

Article

RAG-Enhanced Vision–Language Framework and Dataset for Railway Signal Cognition and Safety Reasoning

Qunbo Wang ¹, Shiyi Xiong ¹, Jiawei Li ¹, Weiliang Li ^{1,*}, Chu Huang ^{1,*}, Sen Zhang ¹, Xize Guo ², Chao Fan ² and Wenjun Wu ³

¹ School of Automation and Intelligence, Beijing Jiaotong University, Beijing 100044, China; 10311@bjtu.edu.cn (Q.W.); syxiong225@outlook.com (S.X.); lijwmaster@163.com (J.L.); szhang02004@outlook.com (S.Z.)

² School of Software, Northwestern Polytechnical University, Xi'an 710072, China; fanchao_email@mail.nwpu.edu.cn (C.F.)

³ School of Artificial Intelligence, Beihang University, Beijing 100191, China; wwj09315@buaa.edu.cn

* Correspondence: weiliang_li04@163.com (W.L.); huang_chu7922@outlook.com (C.H.)

Abstract

Railway scene understanding is critical for ensuring train operational safety and advancing intelligent railway systems. Existing railway vision methods mainly focus on perception and classification, while lacking regulation-guided semantic reasoning capabilities in complex environments. To address these limitations, this paper proposes a retrieval-augmented generation (RAG)-enhanced vision–language framework for railway signal cognition and safety reasoning. The proposed method integrates railway signal perception, regulatory knowledge retrieval, and multi-modal reasoning to improve factual consistency, reasoning reliability, and operational interpretability. In addition, a dedicated railway signal dataset comprising 500 standardized railway scene images with structured QA annotations is constructed to support regulation-oriented multi-modal recognition evaluation. Experimental results show that the proposed framework improves reasoning accuracy from 28.40% to 67.20% with an average end-to-end inference latency of 11.31 s per sample, and the inference speed can be further improved by adjusting experimental configurations to trade off between efficiency and accuracy, demonstrating the potential of RAG-enhanced architectures as a foundational step toward reliable multi-modal cognition in intelligent railway systems.

Keywords: railway safety; railway signal; retrieval-augmented generation; vision–language model



Academic Editor: Paolo Bellavista

Received: 31 May 2026

Revised: 19 June 2026

Accepted: 23 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Railway transportation is a safety-critical system in which train operation decisions depend heavily on signal indications, track conditions, and surrounding environments [1,2]. Among various perception modalities, front-view visual information provides the most direct observation of railway operational states, especially railway signal indications, track conditions, and surrounding infrastructure [3]. In practical train driving scenarios, accurate signal cognition is closely related to operational safety and dispatching efficiency [4]. Therefore, reliable railway signal cognition from front-view visual scenes has become a key problem in intelligent railway systems.

It should be noted that railway safety reasoning poses fundamentally different challenges compared to general road transportation, as it demands strict, rule-constrained

decision-making where reasoning results must be deterministically mapped to precise regulatory clauses to ensure compliance with railway safety operational requirements.

Traditional railway visual analysis methods primarily rely on handcrafted features, rule-based systems, or conventional deep learning models [5]. While these approaches can perform acceptably in constrained environments, they still show limited robustness and insufficient semantic understanding under complex railway conditions [6]. To overcome similar robustness limitations, researchers have increasingly adopted domain-guided and data-driven modeling approaches in other physically complex and safety-critical applications, such as constraint-embedded data-driven prediction in transportation infrastructure [7], physically constrained neural network architectures for geophysical inverse problems [8], and deep feature learning for complex spatial pattern recognition in seismic interpretation [9]. Building on this broader trend of domain-aware cognition. Similarly, in safety-critical robotic and estimation systems, knowledge-informed trajectory planning combining expert demonstration encoding with real-time safety constraints [10], and physics-constrained adaptive estimation frameworks with domain dynamics priors [11], have demonstrated the effectiveness of integrating structured domain knowledge with explicit constraint mechanisms to ensure reliable system outputs. Recent advances in computer vision and deep learning have significantly improved railway perception tasks, including object detection, scene segmentation, and signal recognition [12]. However, most existing studies still focus primarily on visual perception and classification tasks [13], with limited attention to high-level semantic reasoning and explicit integration with railway operational regulations. Consequently, these methods are difficult to directly support safety-oriented operational cognition, where visual observations need to be interpreted together with railway regulations.

Recently, large language models (LLMs) and vision–language models (VLMs) have demonstrated strong multi-modal understanding and reasoning capabilities [14–16]. By combining visual perception with semantic reasoning, these models provide new opportunities for intelligent railway scene cognition. Nevertheless, applying these general models to safety-critical railway scenarios still faces several limitations. First, publicly available railway datasets suitable for regulation-oriented signal cognition remain extremely limited [17]. Second, due to their reliance on general-purpose corpora, current large models often lack explicit knowledge of railway operational regulations and signaling rules, which limits their reliability in safety-critical scenarios [14]. Third, existing multi-modal approaches mainly emphasize standard visual–language alignment while neglecting standardized railway signal semantics and regulation-grounded reasoning mechanisms [18,19]. To address similar knowledge gaps, existing studies in broader transportation scenarios have shown that external knowledge augmentation can significantly improve reasoning consistency and contextual understanding [18,20,21]. However, how to effectively integrate structured railway regulations into multi-modal reasoning frameworks remains an open challenge.

To address these issues, this study proposes a retrieval-augmented generation (RAG)-enhanced vision–language framework for railway signal cognition and safety reasoning in train driving scenarios. The proposed framework integrates railway signal perception, regulatory knowledge retrieval, and vision–language reasoning to support signal understanding and operational risk analysis. In particular, this work constructs a dedicated railway signal dataset with standardized signal representations and introduces a regulation-guided retrieval-augmented reasoning mechanism based on the *Railway Technical Management Regulations* [22]. By combining visual signal cognition with domain knowledge retrieval, the proposed framework aims to improve reasoning reliability, factual consistency, and operational interpretability in complex railway environments.

The main contributions of this work are summarized as follows:

- Propose a RAG-enhanced railway signal cognition framework that integrates vision–language reasoning with railway operational regulations, enabling regulation-guided signal understanding and safety reasoning in train driving scenarios.
- Construct a dedicated railway signal dataset with standardized signal representations and structured QA annotations, addressing the lack of publicly available datasets for regulation-oriented railway signal cognition.
- Design a multi-level evaluation benchmark covering retrieval quality, reasoning accuracy, end-to-end latency, per-sample processing time, and system reliability, validating the effectiveness and robustness of the proposed framework in complex railway environments.

2. Related Work

2.1. Machine Vision and Multi-Modal Perception for Railway Systems

Machine vision has been widely used in intelligent transportation systems and railway monitoring tasks, including object detection, infrastructure inspection, behavior analysis, and operational state estimation [23,24]. However, complex railway environments and uncertain operating conditions expose the limitations of traditional single-modal perception frameworks, which has motivated the development of multi-modal collaborative perception methods [25].

Recent studies have shown that integrating heterogeneous sensing modalities, such as cameras, LiDAR, and radar, can substantially improve perception robustness under adverse weather and low-visibility conditions [25,26]. For example, Huang et al. proposed the V2X-R dataset and a multi-modal diffusion-denoising framework to maintain stable perception performance in foggy and snowy environments, demonstrating the potential of multi-modal fusion and generative modeling for all-weather transportation perception [25].

