ExpressMind: A Multimodal Pretrained Large Language Model for Expressway Operation

Zihe Wang
Beihang University
Beijing, China
by2313310@buaa.edu.cn

Yihuan Wang
Beihang University
Beijing, China
yhuanwang@buaa.edu.cn

Haiyang Yu
Beihang University
Beijing, China
hyyu@buaa.edu.cn

Zhiyong Cui*
Beihang University
Beijing, China
zhiyongc@buaa.edu.cn

Xiaojian Liao
Beihang University
Beijing, China
liaoxj@buaa.edu.cn

Chengcheng Wang
Shandong Hi-speed Group Co., Ltd
Jinan, China
wangchengcheng@sdhsg.com

Yonglin Tian
Institute of automation, Chinese Academy of Sciences
Beijing, China
tyldyx@mail.ustc.edu.cn

Yongxin Tong
Beihang University
Beijing, China
yxtong@buaa.edu.cn

Abstract

The current expressway operation relies on rule-based and isolated models, which limits the ability to jointly analyze knowledge across different systems. Meanwhile, Large Language Models (LLMs) are increasingly applied in intelligent transportation, advancing traffic models from algorithmic to cognitive intelligence. However, general LLMs are unable to effectively understand the regulations and causal relationships of events in unconventional scenarios in the expressway field. Therefore, this paper constructs a pre-trained multimodal large language model (MLLM) for expressways, ExpressMind, which serves as the cognitive core for intelligent expressway operations. This paper constructs the industry's first full-stack expressway dataset, encompassing traffic knowledge texts, emergency reasoning chains, and annotated video events to overcome data scarcity. This paper proposes a dual-layer LLM pre-training paradigm based on self-supervised training and unsupervised learning. Additionally, this study introduces a Graph-Augmented RAG framework to dynamically index the expressway knowledge base. To enhance reasoning for expressway incident response strategies, we develop a RL-aligned Chain-of-Thought (RL-CoT) mechanism that enforces consistency between model reasoning and expert problem-solving heuristics for incident handling. Finally, ExpressMind integrates a cross-modal encoder to align the dynamic feature sequences under the visual and textual channels, enabling it to understand traffic scenes in both video and image modalities. Extensive experiments on our newly released multi-modal expressway benchmark demonstrate that ExpressMind comprehensively outperforms existing baselines in event detection, safety response generation, and complex traffic analysis. The code and data are available at: https://wanderhee.github.io/ExpressMind/.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Large Language Models, Intelligent Expressway Operations, Pretraining Paradigm, Chain-of-Thought, Multimodal Understanding

ACM Reference Format:
Zihe Wang, Yihuan Wang, Haiyang Yu, Zhiyong Cui, Xiaojian Liao, Chengcheng Wang, Yonglin Tian, and Yongxin Tong. 2026. ExpressMind: A Multimodal Pretrained Large Language Model for Expressway Operation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

With the continuous advancement of intelligent transportation systems (ITS), expressway operation is evolving from the reactive and rule-based paradigm towards intelligent agents endowed with deep cognitive reasoning capabilities. Breakthroughs in artificial intelligence [36], particularly the emergence of Large Language Models (LLMs) with exceptional computational and reasoning abilities [38], are profoundly shaping the intelligent transformation across ITS industries. However, the expressway sector still lacks a domain-specific LLM which is capable of deeply adapting to complex expressway operational needs. At the current stage, advancing foundation models' core capabilities such as expressway knowledge integration, multimodal scene understanding, and autonomous incident reasoning remains a critical bottleneck, necessitating a fundamental methodological shift.

*Corresponding author

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

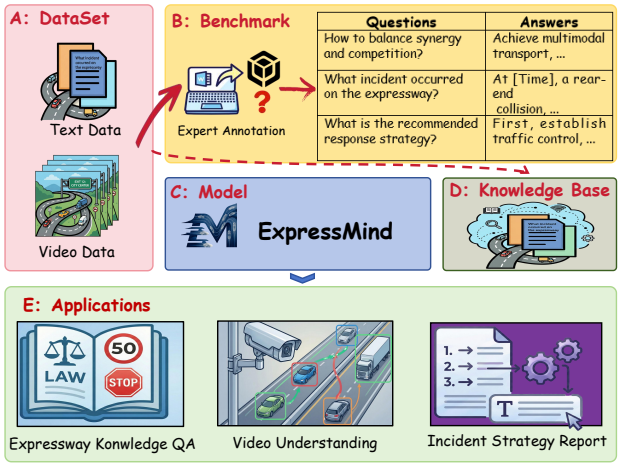


Figure 1: Overview of ExpressMind.

Obviously, most open-source LLMs [20] lack a deep understanding of publicly inaccessible specialized knowledge in the expressway domain, such as technical standards and regulations. Dynamic information and professional terminology cannot be promptly fed into LLMs and, thus, their reasoning and decision-making processes struggle to guarantee safety and efficiency requirements. Furthermore, existing methods exhibit shortcomings in key visual feature extraction and traffic-related reasoning within multimodal scenarios. Therefore, current approaches fail to meet the dynamic and precise operation requirements of expressway, lacking an intelligent central hub capable of multi-task collaborative cognition and deep industry understanding.

However, constructing an expressway domain-specific multimodal large model as such intelligent central hub faces considerable difficulties and unique challenges. The highly heterogeneous and complex multimodal data in this field is the first hurdle. The unstructured data such as real-time monitoring videos and semi-structured data including incident records, making it extremely difficult to achieve effective alignment and fusion between different modalities. Additionally, the expressway domain has strict requirements on safety and accuracy. The desired foundation model needs to accurately grasp professional knowledge such as traffic engineering principles and emergency disposal specifications, which brings great challenges to the in-depth integration of domain knowledge and the design of multimodal architecture. More importantly, the scarcity of high-quality labeled multimodal data in the expressway field, coupled with the privacy and security constraints, further increases the difficulty of model training and optimization, becoming a key obstacle to building a high-performance domain-specific multimodal large model.

To address these challenges, this paper introduces ExpressMind, a domain multimodal LLM for expressway operation. We construct the first full-stack expressway dataset and propose a two-stage pretraining paradigm for the internalization of expressway-domain knowledge. This study also develops a Reinforcement Learning (RL)-based Chain-of-Thought (CoT) alignment mechanism to strengthen domain reasoning. Furthermore, a visual-enhanced cross-modal encoder is incorporated and a graph-based retrieval-augmented generation (RAG) is proposed to enhance the extraction of key traffic scene characteristics and dynamic knowledge. The integration of these modules as a whole build the foundation of the multimodal pretrained LLM to process multi-source expressway data and provide efficient expressway operation decision support.

