenges. *arXiv preprint arXiv:2407.03374*, 2024.

[79] Pittawat Taveekitworachai, Mustafa Can Gursesli, Febri Abdullah, Siyuan Chen, Federico Cala, Andrea Guazzini, Antonio Lanata, and Ruck Thawonmas. Journey of chatgpt from prompts to stories in games: the positive, the negative, and the neutral. In *2023 IEEE 13th International Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, pages 202–203, 2023.

[80] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[81] Adithya Thaduri, Ajit Kumar Verma, and Uday Kumar. Maintenance of railway infrastructure using cyber-physical systems. *Decision Analytics Applications in Industry*, pages 521–540, 2020.

[82] Alexandre Trilla and Xavier Cabré. Determining the equivalent conicity for railway wheelset maintenance with deep ensembles. In *Annual Conference of the PHM Society*, volume 10, 2018.

[83] Perin Unal, Özlem Albayrak, Moez Jomâa, and Arne J. Berre. Data-driven artificial intelligence and predictive analytics for the maintenance of industrial machinery with hybrid and cognitive digital twins. In *Technologies and Applications for Big Data Value*, pages 299–319. Springer, 2022.

[84] Huan Wang and Yan-Fu Li. Large-scale language models for phm in railway systems-potential applications, limitations, and solutions. In *International Conference on Electrical and Information Technologies for Rail Transportation*, pages 591–599. Springer, 2023.

[85] Huan Wang, Yan-Fu Li, and Min Xie. Empowering chatgpt-like large-scale language models with local knowledge base for industrial prognostics and health management. *arXiv preprint arXiv:2312.14945*, 2023.

[86] Qi Wang, Siqi Bu, and Zhengyou He. Achieving predictive and proactive maintenance for high-speed railway power equipment with lstm-rnn. *IEEE Transactions on Industrial Informatics*, 16(10):6509–6517, 2020.

[87] Daniel N. Wilke, Daniel Fourie, and Petrus Johannes Gräbe. Towards a railway infrastructure digital twin framework for african railway lifecycle management. In *International Congress and Workshop on Industrial AI and eMaintenance 2023*, pages 101–113. Springer, 2024.

[88] Xingtang Wu, Wenbo Lian, Min Zhou, Haifeng Song, and Hairong Dong. A digital twin-based fault diagnosis framework for bogies of high-speed trains. *IEEE Journal of Radio Frequency Identification*, 7:203–207, 2023.

[89] Xingtang Wu, Wenbo Lian, Min Zhou, Haifeng Song, and Hairong Dong. A digital twin-based fault diagnosis framework for bogies of high-speed trains. *IEEE Journal of Radio Frequency Identification*, 7:203–207, 2023.

[90] Wang Xiaodong, Liu Feng, Ren Junhua, and Liang Rongyu. A survey of digital twin technology for phm. In *Recent Trends in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2018*, pages 397–403. Springer, 2020.

[91] Jianpeng Xu and Bo Ai. Artificial intelligence empowered power allocation for smart railway. *IEEE Communications Magazine*, 59(2):28–33, 2021.

[92] Chunsheng Yang, Rahatara Ferdousi, Abdulmotaleb El Saddik, Yifeng Li, Zheng Liu, and Min Liao. Lifetime learning-enabled modelling framework for digital twin. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 1761–1766, 2022.

[93] Chunsheng Yang, Rahatara Ferdousi, Abdulmotaleb El Saddik, Yifeng Li, Zheng Liu, and Min Liao. Lifetime learning-enabled modelling framework for digital twin. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 1761–1766. IEEE, 2022.

[94] Yizhe Yang, Huashan Sun, Jiawei Li, Runheng Liu, Yinghao Li, Yuhang Liu, Heyan Huang, and Yang Gao. Mindllm: Pre-training lightweight large language model from scratch, evaluations and domain applications. *arXiv preprint arXiv:2310.15777*, 2023.

[95] Zhi Yang, Krishna Mellachervu, Igor Arsovski, Clint Harames, and Jim Miller. Optimized simulation methodology of warpage and localized stress hotspot prediction for assembly risk assessment. In *Proceedings of the IEEE 74th Electronic Components and Technology Conference (ECTC)*, pages 1234–1240, 2024.

[96] Jian Yao, Wei Bai, Guoyuan Yang, Zhikang Meng, and Kaixuan Su. Assessment and prediction of railway station equipment health status based on graph neural network. *Frontiers in Physics*, 10:1080972, 2022.

[97] Shiyao Zhang, Hairong Dong, Ulrich Maschek, and Haifeng Song. A digital-twin-assisted fault diagnosis of railway point machine. In *2021 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 430–433. IEEE, 2021.

[98] Di Zhu, Wei Shao, Jing Li, Zijian Zhou, Siyuan Han, and Yunpeng Feng. Fault diagnosis of railway rolling bearings based on digital twin and transfer learning. *Journal of Rail Transport Planning & Management*, 11:100277, 2021.