Beyond spatial perception, temporal modeling techniques have also been introduced to capture dynamic operational states from sequential observations [27,28]. Samani et al. employed BiLSTM networks to estimate railway infrastructure conditions from train vibration responses, showing the feasibility of deep temporal models for railway structural health monitoring and operational safety analysis [27].

Beyond railway-specific methods, recent advances have also addressed robust feature learning under varying scene complexities through domain transform strategies [29], and explored mask-guided multi-modal fusion mechanisms that selectively integrate complementary information across modalities [30], both of which provide methodological references for railway scene cognition under complex operating conditions.

Overall, existing studies indicate that railway perception is gradually evolving from isolated visual sensing toward multi-modal and temporally-aware cognition frameworks with enhanced environmental robustness and operational reliability.

2.2. Large Language Models and Vision–Language Models in Transportation

The emergence of foundation models, including CLIP, SAM, large language models (LLMs), and vision–language models (VLMs), has promoted the development of intelligent transportation systems from conventional closed-set perception toward open-vocabulary understanding, semantic reasoning, and decision-oriented cognition [12,18,31,32]. Existing studies suggest that such models can alleviate data scarcity and limitations in semantic representation, while enabling transportation systems to evolve from deterministic perception pipelines toward multi-modal and reasoning-aware intelligent frameworks [25,33].

Consequently, recent research increasingly emphasizes planning-aware perception and reasoning-guided decision support rather than isolated object detection and scene recognition [34]. In autonomous driving, planning-oriented perception architectures have

been proposed to directly couple scene understanding with downstream decision-making processes [12]. Hu et al. incorporated planning constraints into end-to-end perception and prediction frameworks, enabling perception outputs to directly support motion planning tasks [12]. These studies reflect a broader transition from perception-centered systems toward cognition-oriented intelligent frameworks.

At the multi-modal level, VLMs further integrate visual scene understanding with rule-aware semantic reasoning through language-space representations [18,35]. These models can encode causal relationships, operational logic, and high-level semantic information beyond conventional visual features [36]. Wang et al. proposed the OmniDrive dataset and a dual-path agent framework that aligns visual perception with trajectory-level planning signals, showing the potential of multi-modal reasoning for transportation decision-making [18].

Overall, existing studies suggest that intelligent transportation systems are evolving toward trustworthy multi-modal reasoning frameworks emphasizing interpretability, factual grounding, and operational safety [36]. Nevertheless, effectively integrating multi-modal reasoning with strict railway operational constraints remains an open challenge [37].

While the above studies demonstrate the potential of foundation models for intelligent transportation, it is important to note that autonomous driving primarily focuses on dynamic spatial planning and trajectory-level decision-making, where approximate or probabilistic reasoning is often acceptable. In contrast, railway cognition must prioritize strict, deterministic mapping of visual signal combinations to rigid regulatory codes, where any ambiguity or hallucination in the reasoning chain may directly compromise operational safety. This fundamental difference necessitates a unique, highly constrained reasoning architecture that goes beyond conventional planning-oriented perception frameworks, motivating the regulation-grounded approach proposed in this work.

Nevertheless, while autonomous driving primarily focuses on dynamic spatial planning where approximate reasoning is often acceptable, railway cognition demands strict, deterministic mapping of visual signal combinations to precise regulatory clauses, as any hallucination in the reasoning chain may directly compromise operational safety. This fundamental difference motivates the regulation-grounded reasoning architecture proposed in this work.

2.3. Knowledge-Enhanced Multi-Modal Reasoning for Transportation

In traditional railway safety systems, Automatic Train Protection (ATP) and interlocking logic enforce predefined speed limits and route conflict resolution through hardware-based rules. While these systems provide reliable safety guarantees under normal operating conditions, they operate on rigid rule sets and cannot adapt to novel or ambiguous visual scenarios beyond their predefined scope. This limitation motivates the development of intelligent cognition systems capable of reasoning with flexible domain knowledge to complement traditional safety mechanisms.

Despite their strong multi-modal understanding capabilities, vision-language models (VLMs) still suffer from reliability and hallucination issues in safety-critical transportation scenarios, particularly when interpreting complex situations involving signal abnormalities, track occupancy, and emergency events [21,38]. To improve reasoning robustness and interpretability, recent studies have increasingly integrated retrieval-augmented generation (RAG) and external knowledge into multi-modal reasoning frameworks [39].

In intelligent transportation systems, knowledge-enhanced reasoning has emerged as an effective approach for grounding model decisions in domain regulations and operational rules [40]. For example, SafeDriveRAG incorporates traffic regulations and driving manuals into a retrieval-augmented reasoning pipeline through heterogeneous multi-modal knowl-

edge graphs, significantly improving reasoning reliability in complex driving scenarios [21]. Similarly, SlowFastVAD introduces retrieval-based knowledge augmentation for video anomaly detection by leveraging repositories of normal and abnormal operational patterns, thereby enhancing anomaly discrimination and reducing hallucination-induced false alarms [41,42].

SafeDriveRAG prioritized semantic consistency as its core evaluation focus, employing SEMScore alongside classification accuracy to compose a composite SafeDrive Score, with subgraph retrieval time serving as a supplementary metric for retrieval efficiency. SlowFastVAD centered evaluation on detection accuracy and inference speed as its primary metrics, while relying on qualitative visualization for interpretability assessment. These two contrasting evaluation paradigms, emphasizing generation quality and detection efficiency respectively, motivate the multi-dimensional evaluation benchmark proposed in this work, which jointly assesses retrieval quality, reasoning accuracy, inference efficiency, and system reliability.

In the railway domain, increasing attention has been devoted to constructing structured knowledge repositories that encode operational regulations, maintenance standards, emergency procedures, and safety management rules [43]. Such domain-specific knowledge provides an important foundation for regulation-aware multi-modal reasoning and decision support in complex railway environments [44]. Nevertheless, existing studies primarily focus on knowledge management or retrieval, while the integration of railway knowledge with multi-modal perception and interpretable reasoning remains insufficiently explored. Developing reliable, explainable, and regulation-grounded cognitive frameworks therefore remains a critical challenge for intelligent railway systems [33,36].

3. Methodology

3.1. Problem Analysis and Overall Framework

Train driving safety highly depends on accurate understanding of railway signals, track conditions, and surrounding environments from front-view driving scenes. Existing railway perception methods mainly focus on visual recognition and predefined rule matching, which often lack high-level semantic reasoning and explicit integration with railway operational regulations. As a result, their reliability and interpretability remain limited in complex safety-critical scenarios.

To address these challenges, this study proposes a retrieval-augmented vision-language cognition framework for railway signal understanding and safety reasoning. The framework integrates multi-modal perception with external railway regulatory knowledge to support regulation-grounded operational analysis. As illustrated in Figure 1, the proposed framework consists of four main stages: visual perception, knowledge retrieval, multi-modal reasoning, and decision arbitration. First, a vision-language model extracts semantic information from front-view railway scenes. Relevant operational regulations are then retrieved from an external knowledge base through semantic retrieval. The visual observations and retrieved regulations are jointly incorporated into multi-modal reasoning to infer operational states and safety risks. Finally, a verification and voting mechanism is introduced to improve reasoning consistency and reliability.

The proposed framework establishes an end-to-end pipeline from railway scene perception to interpretable safety decision-making, providing regulation-guided multi-modal cognition for intelligent railway systems.