The overview of ExpressMind is illustrated in Figure. 1, and its five core contributions are summarized as follows:

- **Full-stack Expressway dataset**: This study constructs the first industry's full-stack expressway dataset spanning text cognition, logical reasoning, and visual perception, including three specialized subsets: traffic knowledge texts, emergency response reasoning, and event video scene understanding.
- **RL-aligned CoT Reasoning**: We design a RL-based expressway strategy alignment strategy in LLM training, which can significantly enhance the model's logical reasoning and self-correction capabilities.
- **Graph-Augmented Retrieval**: A graph RAG-based dynamic knowledge base is established for critical expressway information retrieval and indexing.
- **Multimodal Alignment mechanism**: A Visual-Prior Alignment mechanism is designed by enforcing alignment and reweighting of visual tokens to enhance the understanding of visual features.
- **Multi-modal Benchmark**: The multi-modal Benchmark for evaluating LLMs within the expressway domain is released, encompassing four evaluation subsets: basic knowledge comprehension, video incident detection, safety response generation, and traffic analysis reporting.

2 Relatedwork

Traffic related foundation model: Current mainstream LLM architectures are primarily categorized into three paradigms: Encoder-only (e.g., BERT [6]), Encoder-Decoder (e.g., GLM [8]), and Decoder-only (e.g., GPT [1], LLaMA [7]). To adapt general-purpose models to vertical domains, methodologies such as Supervised Fine-Tuning (SFT) [22] and Parameter-Efficient Fine-Tuning (PEFT) [12] have been introduced. Within the transportation sector, models such as TransGPT [28] and TrafficGPT [37] advances traffic safety analysis via domain-adaptive training. Furthermore, UrbanGPT [17] integrates spatio-temporal encoders with LLMs through instruction tuning for general urban analysis. However, a domain LLM for expressway tasks has yet to emerge.

MLLMs: CLIP [24] established the foundation of multimodal learning by aligning the representation spaces of images and texts via contrastive learning. To endow LLMs with visual comprehension capabilities, LLaVA [19] introduces a linear projection layer to map visual features into token embeddings processable by language models. BLIP-2 [16] proposes the Q-former architecture to extract text-relevant features from frozen visual encoders.Following this research trajectory, the Qwen-VL series [4] is subsequently proposed to align visual and linguistic representations, In traffic scenarios, MLLMs have been applied to tasks such as accident analysis (TrafficLens [2], MoTIF [30]) and anomaly detection (Anomaly-OneVision [32]) by facilitating semantic understanding of surveillance footage.

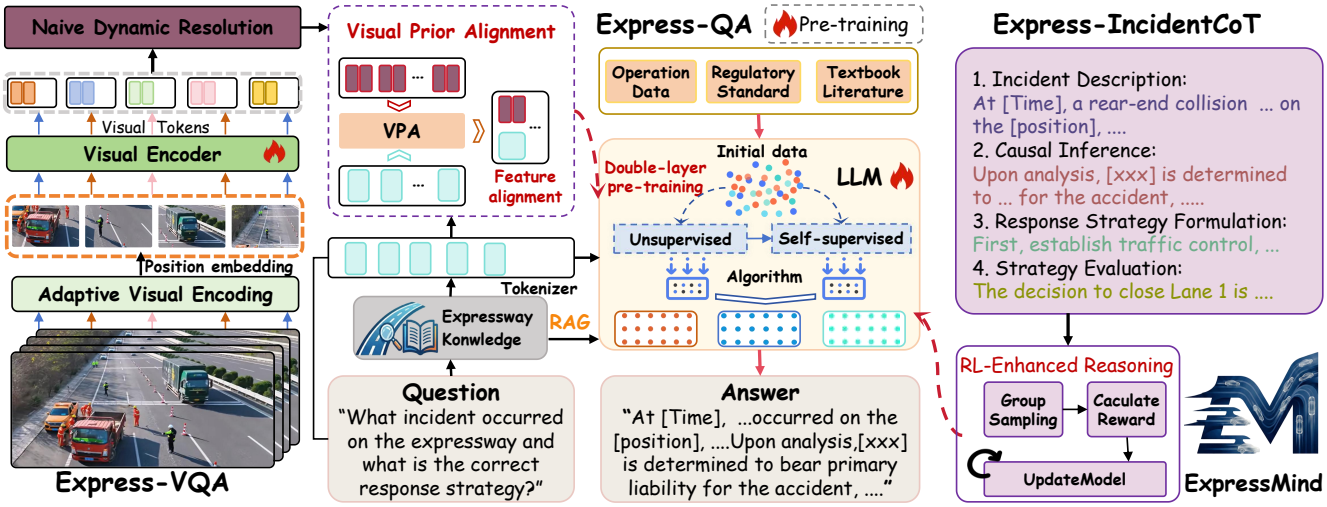


Figure 2: The Overall Framework of ExpressMind.

Reinforcement Learning for Reasoning: Reinforcement Learning from Human Feedback [22] has emerged as the prevailing paradigm for aligning LLMs with human intent. Algorithms such as DPO [25], CPO [31], and GRPO [26] leverage preference information inherent in CoT processes to further optimize reasoning trajectories and training efficiency. To mitigate reward hacking, DreamPRM [5] introduces a domain re-weighting mechanism. Recently, integrating the semantic comprehension of LLMs with the decision-making capabilities of RL has moved to the forefront of transportation research. Traffic-R1 [40] and LLMLight [15] employ RL to enhance the generalization of LLMs in signal control tasks. AgentsCoMerge [13] uses RL with ramp and density rewards for traffic optimization, Time-LLM [14] transforms time-series into text via RL, achieving SOTA in traffic forecasting. While RL has been applied in related traffic tasks, its use for enhancing reasoning in the expressway domain remains unexplored.

3 Methodology

This study proposes ExpressMind, a domain-specific MLLM tailored for expressway operation. The overall framework is illustrated in Figure. 2 and the key components are introduced as follows:

3.1 Task-oriented Domain Data Profiling

To address the domain-specific tasks depicted above, which include Expressway Knowledge QA, Video Understanding, and Incident Strategy Report, this study collects four distinct types of data to support the complete training pipeline of ExpressMind, as illustrated in Figure 3.