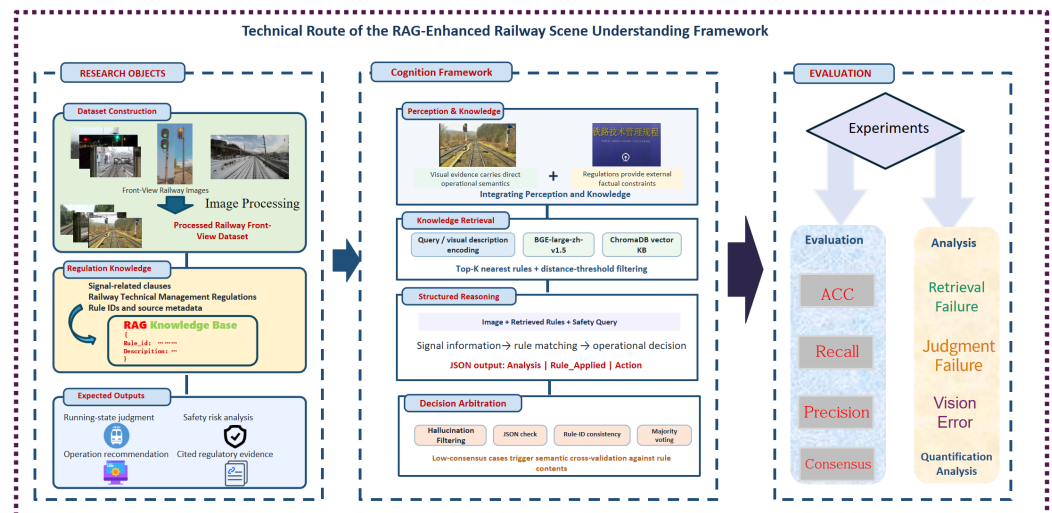


Figure 1. Technical Route of the RAG-Enhanced Railway Safety Cognition Framework.

3.2. Retrieval-Augmented Knowledge Architecture

To provide explicit regulatory knowledge support for railway safety reasoning, this study designs a retrieval-augmented knowledge architecture that integrates railway operational regulations into the multi-modal cognition pipeline. The architecture aims to bridge visual scene understanding with regulation-grounded operational semantics.

3.2.1. Vector Knowledge Base Construction

The *Railway Technical Management Regulations* [22] are adopted as the primary external knowledge source, where signal-related operational rules are selected as the core knowledge domain. During preprocessing, regulation documents are normalized and segmented into semantically independent text chunks while preserving the original rule structure and semantic integrity.

All regulation chunks are encoded into dense semantic embeddings using the BGE-large-zh-v1.5 model and stored in ChromaDB [45] for semantic retrieval. Besides vector representations, each chunk is associated with structured metadata, including rule identifiers and source information, enabling regulation traceability during downstream reasoning. The resulting vector knowledge base supports scalable semantic retrieval and dynamic regulation updates for different railway deployment scenarios.

3.2.2. Semantic Retrieval and Knowledge Injection

Given a railway scene query, the retrieval module performs semantic matching between scene descriptions and regulation embeddings to identify the most relevant operational rules. Retrieved regulations are filtered through similarity constraints to suppress semantically irrelevant knowledge and improve retrieval precision.

The retrieved regulations are subsequently organized into structured prompts and injected into the multi-modal reasoning process as external knowledge constraints. In this way, the reasoning model no longer relies solely on internal parametric knowledge, but instead performs operational inference under explicit regulation guidance. Consequently, railway signal understanding and operational analysis become traceable to external regulatory knowledge, improving factual consistency and interpretability in safety-critical railway scenarios.

3.3. Multi-Path Consensus Reasoning for Railway Scene Understanding

Building upon the retrieved regulatory knowledge, we develop a multi-path consensus reasoning framework for reliable railway safety assessment.

Firstly, visual observations and retrieved regulations are jointly provided to a vision–language model, enabling end-to-end reasoning from scene perception to operational decision-making without introducing information loss caused by intermediate symbolic abstraction. Considering both reasoning capability and deployment efficiency, Qwen2.5-VL-3B-Instruct is adopted as the reasoning backbone.

Each reasoning result is represented in a structured JSON format containing three fields: *Analysis*, which records the reasoning process; *Rule_Applied*, which identifies the referenced regulation; and *Action*, which specifies the inferred operational instruction. This structured representation provides a unified basis for subsequent verification and arbitration.

To improve decision robustness, the framework employs multi-path parallel reasoning rather than relying on a single model output. Multiple reasoning paths are independently generated under low-temperature sampling settings, producing diverse yet regulation-constrained interpretations of the same scene. Then, consensus voting is performed over all valid reasoning paths. When all reasoning paths yield identical operational instructions, the corresponding operational instruction is directly selected as the final result. When inconsistencies arise among the paths, a regulation-guided cross-validation mechanism is activated as a secondary arbitration step. Specifically, decision-critical visual attributes extracted from the *Analysis* field, such as signal color and signal count, are compared with the descriptions of the associated retrieved regulations. A regulation-consistency score is then computed based on the co-occurrence count of these visual attributes within the regulatory texts. The candidate with the highest consistency score is selected as the final decision.

Overall, the framework performs verification and arbitration in a progressive manner. Structured-format verification ensures output validity, multi-path parallel reasoning improves decision robustness, and regulation-guided cross-validation resolves low-consensus cases using external regulatory knowledge. Through the integration of these mechanisms, the proposed framework improves the reliability, interpretability, and robustness under complex operational conditions.

4. Standardized Visual Railway Signal Dataset Construction

The performance evaluation of the proposed framework relies on high-quality multi-modal data resources and a reliable evaluation benchmark. Since the quality of input data directly affects retrieval and reasoning reliability, this study constructs a standardized railway visual-signal dataset and corresponding evaluation framework for railway safety reasoning tasks. Figure 2 illustrates the overall construction pipeline of the proposed standardized dataset, and the following subsections detail each of its three main components.

4.1. Construction of the Standardized Railway Visual-Signal Dataset

The proposed dataset consists of three components: railway scene images, a regulatory knowledge base, and structured question–answer (QA) pairs. The visual images provide first-view railway operational scenes, the knowledge base contains standardized railway regulations, and the QA pairs establish semantic associations between visual scenes and operational rules.

4.1.1. Railway Scene Image Collection and Selection

Existing public railway datasets remain insufficient for railway safety reasoning research, especially for large-scale datasets aligned with Chinese railway operational standards. Therefore, this study investigated available railway datasets, including RailSem19

and FRSign, and selected RailSem19 as the primary visual source due to its large scale and scene diversity.

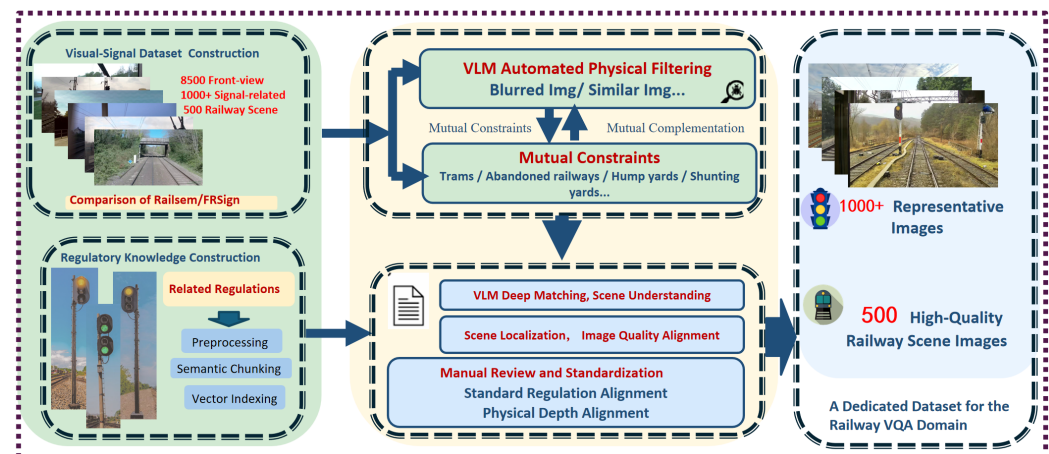


Figure 2. Schematic Diagram Of The Standardized Dataset Construction Process.

RailSem19 is the first publicly available semantic scene understanding dataset for railway environments, containing 8500 annotated front-view railway images collected from train and tram driving cabins. The dataset covers diverse railway infrastructures, weather conditions, illumination settings, and operational scenarios, including railway crossings, station-entry operations, and shunting tasks. Such diversity provides strong generalization capability for railway safety reasoning.

However, preliminary experiments revealed that some signal-related images in RailSem19 suffer from low clarity, incomplete signal display, and inconsistencies with Chinese railway signal standards. To address these limitations, this study manually screened more than 1000 representative images and refined them according to the following criteria: sufficient visual clarity, identifiable scene elements, explicit signal states, and diverse operational conditions.

To further improve signal standardization, the selected images were manually adjusted using standard signal references derived from the *Railway Technical Management Regulations*. Signal colors and appearances were refined to better match standardized Chinese railway signal representations. After processing, 500 high-quality railway scene images with standardized visual signal characteristics were obtained. The dataset was further validated using the Qwen3-VL model to ensure adequate image quality and signal visibility for reasoning tasks.

4.1.2. Regulatory Knowledge Base Construction

The retrieval-augmented generation module requires a structured regulatory knowledge source. Therefore, this study constructed the knowledge base using the *Railway Technical Management Regulations*, which provide authoritative operational and technical standards for railway transportation in China.

Since train drivers primarily rely on visual signal states during operation, this study mainly adopted signal-related provisions from Chapter 3 of the regulations. The knowledge base construction process includes document preprocessing, semantic chunking, and vector indexing.

During preprocessing, the regulation documents were cleaned and standardized through title extraction, whitespace normalization, and formatting correction while preserving semantic integrity. In the semantic chunking stage, a hierarchical segmentation strategy based on paragraphs and sentence-level punctuation was adopted to preserve contextual coherence and avoid semantic fragmentation.

For vector indexing, all knowledge chunks were encoded using the BGE-large-zh-v1.5 embedding model to generate dense semantic representations. The same embedding model was used during both indexing and retrieval stages to maintain semantic consistency. Each knowledge chunk retained its corresponding rule identifier, enabling fact verification during reasoning and suppressing hallucinated outputs.

4.1.3. Structured QA Dataset Construction

Based on the railway scene images and the regulatory knowledge base, this study further constructed a structured QA dataset for railway safety reasoning tasks. Each QA sample contains four components: image file name, question, ground-truth rule ID, and reference answer.

The questions mainly focus on train operational states and potential safety risks, including signal recognition, risk assessment, and operational safety recommendations. The rule IDs correspond to specific clauses in the *Railway Technical Management Regulations*, ensuring that each QA pair is supported by authoritative regulatory evidence.

To balance dataset scale and quality, a semi-automatic QA generation workflow was adopted. First, a large language model generated preliminary QA samples based on scene content and associated regulations using few-shot prompting. Subsequently, format verification and semantic consistency checks were performed automatically to remove invalid or inconsistent samples. Finally, ambiguous entries were manually reviewed and corrected to ensure dataset reliability.

The final dataset consists of 500 manually refined railway scene images, a regulatory knowledge base constructed from the *Railway Technical Management Regulations*, and structured QA pairs covering diverse railway operational scenarios and risk levels.

4.2. Evaluation Benchmark

To objectively evaluate the proposed framework, this study designs a comprehensive evaluation benchmark covering retrieval quality, reasoning accuracy, inference efficiency, and system reliability. The benchmark follows the principles of quantifiability, reproducibility, and practical relevance.

The evaluation framework includes four primary metrics:

- **Rule Matching Accuracy**, which evaluates whether the generated reasoning results correctly identify relevant regulatory clauses;
- **Retrieval Quality**, which measures the effectiveness of the retrieval module in obtaining semantically relevant knowledge;
- **Inference Efficiency**, which evaluates system response speed and reasoning latency;
- **System Reliability**, which assesses reasoning robustness and hallucination suppression capability under complex railway scenarios.

These metrics collectively provide a multidimensional evaluation of the proposed railway safety reasoning framework.

5. Experiments

This section evaluates the proposed framework from four perspectives: retrieval effectiveness, reasoning strategy optimization, component contribution, and explainability analysis. All experiments follow the controlled-variable principle to analyze the contribution of each module.

5.1. Experimental Settings

All experiments were conducted on a server equipped with NVIDIA GPUs. Qwen2.5-VL-3B-Instruct was adopted as the vision-language model, BGE-large-zh-v1.5 was used

as the text embedding model, and ChromaDB served as the vector database. For the configuration, the retrieval module employs a *TOP_K_RETRIEVAL* of 10 and a distance threshold of 2.0. The Visual Language Model (VLM) is deployed with $N = 16$ sequences, a temperature of 0.1, a top- p of 0.8, a maximum of 512 tokens, and a GPU utilization factor of 0.85.

It should be noted that the proposed framework operates in a zero-shot inference manner, where the VLM is used directly for reasoning without any task-specific fine-tuning or training on the railway dataset. This training-free paradigm, which enables seamless adaptation to different large models, reflects the adaptability and decoupled architectural design of the proposed approach, and allows the system to accommodate diverse deployment scenarios by configuring different models. Therefore, all 500 samples are used as the evaluation set, and no training/validation/test split is required.

5.2. Necessity of Retrieval-Augmented Reasoning

We establish two experimental setups to evaluate the impact of the retrieval mechanism: (1) *Baseline-NoRAG*, a baseline comparison where no retrieval results are provided during inference and the model relies entirely on parametric memory; and (2) *Ours-Full*, the full system implementing the complete retrieval pipeline with BGE text retrieval for regulation recall. The experimental results are presented in Table 1.

Table 1. Statistical Results of Ablation Study on the Necessity of Retrieval Methods.

Experiment Group	Total Samples	Correct	Accuracy (%)	Retrieval Failures	Reasoning Failures
Baseline-NoRAG	500	142	28.40	-	358
Ours-Full	500	336	67.20	21	143

Note: In the Baseline-NoRAG group, the retrieval failure metric is denoted by “-” as no retrieval mechanism is involved by experimental design, rendering this metric inapplicable.

The experimental results show that the model without retrieval augmentation achieved only 28.40% accuracy, while the complete framework achieved 67.20% accuracy. This large performance gap indicates that general-purpose large models lack sufficient knowledge of the railway domain, especially when mapping signal combinations to corresponding operational regulations. Therefore, introducing external regulatory knowledge is essential for railway safety reasoning.

The complete framework achieved a retrieval failure rate of 4.20% (21/500), indicating that the retrieval module successfully recalled relevant regulations in most cases. Retrieval failures mainly resulted from the distance-threshold filtering mechanism, which removed semantically weak retrieval results. Reasoning failures remained the dominant error source, accounting for 28.60% (143/500) of all samples.