- **Textual Data**: To establish the model's fundamental domain understanding, textual data, including policy documents, expert knowledge, and SFT QA pairs, are used in the **3.2 Pre-training Stage**.
- **Incident CoT Data**: To refine reasoning trajectories via reinforcement learning, incident CoT data comprises incident descriptions, causal reasoning, response strategies, and evaluations. It is employed during the **3.3 RL Alignment Stage**.
- **Dynamic Knowledge Base**: To ensure model responses remain aligned with the latest operational scenarios, it contains real-time traffic conditions, incident reports, and traffic flow data, providing **3.4 real-time retrieval augmentation** across all training stages.
- **Multimodal Data**: To achieve video-language understanding, data such as accident images and congestion videos are introduced in the **3.5 Cross-modal Alignment Stage** to achieve video-language understanding.

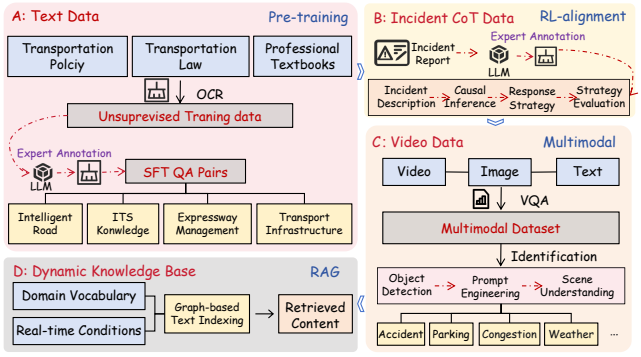


Figure 3: Task-oriented Domain Data Profiling.

3.2 Training Paradigm of Pretrained LLM

To ensure the model acquires high-quality foundational knowledge for expressway scenarios, we constructed a dedicated dataset containing unlabeled text and self-supervised QA pairs, with all data undergoing rigorous deduplication and standardization. ExpressMind, built upon the Qwen foundation model, adopts a two-phase

pre-training strategy: the first phase establishes fundamental scenario knowledge, and the second phase adapts the model to handle complex domain-specific tasks, respectively.

Stage 1: Unsupervised Training. In this phase, model parameters θ are optimized by minimizing the negative log-likelihood loss. Given an input sequence $x = \{x_1, x_2, \ldots, x_T\}$ derived from domain-specific corpora, the pre-training loss function $\mathcal{L}_{PT}$ is formulated as:

$$\mathcal{L}_{PT}(\theta) = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t}; \theta) \tag{1}$$

where $x_{<t}$ denotes the context sequence preceding time step t, and $P(x_t \mid x_{<t}; \theta)$ represents the conditional probability of the model predicting the next token given the current parameters.

Stage 2: Full-Parameter Supervised Fine-Tuning. Following the acquisition of foundational domain knowledge, full-parameter SFT is conducted to align the model with specific tasks and instructions in the expressway transportation domain. During training, To ensure the model focuses on response generation, a masked loss strategy is employed by introducing a binary mask vector M, where $M_t = 0$ corresponds to instruction tokens and $M_t = 1$ to response tokens. Consequently, the loss function for supervised fine-tuning, denoted as $\mathcal{L}_{SFT}$, is formulated as:

$$\mathcal{L}_{SFT}(\theta) = -\frac{1}{\sum_{t=1}^{T} M_t} \sum_{t=1}^{T} M_t \cdot \log P(x_t \mid x_{<t}; \theta) \tag{2}$$

This two-stage training equips the model with an in-depth mastery of expressway domain knowledge, providing a basis for the following alignment and reasoning tasks.

3.3 RL for Expressway Strategy Alignment

Although LLMs have acquired fundamental domain cognition through full-parameter pre-training, when dealing with unseen complex expressway accident scenarios, their generated response strategies often fail to establish a complete logical chain from scene analysis to strategy formulation and evaluation. It notably lacks deep logical deduction and fails to ensure strategic optimality. As shown in Figure 4, to enforce the "Perception-Analysis-Decision-Reflection" cognitive loop and address its reasoning bottleneck, we leverage a CoT dataset derived from real-world expressway emergency responses and employ the Group Relative Policy Optimization (GRPO) algorithm to mine underlying logical patterns, thereby significantly enhancing the model's reasoning capabilities.

The core mechanism of GRPO involves sampling a group of candidate outputs $o_1, o_2, \ldots, o_G$ for a given query q and computing gradients by evaluating the relative scores within the group. Here, q is from the set of all possible queries Q. The objective function is formulated as follows:

$$J_{GRPO}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}} [\mathcal{L}] \tag{3}$$

$$\mathcal{L} = \frac{1}{G}\sum_{i=1}^{G} \left(\min\left(r_i A_i, \text{clip}(r_i, 1-\epsilon, 1+\epsilon)A_i\right) - \beta D_{KL}(\pi_\theta || \pi_{ref})\right) \tag{4}$$

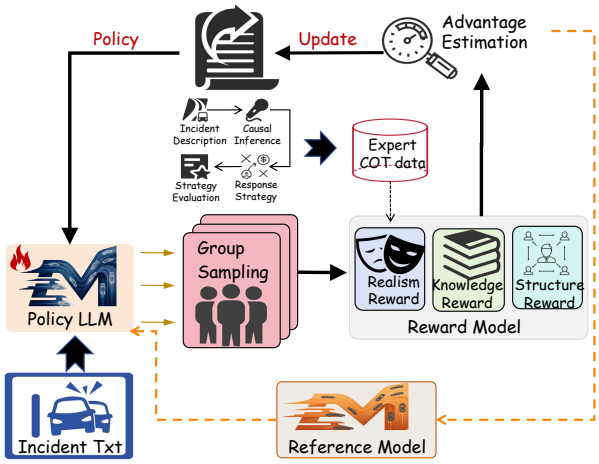


Figure 4: Schematic of the RL-based Reasoning Enhancement.

where A_i denotes the advantage function. The term βD_{KL} acts as a regularization constraint to mitigate catastrophic forgetting during the reinforcement learning process, explicitly ensuring that the model's linguistic generation remains aligned with the standardized traffic terminology acquired during the SFT phase.

Through this mechanism, the algorithm effectively reduces the variance of gradient estimation, ensuring that the model prioritizes learning the relative superiority of strategies over absolute scores. Therefore, this enables stable policy iteration within the complex reasoning space of traffic incident disposal.