From a comparative perspective, the Baseline-NoRAG and Baseline-Single configurations serve as indirect references against non-augmented and single-inference VLM baselines. The accuracy gap between Baseline-NoRAG and Ours-Full demonstrates the critical contribution of retrieval augmentation, while the improvement from Baseline-Single to Ours-Full validates the consensus-based reasoning strategy. Given that RAG-based reasoning architectures for railway field-of-view perception remain largely unexplored, self-contained ablation experiments were employed to strengthen the evaluation rigor.

5.3. Reasoning Parameter and Strategy Optimization

This experiment investigates the joint influence of retrieval parameters and reasoning strategies on system accuracy. Five groups of controlled experiments were designed to

analyze the effects of retrieval scale, multi-sequence reasoning, and consensus mechanisms. The variable settings are listed in Table 2.

Table 2. Parameter Settings for Reasoning Strategy Optimization Experiments.

Experiment Group	TOP-K	N Sequence	Temperature	Description
Baseline-Single	10	1	0.2	Single inference baseline
Param-T5	5	16	0.1	Low TOP-K
Param-T15	15	16	0.1	High TOP-K
Param-N8	10	8	0.1	Medium sequence length
Ours-Full	10	16	0.1	Current configuration

The experimental results for different settings are presented in Table 3.

Table 3. Statistics of inference parameters and strategy optimization experiments.

Method	Total Samples	Correct	Accuracy (%)	Voting Consensus (%)	Effective Sequence Rate (%)	Correction Triggers
Baseline-Single	500	333	66.6	-	6.25	0
Param-T5	500	301	60.2	75.29	99.60	13
Param-T15	500	325	65.0	74.33	99.22	15
Param-N8	500	334	66.8	77.49	49.52	11
Ours-Full	500	336	67.2	75.0	99.40	13

Note: In the Baseline-Single group, the voting consensus metric is denoted by “-” as only a single inference sequence is generated by experimental design, rendering this metric inapplicable.

Effect of TOP_K Retrieval Parameter. The TOP_K parameter significantly affects system performance. When TOP_K was reduced to 5, accuracy decreased to 60.20%, indicating that insufficient retrieval results may omit relevant regulations and lead to incomplete reasoning contexts. In contrast, increasing TOP_K to 15 achieved 65.00% accuracy, slightly lower than the baseline due to the introduction of semantically irrelevant noise regulations that interfered with reasoning. These results suggest that excessively small retrieval ranges limit knowledge coverage, whereas overly large retrieval ranges introduce distracting information. Therefore, a moderate retrieval scale provides the best balance between knowledge completeness and semantic relevance.

Effect of Multi-Sequence Reasoning. The sequence number parameter N controls the number of parallel reasoning paths generated during inference and forms the basis of the self-consistency voting mechanism. When $N = 1$, the framework achieved 66.60% accuracy. Increasing N to 16 improved accuracy to 67.20%. The improvement demonstrates that multi-sequence reasoning can reduce inference randomness and improve output stability through majority consensus. Some reasoning paths inevitably produce incorrect conclusions due to stochastic sampling, while consensus voting effectively suppresses unstable outputs. When $N = 8$, accuracy reached 66.80%, indicating diminishing returns as the number of reasoning sequences increases. Therefore, practical deployment should balance reasoning performance and computational cost.

Temperature and Consensus Mechanism. The single-sequence baseline used a temperature of 0.2, whereas all multi-sequence settings adopted a lower temperature of 0.1. Lower temperature values produce more concentrated output distributions, which improve convergence among reasoning sequences and provide a more stable basis for consensus voting. The voting consensus rate of multi-sequence reasoning ranged from 74.33% to 77.49%, indicating partial diversity among reasoning paths. The proposed framework

achieved an effective sequence rate of 99.40%, allowing majority voting to operate on sufficient valid samples and thereby improving statistical reliability.

Voting Consensus Analysis. Since Baseline-Single generates only a single sequence, the voting process is not involved, making the consensus metric inapplicable. For multi-sequence inference, the voting consensus spans from 74.33% to 77.49%, indicating a certain degree of divergence among the generated sequences. Specifically, Param-N8 achieves the highest consensus but suffers from an insufficient voting base, as its effective sequence rate is merely 49.52%, meaning nearly half of the sequences are eliminated during format verification or hallucination filtering. In contrast, the Ours-Full group achieves an effective sequence rate of 99.40%. This high yield ensures that majority voting is conducted over a sufficient sample size, thereby lending greater statistical significance to the consensus conclusions.

Cross-Validation Mechanism. Based on the parameter optimization experiments, comparative experiments were further conducted to evaluate the effectiveness of the cross-validation mechanism. The experimental results are shown in Table 4. The proposed framework further introduces an internal cross-validation mechanism for low-consensus scenarios. When the confidence of majority voting is insufficient, the system performs secondary verification using explicit visual signal attributes such as signal color and signal count. The cross-validation mechanism improved accuracy from 66.0% to 67.20%, demonstrating its effectiveness in correcting ambiguous reasoning outputs. This mechanism mainly contributes in low-consensus situations where statistical voting alone is insufficiently reliable.

The multi-path consensus voting mechanism operates most effectively when the majority of reasoning paths converge toward the correct answer, which is supported by the high effective sequence rate (99.40%) and the voting consensus range (74.33–77.49%) observed in our experiments. Similarly, for the cross-validation mechanism, the experimental results indicate that this mechanism is capable of improving system accuracy to a certain extent at a relatively low computational cost. However, it is worth noting that for retrieval entries with highly similar semantic descriptions, both mechanisms still cannot achieve correct consensus selection or substantial correction of system reasoning results under extreme cases. Addressing such limitations may require further improvements in the model’s intrinsic reasoning capability and its ability to comprehend semantically critical information.

Table 4. Comparative Experimental Data Statistics of Cross-Validation Mechanism Effectiveness.

Experiment Group	Cross-Validation	Total Samples	Correct Count	Accuracy (%)	Correction Trigger Times
Ablation-NoCrossover	Disabled	500	330	66.0	0
Ours-Full	Enabled	500	336	67.2	13

5.4. Component Ablation Study

This experiment evaluates the contribution of each system component by selectively removing specific modules from the complete framework. The experimental design and dataset are summarized in Table 5.

Retrieval-Augmented Reasoning. Removing retrieval augmentation reduced accuracy from 67.20% to 28.40%, representing the largest performance degradation among all ablation settings. This result demonstrates that railway safety reasoning strongly depends on domain-specific regulatory knowledge and cannot be reliably handled by general-purpose large models alone.

Table 5. Component Contribution Ablation Experimental Data Statistics.

Experiment Group	Description	Accuracy (%)	Relative Change from Baseline (%)
Ours-Full	Full System (Baseline)	67.2	-
Ablation-NoRAG	Remove Retrieval-Augmented Generation	28.4	−38.8
Ablation-NoVoting	Remove Majority Voting	65.0	−2.2
Ablation-NoCrossover	Remove Cross-Validation	66.0	−1.2
Ablation-AllSeq1	Number of Inference Sequences = 1	66.6	−0.6

Note: In the Ours-Full group, the relative change metric is denoted by “-” as this group serves as the baseline by experimental design, rendering this metric inapplicable.

Multi-Sequence Reasoning. Removing the multi-sequence reasoning mechanism reduced accuracy from 67.20% to 66.60%. Although the performance decrease is relatively small, the results indicate that multi-path reasoning mainly improves robustness for difficult and ambiguous samples by reducing stochastic inference errors.