To steer the model towards structured accident response Logic, we design a multi-dimensional reward $R_{total} = \lambda_1 R_{struct} + \lambda_2 R_{know} + \lambda_3 R_{sem}$, comprising three decoupled terms:

- **Structural Integrity (R_{struct}):** To enforce the "Perception-Analysis-Decision-Reflection" cognitive loop, we employ a gated counting mechanism. The reward accumulates only if the four stage-specific tags $S_{1..4}$ appear in a strict monotonic order:

$$R_{struct} = \left(\sum_{k=1}^{4} \mathbb{I}(S_k \in O)\right) \cdot \mathbb{I}\left(\text{idx}(S_1) < \text{idx}(S_2) < \text{idx}(S_3) < \text{idx}(S_4)\right) \tag{5}$$

- **Domain Alignment (R_{know}):** We maximize the coverage of stage-specific expert terminology $\mathcal{V}_k$ while penalizing linguistic degradation via a perplexity (PPL) constraint:

$$R_{know} = \frac{1}{K}\sum_{k=1}^{K} \omega_k \frac{|S_k \cap \mathcal{V}_k|}{|S_k|} - \eta \cdot \text{ReLU}\big(\text{PPL}(O) - \tau_{ppl}\big) \tag{6}$$

- **Semantic Consistency (R_{sem}):** To ensure strategic optimality, we compute the cosine similarity between the model's decision logic and a reference set $\mathcal{D}_{ref}$ containing expert records and teacher traces in the embedding space:

$$R_{sem} = \max_{d\in\mathcal{D}_{ref}} \cos\left(\phi(S_2 \oplus S_3), \phi(d)\right) \tag{7}$$

The overall algorithm process for RL reasoning alignment is presented in Table 1. At its core, the study generates decision-making strategies equipped with complete and verifiable expressway emergency response processes. This explicit reasoning trace enhances the interpretability and reliability of the model's outputs. Implementation details regarding the expert vocabulary $\mathcal{V}_k$ and hyperparameters are provided in Appendix B.

Algorithm 1 Traffic Incident Strategy Alignment via GRPO

Require: unstructured traffic incident text description $q \sim P(Q)$
Ensure: optimized traffic disposal policy π^*
Initialize policy model π_θ, reference model π_{ref}, expert database $\mathcal{D}_{ref}$
for training epoch $t = 1$ **to** T **do**
  // Step 1: Candidate Response Sampling
  Sample G candidate responses from current policy:
  $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{t-1}}(\cdot \mid q)$
  // Step 2: Multi-dimensional Reward Evaluation
  for each candidate response o_i **do**
    Compute reward: $r_i = \sum_j \lambda_j R_j(o_i)$
    where $R \in \{R_{struct}, R_{know}, R_{sem}\}$
  end for
  // Step 3: Advantage Normalization within Group
  Compute group mean: $\mu_{\text{group}} = \frac{1}{G}\sum_{i=1}^{G} r_i$
  Compute group standard deviation:
  $\sigma_{\text{group}} = \sqrt{\frac{1}{G}\sum_{i=1}^{G}(r_i - \mu_{\text{group}})^2}$
  for each candidate response o_i **do**
    Compute advantage: $A_i = \frac{r_i - \mu_{\text{group}}}{\sigma_{\text{group}} + \epsilon}$
  end for
  // Step 4: Policy Update via GRPO Objective
  Update parameters by maximizing GRPO objective:
  $\theta_t \leftarrow \text{argmax}_\theta \; J_{\text{GRPO}}(\theta; A_i, \pi_{\theta_{t-1}}, \pi_{ref})$
end for
return optimized policy $\pi^* \leftarrow \pi_{\theta_T}$

3.4 Knowledge Graph-Augmented Retrieval

The static parameters of LLMs cannot capture dynamic information and professional vocabulary. Therefore, this paper constructs a expressway knowledge base to assist LLMs in learning these knowledge. Traffic knowledge is unstructured data so that a graph-based RAG must be adopted to retrieve the knowledge base. This study employs LightRAG [11] to enhance the ability to update incremental knowledge which introduces a dual-layer retrieval mechanism by constructing a structured graph index

During the indexing phase, an unstructured traffic corpus D is transformed into a structured and incrementally updatable knowledge graph $\widehat{G} = (\widehat{V}, \widehat{E})$. Here, each node $\widehat{V}$ represents a standardized traffic term, and each edge $\widehat{E}$ encodes semantic relations.

This paper designs an entity and relation extraction module to identify the entities specific to the transportation field and their interrelationships. The proposed LLM profiling is the generation of a structured key-value pair (K, L) for every node $v \in V$ and edge $e \in E$, where the key K serves as a normalized identifier for efficient retrieval, and the value L is an LLM-generated summary that integrates multi-source definitions and usage contexts. Deduplication merges redundant nodes and edges of different text fragments through semantic similarity comparison. Given a new document D', its corresponding subgraph $\widehat{G'} = (\widehat{V'}, \widehat{E'})$ is generated independently and merged into the existing graph via set union operations $\widehat{V} \cup \widehat{V'}, \quad \widehat{E} \cup \widehat{E'}$.

In the retrieval and generation phase, a dual-level retrieval paradigm is employed to jointly capture concrete facts and abstract concepts. Given an initial non-standardized response $\hat{q}_{\text{raw}}$, an LLM first extracts two types of keywords: local terms $k^{(l)}$ (e.g., "long queue", "red-green light") and global semantic cues $k^{(g)}$ (e.g., "traffic congestion", "signal control"). Two parallel retrieval paths are then activated:

Low-level retrieval focuses on exact or near-exact matching by computing the similarity between the embedding of $k^{(l)}$ and the key embedding of entity nodes:

$$s_{\text{low}}(v) = \text{sim}\big(e(k^{(l)}), e(K_v)\big) \tag{8}$$

where $e(\cdot)$ denotes an embedding function, $\text{sim}(\cdot)$ is typically cosine similarity, and K_v, K_e denote the retrieval keys of node v and edge e, respectively.

High-level retrieval operates at the conceptual level by matching $k^{(g)}$ against topic keys associated with relation edges:

$$s_{\text{high}}(e) = \text{sim}\big(e(k^{(g)}), e(K_e)\big) \tag{9}$$

All retrieved structured descriptions and associated text snippets are concatenated into a unified context C, which is fed into the LLM to perform term-level replacement along with $\hat{q}_{\text{raw}}$. The final normalized output $\hat{q}_{\text{norm}}$ is generated according to the RAG formulation: $p(\hat{q}_{\text{norm}} \mid \hat{q}_{\text{raw}}, C) \propto \exp\big(f_{\text{LLM}}(\hat{q}_{\text{raw}}; C)\big)$, where f_{LLM} denotes the LLM internal scoring function.

3.5 Multimodal Understanding with VPA

End-to-end Multimodal understanding of expressway is a critical task for expressway supervision. This paper combines a visual encoder with ExpressMind to form a MLLM with video understanding capabilities. In order to enhance the ability of visual feature extraction, this paper introduces a novel visual encoding architecture integrated with a Visual-Prior Alignment (VPA) mechanism, as illustrated in Figure 5. The feature of the visual encoder is $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_t\}, \quad \mathbf{I} \in \mathbb{R}^{N_v \times d_v}$, where N_v is the total number of visual tokens and d_v is the dimension of each visual feature vector.

We design a cross-modal projection network and use layer normalization to stabilize the training process as shown in the following equation:

$$\mathbf{H}_v = \text{LayerNorm}\left(\mathbf{W}_2 \cdot \text{GELU}\left(\mathbf{W}_1 \cdot \mathbf{I} + \mathbf{b}_1\right) + \mathbf{b}_2\right) \tag{10}$$

where, $H_v \in \mathbb{R}^{N_v \times d_{\text{LLM}}}$, $\mathbf{W}_1 \in \mathbb{R}^{d_v \times d_h}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_h \times d_{\text{LLM}}}$ are learnable weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ are bias terms, and d_h denotes the intermediate hidden dimension. The output dimension d_{LLM} matches the hidden size of the LLM.

This paper employs MRoPE to address the issue of uneven frequency allocation in video understanding. It allocates feature channels to the temporal, height, and width axes in a fine-grained polling manner, ensuring that each positional axis is encoded with a full

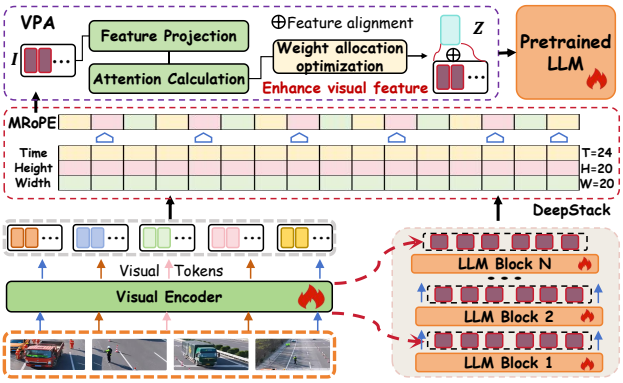


Figure 5: Multimodal Encoding Framework.

frequency spectrum ranging from high to low frequencies. Furthermore, this study introduces the DeepStack[21] mechanism to extract feature maps and perform cross-layer fusion at different depths, which enables the language decoder to access a complete visual hierarchy spanning from pixel-level to semantic-level information.

The visual features may suffer from feature attenuation during the sequence alignment after encoding due to the compression of sequence length. This paper innovatively proposes the VPA mechanism to overcome this problem. It introduces learnable cross-modal attention reweighting to achieve dynamic alignment of visual and language features. VPA explicitly enhances the computational weight of visual tokens in multimodal fusion, establishing a visual-priority inductive preference. Given the projected visual features $\mathbf{H}_v$ and the text embeddings $\mathbf{U} \in \mathbb{R}^{L_t \times d_{\text{LLM}}}$ from the instruction prompt, L_t is the text sequence length. The formula for adjusting the learnable cross-modal attention weights in VPA is as follows:

$$\hat{\mathbf{H}}_v = \text{softmax}\left(\frac{\mathbf{H}_v \mathbf{W}_p \mathbf{U}^\top}{\sqrt{d_{\text{LLM}}}}\right) \cdot \mathbf{H}_v \tag{11}$$

where $\mathbf{W}_p \in \mathbb{R}^{d_{\text{LLM}} \times d_{\text{LLM}}}$ is a learnable projection matrix. This method increases the weight of visual features when aligning visual tokens and text, establishing the inductive bias of visual priors during the feature fusion process.

The enhanced visual features $\hat{\mathbf{H}}_v$ are concatenated with the text embeddings $\mathbf{U}$ to form a unified multimodal input sequence:

$$\mathbf{Z} = \left[\hat{\mathbf{H}}_v; \mathbf{U}\right] \in \mathbb{R}^{(N_v + L_t) \times d_{\text{LLM}}} \tag{12}$$

The combined representation $\mathbf{Z}$ is fed into the LLM and generates natural language description of video, which achieves understanding of traffic scenarios.

4 Experiments

4.1 Dataset

This study released a comprehensive open-source expressway dataset comprising four specialized sub-datasets and a standardized benchmark, for the training and evaluation of **ExpressMind**, focusing on multi-modal capabilities such as expressway domain knowledge understanding, incident response strategy reasoning, video understanding and incident detection.

1: Express-Insight contains over 7 million tokens of high-quality text where the content serves as a domain-specific corpus for unsupervised pre-training. The dataset is derived from web-crawled resources including traffic laws, expressway policy documents, and theoretical books on Smart expressways and Intelligent Transportation Systems.

2: Express-QA contains over 870,000 samples where QA pairs are obtained through a quality-aware generation and refinement process using DeepSeek-V3 with designed prompts.

3: Express-IncidentCoT contains 1,786 incident response strategy Chain-of-Thought samples derived from real-world incident reports of Shandong Expressway, structured into a four-stage cognitive chain: [Incident Description] → [Causal Inference] → [Response Strategy Formulation] → [Strategy Evaluation].

4: Express-VQA a multi-modal dataset for expressway visual reasoning, integrating 1627 surveillance videos from expressways in Shandong and Guangdong, China. The average duration of the videos exceeds 2 minutes. There are over 3,200 pairs of VQA pairs. There are also 12 sets of surveillance video from Tianjin that cover two consecutive days. The data is collected across 70 roads at 1920×1080 resolution, encompassing diverse times and weather conditions to evaluate model robustness against seven core traffic anomalies such as accidents, congestion, and construction.

4.2 Experiment Setup

The experimental environment is configured with the following specifications: All experimental workflows—including model training, testing, and inference—were executed on a server node equipped with 8 NVIDIA H20 GPUs. The software environment is built upon Python 3.10+, PyTorch 2.4.0+, and CUDA 12.4+. The transformers, tokenizers, torchvision, and opencv-python libraries are employed for processing text, model, and image data, respectively. Stable versions of DeepSpeed, Accelerate, PEFT, and Flash-Attention 2 are utilized to facilitate efficient distributed fine-tuning, thereby ensuring high training efficiency and stability. Under this configuration, the end-to-end training of the framework required approximately 700 gpu hours. In particular, the hyperparameter configurations for each training stage are detailed in Appendix A.

4.3 Quantitative Results

4.3.1 Pre-training Results. The results presented in Table 1 summarize the performance of ExpressMind across three specialized tasks (totaling 20,000 test questions): Expressway Laws & Regulations QA, Smart Expressway Knowledge QA, and Intelligent Transportation System Knowledge QA. For each task, We used a set of multi-dimensional metrics, including Accuracy, F1-Score, Embedding Similarity, and GPT-Score, to benchmark our model againstestablished open-source baseline models such as Qwen-32B [33], Llama-3.3-70B [9], and the DeepSeek-R1-Distill series [10]. Despite its specialized focus, ExpressMind consistently outperforms these LLM across all evaluation dimensions. Notably, in the *Expressway Laws & Regulations QA* task, our model achieves a peak MCQ

Table 1: Pre-training results on QA tasks.