Majority Voting Mechanism. Removing majority voting reduced accuracy from 67.20% to 65.00%. Without consensus selection, the system loses the ability to filter inconsistent reasoning outputs, leading to increased instability. Majority voting and multi-sequence reasoning jointly form the self-consistency reasoning mechanism of the proposed framework.

Overall, retrieval augmentation is the most critical component, while multi-sequence reasoning, majority voting, and cross-validation jointly improve reasoning robustness and consistency.

5.5. Error Analysis and Explainability Evaluation

This experiment analyzes system errors from both quantitative and qualitative perspectives and evaluates the explainability of model outputs, as presented in Table 6.

Table 6. Error Type Attribution Data Statistics.

Error Type	Count	Subdivided Source
Retrieval Failure	21	Threshold Filtering/Knowledge Base Missing
Inference Failure	143	High-frequency Confusing Rules

Among all 164 error samples, retrieval failures accounted for 12.80%, while reasoning failures accounted for 87.20%. Retrieval failures occurred when the correct regulatory clauses were not recalled or were removed during distance-threshold filtering, causing the reasoning model to lose critical references. In contrast, reasoning failures occurred even when the correct regulations had already been retrieved, mainly due to rule confusion or visual misinterpretation. For reasoning failure cases, high-frequency pairs of confusing rules were further analyzed, as presented in Table 7.

In addition, to systematically evaluate the interpretability of the system’s structured JSON outputs, we design a multi-dimensional framework grounded in the core decision-making pipeline. This framework maps interpretability onto four complementary dimensions, each corresponding to a critical field in the JSON structure. The specific evaluation criteria are detailed in Table 8.

Perception is the foundational step of the decision-making pipeline, providing visual observations for subsequent rule citation and action reasoning. Perceptual errors propagate through the pipeline and undermine downstream decision validity. The perception accuracy evaluation criteria is presented in Table 9.

Table 7. Interpretability Evaluation Data of High-Frequency Confusion Pairs.

Confused Rule Pair	Misjudgment Count	Difficulty Level	Typical Feature
Rule 16 $\Rightarrow$ Rule 41	14	High	Different Actions: Pass $\leftrightarrow$ Stop
Rule 41 $\Rightarrow$ Rule 40	14	High	Different Actions: Stop $\leftrightarrow$ Pass
Rule 43 $\Rightarrow$ Rule 42	13	High	Correct Rule Contains Yellow, Misjudged Rule Does Not
Rule 15 $\Rightarrow$ Rule 40	9	Medium	Different Actions: Permit $\leftrightarrow$ Pass
Rule 39 $\Rightarrow$ Rule 40	9	Medium	Different Actions: Permit $\leftrightarrow$ Pass

Note: In this table, “ $\Rightarrow$ ” denotes “misclassified as” and “ $\leftrightarrow$ ” denotes “specific misclassification content of the rule”.

Table 8. Interpretability Evaluation Criteria Setting.

Evaluation Dimension	Corresponding Field	Scoring Standard (1–5)	Description
Perception Accuracy	Analysis	1–5	Whether the signal recognition and scene understanding is correct
Rule Citation Accuracy	Rule_Applied	1–5	Whether the Cited Rule ID Matches the Description
Operation Suggestion Rationality	Action	1–5	Whether the Action Matches the Rule
Overall Consistency	Comprehensive	1–5	Whether the Three Fields Are Self-Consistent

Table 9. Perception Accuracy Evaluation Criteria.

Score	Criterion
5	Complete Correct Description of Visual Information
4	Slight Quantity Deviation, or Imprecise Wording
3	Correct Core Visual Information with Additional Wrong Description
2	Most Descriptions Are Wrong, Only a Small Amount of Information Is Correct
1	Completely Wrong

Rule citation serves as the critical bridge between visual perception and operational decision-making, translating observed conditions into rule-based constraints. Rule ID matching accuracy directly determines the legality and relevance of subsequent action recommendations. The rule citation evaluation criteria is presented in Table 10.

Table 10. Rule Citation Evaluation Criteria.

Score	Criterion
5	ID Completely Consistent
1	ID Inconsistent

Operational action recommendations are the final output of the decision-making process, directly dictating vehicle behavior and carrying significant safety implications. A 5-level scoring rubric is developed to evaluate alignment between actions, perceived conditions and cited rules. The rationality evaluation criteria for inference effectiveness is presented in Table 11.

Table 11. Rationality Evaluation Criteria for Inference Effectiveness.

Score	Criterion
5	Action is completely consistent with the requirements of the correct rule
4	Direction is correct but wording lacks precision
3	Partially reasonable; key actions are omitted or direction deviates
2	Incorrect action direction
1	Completely unreasonable or missing

The overall consistency, which evaluates the logical coherence among the Analysis, RuleApplied, and Action fields, is introduced in Table 12.

Table 12. Overall Consistency Evaluation.

Score	Criteria
5	Fully self-consistent logic between actions and three fields: complete chain of Analysis description, corresponding Rule, and corresponding Action
4	Generally self-consistent, with only minor contradictions at the wording level
3	Partially self-consistent: Analysis description generally corresponds to Rule, but Action deviates from the first two
2	Obvious contradiction: e.g., Analysis describes “red light” but Action judges “permit to pass”
1	Severe contradiction: three fields conflict with each other, completely illogical

Here, the Qwen_Plus large language model is used to interpret the output results for comprehensive interpretability evaluation. For each evaluation dimension $d \in \{\text{perception, rule, action, consistency}\}$, where $s_{d,i} \in [1,5]$ denotes the score of the i -th sample on dimension d , the global average score across all N samples is computed as:

$$\overline{S}_d = \frac{1}{N} \sum_{i=1}^N s_{d,i} \quad (1)$$

To examine whether scoring patterns differ between correct and incorrect predictions, we further compute grouped averages. Let C^+ denote the set of samples where the predicted rule matches the ground truth (`is_correct = True`), and C^- the set of error samples (`is_correct = False`). The average score within each group is:

$$\overline{S}_d^+ = \frac{1}{N^+} \sum_{i \in C^+} s_{d,i} \quad (2)$$

$$\overline{S}_d^- = \frac{1}{N^-} \sum_{i \in C^-} s_{d,i} \quad (3)$$

To measure the dispersion of per-sample scores on each dimension, the standard deviation is computed as:

$$\sigma_d = \sqrt{\frac{1}{N} \sum_{i=1}^N (s_{d,i} - \overline{S}_d)^2} \quad (4)$$

The interpretability evaluation results are presented in Table 13.

Table 13. Dimension-Level Interpretability Evaluation Results.

Evaluation Dimension	All Samples Average Score ($\overline{S}_d$)	Correct Cases Average Score ($\overline{S}_d^+$)	Error Cases Average Score ($\overline{S}_d^-$)	Standard Deviation (σ_d)
Visual Perception Accuracy	4.96	4.99	4.88	0.29
Rule Citation Accuracy	3.70	5.00	1.00	1.87
Operation Suggestion Rationality	4.66	4.95	4.05	0.72
Logical Consistency	4.16	4.97	2.48	1.22

Experimental results show that perception accuracy remains relatively high even in incorrect cases, indicating that the vision-language model possesses strong signal recognition capability. Most errors originate from rule-level reasoning rather than visual perception

itself. Moreover, operational recommendations often remain directionally reasonable even when rule citation errors occur, suggesting partial robustness in semantic reasoning.

However, incorrect rule references significantly reduce overall logical consistency because they break the reasoning chain among analysis, applied regulations, and operational recommendations. Correct predictions generally exhibit stronger internal consistency and higher explainability quality.