Models	MCQ [%]			True/False [%]			Fill-in-the-Blank [%]		Short Answer [%]	
	Acc	Rec	F1	Acc	Rec	F1	Emb	F1	GPT-Score	F1
Expressway Laws & Regulations QA										
Qwen-32B	97.9	96.5	97.2	96.1	95.8	96.4	93.7	85.3	75.6	82.1
Baichuan-32B	91.3	88.5	89.9	89.5	88.2	89.4	88.8	78.2	69.3	74.5
DeepSeek-R1-Distill-Qwen-32B	96.1	96.3	97.4	96.5	96.0	96.1	92.4	84.1	76.6	82.4
DeepSeek-R1-Distill-Llama-70B	97.0	96.2	96.5	97.0	95.8	96.4	95.9	88.7	85.9	87.8
Llama-3.3-70B	97.5	96.0	96.8	96.5	96.2	96.8	96.6	89.4	86.1	87.2
GLM-4-32B	91.7	89.4	90.5	90.1	89.5	90.3	90.4	80.5	71.4	76.4
ExpressMind-14B	**98.4**	**97.9**	**98.1**	**98.2**	**97.5**	**98.3**	**97.4**	**90.5**	**86.8**	**88.5**
Smart Expressway Knowledge QA										
Qwen-32B	96.5	95.2	95.8	95.5	94.8	95.7	83.1	78.5	76.4	79.5
Baichuan-32B	89.7	87.5	88.6	88.2	87.0	88.1	87.7	77.5	68.4	73.8
DeepSeek-R1-Distill-Qwen-32B	95.4	96.1	94.8	96.7	95.5	96.2	95.6	88.9	83.1	86.0
DeepSeek-R1-Distill-Llama-70B	95.9	96.1	95.5	94.8	96.2	95.7	94.1	88.4	84.2	86.1
Llama-3.3-70B	**96.8**	95.9	96.3	95.8	95.4	96.1	95.7	88.5	84.0	86.9
GLM-4-32B	90.6	88.7	89.6	90.2	89.1	90.4	89.3	79.8	70.9	75.2
ExpressMind-14B	96.7	**96.2**	**96.4**	**97.5**	**97.0**	**97.6**	**96.4**	**89.2**	**85.4**	**87.8**
Intelligent Transport System Knowledge QA										
Qwen-32B	94.3	93.5	93.9	94.8	94.2	95.2	91.7	84.2	71.2	80.5
Baichuan-32B	88.6	86.8	87.7	86.5	85.9	86.4	87.5	76.8	68.9	72.5
DeepSeek-R1-Distill-Qwen-32B	93.4	94.1	93.8	92.7	92.5	93.2	95.6	87.1	83.5	85.5
DeepSeek-R1-Distill-Llama-70B	93.9	94.5	94.5	92.8	93.2	92.7	94.1	87.4	84.2	86.1
Llama-3.3-70B	94.9	94.1	94.5	95.0	94.5	95.1	94.2	87.6	84.7	85.8
GLM-4-32B	90.1	88.5	89.3	89.4	88.8	89.7	88.6	78.9	69.7	74.1
ExpressMind-14B	**95.6**	**95.1**	**95.3**	**96.5**	**96.0**	**96.8**	**95.9**	**88.7**	**84.9**	**86.5**

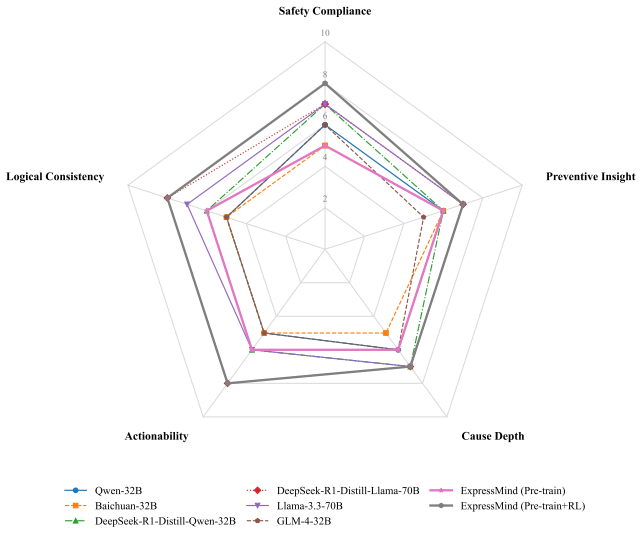


Figure 6: Performance of ExpressMind with RL Alignment.

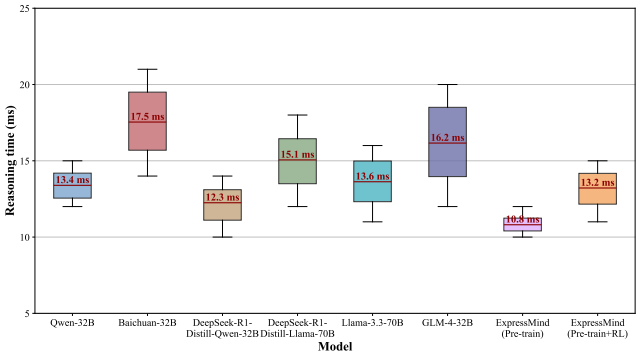


Figure 7: The Reasoning Time of the RL Alignment.

accuracy of 98.4% and a Short Answer F1-score of 88.5%, surpassing the strongest baseline, Llama-3.3-70B, by a significant margin. Furthermore, regarding the GPT-Score, which evaluates deep semantic understanding, ExpressMind maintains a high average of 85.7%, demonstrating superior competitiveness against reasoning-distilled models like DeepSeek-R1-Distill-Llama-70B. These results emphasize ExpressMind's expert-level proficiency and its ability to provide precise, logically coherent responses within the specialized expressway transportation domain.