A deeper analysis of the high-frequency confusion pairs reveals that the primary cause of misclassification lies in the visual similarity between signal configurations that correspond to operationally distinct regulations. For instance, the confusion between Rule 16 (Pass) and Rule 41 (Stop) often occurs when signal color differences are subtle or when multiple signals are partially occluded, making it difficult for the model to reliably distinguish the governing rule. Similarly, confusions involving Rule 43 and Rule 42 stem from the presence or absence of yellow signal indications, which can be ambiguous under varying illumination conditions.

To address these challenges, several optimization strategies can be considered: (1) Fine-grained regulation chunking that separates operationally similar rules into more discriminative knowledge chunks; (2) Hard negative sampling during retrieval to improve the model's ability to distinguish between easily confused rule pairs. These strategies will be explored in future work to further improve the framework's reasoning precision in high-confusion scenarios.

5.6. Inference Efficiency Analysis

To evaluate the practical inference efficiency of the proposed framework, we measure the end-to-end latency on a server equipped with a single NVIDIA RTX A6000 GPU. The framework consists of three model components: Qwen2.5-VL-3B-Instruct as the VLM backbone (3B parameters, bfloat16 precision), BGE-large-zh-v1.5 for text retrieval (325.5 M parameters, CPU-based), and chinese-clip-vit-base-patch16 for visual encoding, with a combined disk footprint of 10.83 GB. During inference, the VLM generates 16 parallel sequences via the vLLM framework, with the actual active GPU memory consumption ranging from approximately 9 to 12 GB.

Under the full configuration, the average end-to-end latency per sample is 11.31 s (median: 10.95 s; min: 7.83 s; max: 17.00 s), covering the complete pipeline of retrieval, visual perception, 16-way parallel inference, consensus voting, and cross-validation, with a total processing time of 5654.71 s for all 500 samples. When reducing the retrieval count to 6 and inference sequences to 5, the average latency decreases to 4.45 s (median: 4.37 s; min: 3.33 s; max: 6.09 s), representing a 60.8% reduction, while accuracy drops to 59.0%. These results indicate a significant trade-off between inference efficiency and reasoning performance, and the configurable sequence parameters provide flexibility for different deployment contexts.

6. Conclusions

This paper presented a retrieval-augmented vision–language framework for railway signal cognition and safety reasoning in train driving scenarios. By integrating front-view railway scene understanding, regulation-oriented knowledge retrieval, structured multi-modal reasoning, and consensus-based decision arbitration, the proposed framework establishes a regulation-grounded cognition pipeline that connects visual signal perception with railway operational rules. In addition, a dedicated railway signal dataset with standardized signal representations and structured QA annotations was constructed to support regulation-oriented multi-modal reasoning and evaluation. Experimental results demonstrate that incorporating railway regulations as external knowledge significantly im-

proves reasoning performance and reliability compared with conventional vision–language reasoning approaches. The ablation studies further confirm the effectiveness of retrieval augmentation and show that consensus-based reasoning strategies contribute to more stable and consistent decision-making. These findings highlight the importance of combining domain-specific regulatory knowledge with multi-modal reasoning for safety-critical railway applications. Future work will focus on improving retrieval precision, expanding regulatory knowledge coverage, and enhancing fine-grained regulation-aware reasoning to further improve the robustness and interpretability of intelligent railway safety cognition systems.

Author Contributions: Conceptualization, Q.W. and C.H.; methodology, Q.W., S.X., J.L. and S.Z.; investigation, Q.W., S.X., J.L. and S.Z.; writing—original draft preparation, Q.W., S.X., J.L. and S.Z.; writing—review and editing, W.L., C.H., X.G., C.F. and W.W.; supervision, W.L. and C.H.; project administration, W.L. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shanghai Aerospace Science and Technology Innovation Foundation (grant number SAST2021-054); the Aviation Science Fund (grant number 20240001053008); the Shanghai Aerospace Science and Technology Innovation Foundation (grant number SAST2023-005); the Beijing Natural Science Foundation (grant number L252146); and the Aviation Science Fund (grant number 201907053004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the corresponding author upon reasonable request. The original railway scene images were derived from the publicly available RailSem19 dataset, and the processed data and structured question–answer pairs are described in the manuscript.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT-4 by OpenAI for language polishing and readability improvement. A large language model was also used only to assist in drafting preliminary candidate QA annotations, which were reviewed and verified by the authors. The authors take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Li, L.; Dao, A.; Zhou, X.; Qiao, Y.; Mai, Z.; Lee, D.; Chen, Z.; Tan, Z.; Bansal, M.; et al. IndustryNav: Exploring Spatial Reasoning of Embodied Agents in Dynamic Industrial Navigation. *arXiv* **2025**, arXiv:2511.17384.
2. D’Ariano, A.; Pacciarelli, D.; Pranzo, M. A branch and bound algorithm for scheduling trains in a railway network. *Eur. J. Oper. Res.* **2007**, *183*, 643–657. [[CrossRef](#)]
3. Cao, Z.; Qin, Y.; Jia, L.; Xie, Z.; Gao, Y.; Wang, Y.; Li, P.; Yu, Z. Railway intrusion detection based on machine vision: A survey, challenges, and perspectives. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 6427–6448. [[CrossRef](#)]
4. Dong, H.; Ning, B.; Cai, B.; Hou, Z. Automatic train control system development and simulation for high-speed railways. *IEEE Circuits Syst. Mag.* **2010**, *10*, 6–18. [[CrossRef](#)]
5. Guo, F.; Liu, J.; Qian, Y.; Xie, Q. Rail surface defect detection using a transformer-based network. *J. Ind. Inf. Integr.* **2024**, *38*, 100584. [[CrossRef](#)]
6. Gibert, X.; Patel, V.M.; Chellappa, R. Deep multitask learning for railway track inspection. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 153–164. [[CrossRef](#)]
7. Li, R.; Prozzi, J.A.; Hong, F. Quantification of Post-Rainfall Moisture Content in Pavement Unbound Layers Using Long-Term Pavement Performance Data. *Transp. Res. Rec. J. Transp. Res. Board* **2025**, *2680*, 660–672. [[CrossRef](#)]
8. Zhan, W.; Chen, Y.; Liu, Q.; Li, J.; Sacchi, M.D.; Zhuang, M.; Liu, Q.H. Simultaneous prediction of petrophysical properties and formation layered thickness from acoustic logging data using a modular cascading residual neural network (MCARNN) with physical constraints. *J. Appl. Geophys.* **2024**, *224*, 105362. [[CrossRef](#)]
9. Wang, Y.; Tan, S.; Mu, X.; Zhang, J. A deep full-sequence convolutional neural network for 3D sedimentary microfacies prediction using time-frequency seismic data. *Geophys. Prospect. Pet.* **2026**, *65*, 469–477. [[CrossRef](#)]