4.3.2 RL Alignment Results. The results presented in Figure 6 indicate that ExpressMind (Pretrain+RL) consistently outperforms existing generalist baseline methods in the specialized task of expressway incident management strategy generation on five metrics, detailed in appendix A. Specifically, in domain-critical metrics such as Safety Compliance and Actionability, ExpressMind (Pretrain+RL) achieves scores in the range of 8.0–9.0, which is notably higher than general baselines like Llama-3.3-70B and Qwen-32B. The ablation study between the two ExpressMind configurations underscores the decisive impact of RL alignment: the base pretrained model without RL tuning yields the weakest performance, whereas its RL-aligned counterpart demonstrates a significant improvement. This enhancement is directly attributable to the model's better-aligned

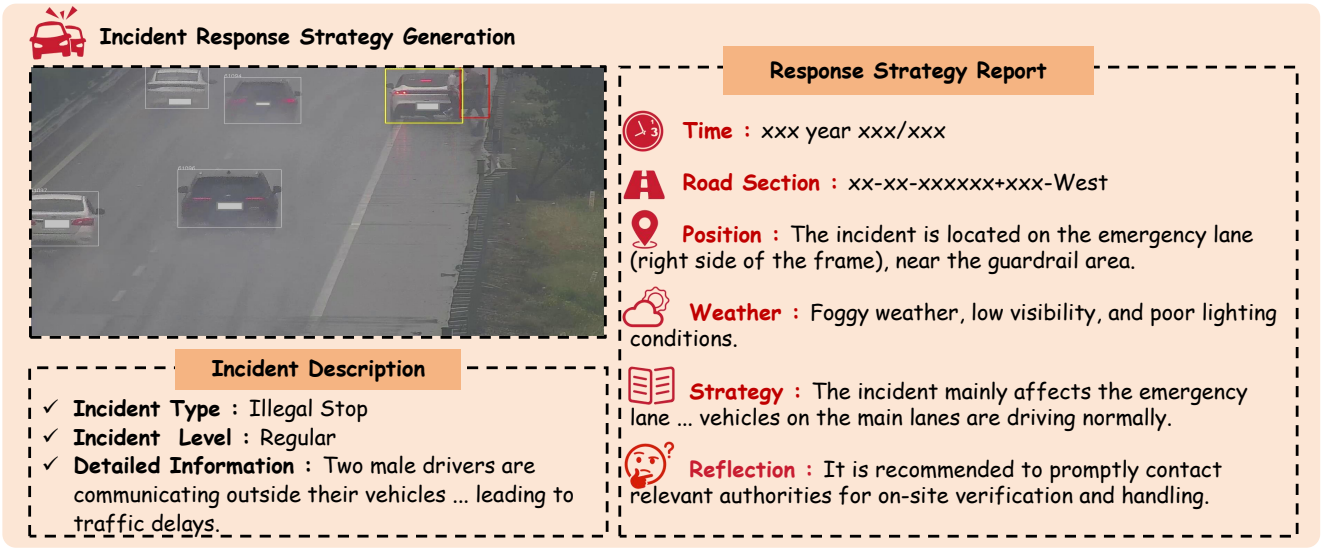


Figure 8: An Example of Incident Response Strategy Generation.

chain-of-thought reasoning with the required procedural knowledge. As shown in Figure 7, ExpressMind (Pre-train+RL) demonstrates exceptional deterministic performance in terms of reasoning efficiency. Experimental results indicate that its average inference latency is reduced to 13.2 ms, achieving a 24.6% acceleration compared to models like Baichuan-32B. More importantly, the tightly clustered distribution in the box plot reveals minimal latency jitter during instruction processing. This combination of low latency and high stability ensures reliable performance for time-sensitive applications, such as millisecond-level response requirements in smart expressway scenarios.

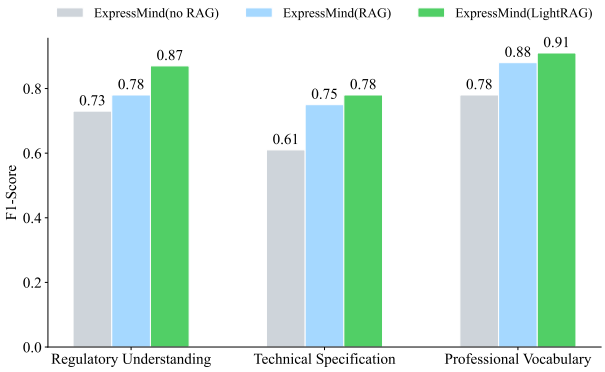
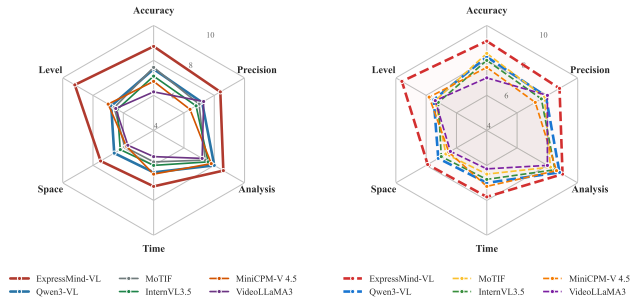


Figure 9: The Ablation Experiment of RAG.

4.3.3 Knowledge Retrieval Ablation Study. This section verifies the improvement in the generation capability of ExpressMind resulting from the use of the expressway knowledge base. We leverage LightRAG to query this repository. The evaluation dataset comprises 200 queries addressing complex traffic regulation comprehension and 100 queries focused on technical specifications.



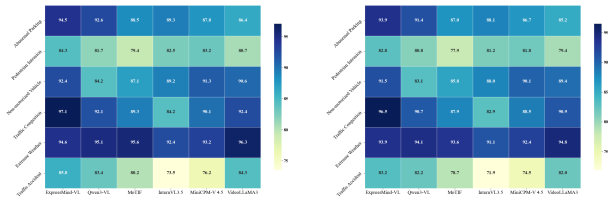
(a) Shandong Expressway **(b) Guangdong Expressway**

Figure 10: Results of Traffic Incident Detection.

We employ the F1-Score to evaluate the factual accuracy of the generated responses. As illustrated in Figure 9, the experimental results demonstrate that ExpressMind exhibits exceptional reasoning capabilities and robust retrieval performance. The Expressway knowledge base can increase the occurrence probability of professional vocabularies by 16.7%.

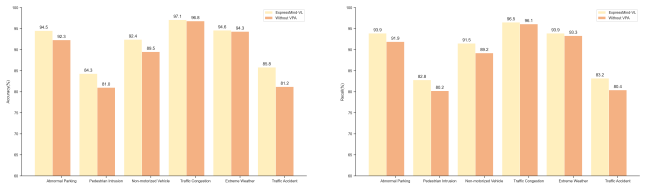
4.3.4 Scene Understanding Comparison. Built upon the ExpressMind backbone and a sophisticated cross-modal encoder, the MLLM, ExpressMind-VL, demonstrates exceptional proficiency in understanding traffic videos. In this study, we evaluate its performance against a range of leading MLLMs, including VideoLLaMA 3 [35], MiniCPM-V 4.5 [34], InternVL 3.5 [29], and Qwen3-VL [3].