10. Su, T.; Zhang, Y.; Liang, X.; Ma, N. Trajectory Planning Method for Fracture Reduction of Parallel Robots Based on DMP and APF. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 4130–4141. [\[CrossRef\]](#)
11. Liang, X.; Yan, Y.; Wang, W.; Su, T.; He, G.; Li, G.; Hou, Z.G. Adaptive human-robot interaction torque estimation with high accuracy and strong tracking ability for a lower limb rehabilitation robot. *IEEE/ASME Trans. Mechatron.* **2024**, *29*, 4814–4825. [\[CrossRef\]](#)
12. Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Los Alamitos, CA, USA, 2023; pp. 17853–17862.
13. Lv, C.; Qi, M.; Liu, L.; Ma, H. T2sg: Traffic topology scene graph for topology reasoning in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: Piscataway, NJ, USA, 2025; pp. 17197–17206.
14. Chen, H.; Xiong, M.; Lu, Y.; Han, W.; Deng, A.; He, Y.; Wu, J.; Li, Y.; Liu, Y.; Hooi, B. Mlr-bench: Evaluating ai agents on open-ended machine learning research. In *Advances in Neural Information Processing Systems 38*; Curran Associates, Inc.: Red Hook, NY, USA, 2026.
15. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 256–274.
16. Yu, C.; Zhang, Y.; Liu, Z.; Ding, Z.; Sun, Y.; Jin, Z. Frame: Feedback-refined agent methodology for enhancing medical research insights. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 7690–7704.
17. Abisheva, G.; Goranin, N.; Razakhova, B.; Aidynov, T.; Satybaldina, D. Specifics of data collection and data processing during formation of RailVista dataset for machine learning-and deep learning-based applications. *Sensors* **2024**, *24*, 5239. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; Alvarez, J.M. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: Piscataway, NJ, USA, 2025; pp. 22442–22452.
19. An, R.; Zeng, K.; Lu, M.; Yang, S.; Zhang, R.; Ji, H.; Liang, H.; Zhang, W. Concept-as-Tree: A Controllable Synthetic Data Framework Makes Stronger Personalized VLMs. *arXiv* **2025**, arXiv:2503.12999.
20. Xing, Z.; Zhang, X.; Hu, Y.; Jiang, B.; He, T.; Zhang, Q.; Long, X.; Yin, W. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: Piscataway, NJ, USA, 2025; pp. 1602–1611.
21. Ye, H.; Qi, M.; Liu, Z.; Liu, L.; Ma, H. Safedriverag: Towards safe autonomous driving with knowledge graph-based retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 11170–11178.
22. Deng, J.; Sun, G. Research on revision of relevant provisions of high-speed railway emergency disposal in the regulations on railway technical management (RRTM). *Railw. Sci.* **2023**, *2*, 539–549. [\[CrossRef\]](#)
23. Hess, G.; Lindström, C.; Fatemi, M.; Petersson, C.; Svensson, L. Splatad: Real-time lidar and camera rendering with 3d gaussian splatting for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: Piscataway, NJ, USA, 2025; pp. 11982–11992.
24. Zhou, W.; Yang, L.; Zhao, L.; Zhang, R.; Cui, Y.; Huang, H.; Qie, K.; Wang, C. Vision technologies with applications in traffic surveillance systems: A holistic survey. *ACM Comput. Surv.* **2025**, *58*, 1–47. [\[CrossRef\]](#)
25. Huang, X.; Wang, J.; Xia, Q.; Chen, S.; Yang, B.; Li, X.; Wang, C.; Wen, C. V2x-r: Cooperative lidar-4d radar fusion with denoising diffusion for 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: Piscataway, NJ, USA, 2025; pp. 27390–27400.
26. Xie, T.; Zhang, J.; Bai, H.; Nowak, R. Deep active learning in the open world. *arXiv* **2024**, arXiv:2411.06353.
27. Riahi Samani, R.; Nunez Vicencio, A.; De Schutter, B. Bidirectional Long Short-Term Memory Approach for Infrastructure Health Monitoring Using On-Board Vibration Response. *Transp. Res. Rec.* **2025**, 2679, 344–356. [\[CrossRef\]](#)
28. Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B.L.; Zimmer, W.; Cao, H.; Knoll, A.C. Vision language models in autonomous driving: A survey and outlook. *IEEE Trans. Intell. Veh.* **2024**, *9*, 1–20. [\[CrossRef\]](#)
29. Liu, Y.; Wang, C.; Lu, M.; Yang, J.; Gui, J.; Zhang, S. From simple to complex scenes: Learning robust feature representations for accurate human parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5450–5462. [\[CrossRef\]](#)
30. Wang, S.; Wang, C.; Shi, C.; Liu, Y.; Lu, M. Mask-guided mamba fusion for drone-based visible-infrared vehicle detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–12. [\[CrossRef\]](#)
31. Yuan, H.; Li, X.; Zhou, C.; Li, Y.; Chen, K.; Loy, C.C. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 419–437.

32. Sinhamahapatra, P.; Bose, S.; Roscher, K.; Günnemann, S. Zero-shot open-vocabulary ood object detection and grounding using vision language models. In *Proceedings of the Northern Lights Deep Learning Conference 2025*; PMLR: New York, NY, USA, 2025.
33. Ghiasi, R.; Malekjafarian, A. Monitoring of railway tracks maintenance needs using dynamic responses collected by an in-service train. *Railw. Eng. Sci.* **2025**, *33*, 553–580. [[CrossRef](#)]
34. Zhang, B.; Li, J.; Song, N.; Zhang, L. Perception in plan: Coupled perception and planning for end-to-end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2026; Volume 40, pp. 12376–12384.
35. Feng, B.; Mei, Z.; Ost, J.; Ghilotti, F.; Li, B.; Girgis, R.; Majumdar, A.; Heide, F. Verdi: Vlm-embedded reasoning for autonomous driving. *arXiv* **2025**, arXiv:2505.15925.
36. Zablocki, É.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of deep vision-based autonomous driving systems: Review and challenges. *Int. J. Comput. Vis.* **2022**, *130*, 2425–2452. [[CrossRef](#)]
37. Zhang, S.; Li, R.; Deng, S.; Zheng, Z.; Zhang, Y.; Li, J.; Zhang, K.; Zhang, T.; Wu, W.; Wang, Q. RailVQA: A Benchmark and Framework for Efficient Interpretable Visual Cognition in Automatic Train Operation. *arXiv* **2026**, arXiv:2603.27112.
38. Xie, S.; Kong, L.; Dong, Y.; Sima, C.; Zhang, W.; Chen, Q.A.; Liu, Z.; Pan, L. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data and Metric Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2025; pp. 6585–6597.
39. Qi, J.; Xu, Z.; Shao, R.; Chen, Y.; Di, J.; Cheng, Y.; Wang, Q.; Huang, L. Rora-vlm: Robust retrieval-augmented vision language models. *arXiv* **2024**, arXiv:2410.08876.
40. Zhang, S.; Huang, W.; Gao, Z.; Chen, H.; Lv, C. Wisead: Knowledge augmented end-to-end autonomous driving with vision-language model. *arXiv* **2024**, arXiv:2412.09951.
41. Ding, Z.; Zhang, H.; Wu, P.; Pang, G.; Yang, Z.; Wang, P.; Zhang, Y. Slowfastvad: Video anomaly detection via integrating simple detector and rag-enhanced vision-language model. *arXiv* **2025**, arXiv:2504.10320.
42. Zhang, J.; Yang, X.; Wang, T.; Yao, Y.; Petiushko, A.; Li, B. Safeauto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. *arXiv* **2025**, arXiv:2503.00211.
43. Liu, J.; Chen, K.; Duan, H.; Li, C. A knowledge graph-based hazard prediction approach for preventing railway operational accidents. *Reliab. Eng. Syst. Saf.* **2024**, *247*, 110126. [[CrossRef](#)]
44. Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; Wu, X. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 3580–3599. [[CrossRef](#)]
45. Xiao, S.; Liu, Z.; Zhang, P.; Muennighoff, N.; Lian, D.; Nie, J.Y. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 641–649.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.