To quantitatively evaluate the quality of the generated descriptions, we employ four standard automated metrics: BLEU-4 [23], ROUGE-L [18], CIDEr [27], and BERTScore [39]. Each metric assesses a distinct dimension of linguistic fidelity, ranging from lexical overlap to semantic similarity. To ensure reproducibility, we



(a) Detection Accuracy (b) Detection Recall

Figure 11: Results of Traffic Incident Detection.



(a) Accuracy (b) Recall

Figure 12: Ablation Results of VPA.

conducted a comprehensive evaluation using a dataset of 670 expressway surveillance videos input into all comparative models.

As presented in Table 2, the experimental results demonstrate that ExpressMind-VL significantly outperforms other generalist models in the descriptive accuracy of expressway scenes. The model exhibits a superior capability to interpret complex traffic scenarios, validating the effectiveness of our domain-specific multimodal alignment.

Table 2: Automated evaluation of quantitative results.

Model	LLM	BLEU-4	ROUGE-L	CIDEr	BERTScore
MoTIF	LLaMA2-7B	82.59	87.45	69.95	88.43
VideoLLaMA3	Qwen 2.5-7B	79.63	85.32	67.42	85.33
MiniCPM-V 4.5	LLaMA 3-8B	81.98	87.13	69.76	88.04
InternVL3.5	Qwen 3-38B	84.53	88.94	71.67	88.97
Qwen3-VL	Qwen 3-32B	84.85	**89.30**	72.96	89.18
ExpressMind-VL	ExpressMind-14B	**85.24**	89.25	**73.36**	**89.28**

Accurate detection and comprehensive analysis of traffic incidents are paramount for intelligent expressway operation. To rigorously evaluate these capabilities, we benchmark ExpressMind-VL against the high-performing Qwen3-VL-32B using a curated dataset of 200 traffic incident videos. The quantitative evaluation employs a multi-dimensional metric suite: Accuracy and Precision for event classification, and F1-scores to assess the semantic fidelity of descriptive texts concerning event severity (Level), causal analysis, spatiotemporal context (Time & Space), and congestion status (Queue). As illustrated in Figure 10, experimental results demonstrate that ExpressMind-VL exhibits significantly superior recognition and reasoning capabilities in traffic incident detection and analysis compared to the baseline. Furthermore, a representative example of Incident Response Strategy Generation, which translates these analytical insights into actionable decisions, is visualized in Figure 11.

To evaluate the practical performance of ExpressMind-VL in real-world expressway scenarios, we assessed its detection capability for six core types of traffic incidents on the Express-VQA dataset. As shown in Figure 11, the accuracy and recall rates of ExpressMind-VL have both exceeded 90% across all tasks, including abnormal parking, pedestrian intrusion, non-motorized vehicle detection, traffic congestion, extreme weather, and traffic incidents. The high precision and recall of ExpressMind-VL can be attributed to the multi-level technical strategies integrated during its development. Pre-training on high-quality video-text pairs established a robust foundation for cross-modal semantic alignment. As shown in Figure 12, the VPA mechanism explicitly enhances the contribution of visual features, ensuring the dominance of dynamic visual cues in the reasoning process. Furthermore, standardized prompt engineering reformulates the detection task into a structured text generation problem, enabling the model to naturally incorporate prior knowledge such as traffic rules into logical reasoning.

4.4 Deployment Analysis

The proposed ExpressMind has already been deployed in the Shandong Expressway Cloud Brain system. It demonstrates domain-specific comprehension in professional knowledge question-answering and the generation of emergency response strategies for traffic incidents. Furthermore, addressing user needs for customized ExpressMind-VL functionality, we have developed an ExpressMind-VL-based expressway incident monitoring and management platform for both Shandong and Zhejiang expressways. This study designs standardized prompt engineering and a video stream detection mechanism, enabling the model to retain 10 seconds of video upon detecting a traffic incident while simultaneously generating a structured incident analysis report. Through this approach, the system can automatically identify traffic incidents, analyze Traffic scenes, and produce response plans based on real-time conditions. Ultimately, ExpressMind-VL achieves a fully autonomous "perception-analysis-decision" processing pipeline for expressway traffic incidents, serving as an intelligent central hub of expressway. The demonstration of the system application is described in the Appendix D.

5 Conclusion

This study introduces ExpressMind, the first domain-specific MLLM designed for expressway scenarios. It is built through multiple technical innovations: a two-stage pretraining paradigm for domain knowledge internalization, a GRPO-enhanced RL framework for safety-critical reasoning alignment, a graph-augmented RAG mechanism for real-time spatiotemporal knowledge retrieval, and a VPA multimodal alignment module for deep video understanding. To support this work, we have open-sourced the first training dataset covering domain knowledge, incident CoT strategy reasoning, and multimodal incident detection VQA. The ExpressMind has been applied in top-tier expressway groups, serving as a representative application case of large models in the expressway domain.

In future work, we will focus on three key improvements: enhancing multimodal spatiotemporal reasoning for dynamic scenario understanding, strengthening chain-of-thought alignment in long-text analysis, and advancing model lightweighting for edge deployment.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Md Adnan Arefeen, Biplob Debnath, and Srimat Chakradhar. 2024. TrafficLens: Multi-Camera Traffic Video Analysis Using LLMs. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 3974–3981.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[5] Qi Cao, Ruiyi Wang, Ruiyi Zhang, Sai Ashish Somayajula, and Pengtao Xie. 2025. DreamPRM: Domain-Reweighted Process Reward Model for Multimodal Reasoning. *arXiv preprint arXiv:2505.20241* (2025).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv–2407.

[8] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[11] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779* (2024).

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[13] Senkang Hu, Zhengru Fang, Zihan Fang, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Tak Wu Kwong. 2025. Agentscomerge: Large language model empowered collaborative decision making for ramp merging. *IEEE Transactions on Mobile Computing* (2025).

[14] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[15] Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. 2025. Llmlight: Large language models as traffic signal control agents. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2335–2346.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[17] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5351–5362.

[18] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[20] Qian Ma, Hongliang Chi, Hengrui Zhang, Kay Liu, Zhiwei Zhang, Lu Cheng, Suhang Wang, Philip S Yu, and Yao Ma. 2025. Overcoming pitfalls in graph contrastive learning evaluation: Toward comprehensive benchmarks. *ACM SIGKDD Explorations Newsletter* 27, 2 (2025), 97–106.

[21] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. 2024. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems* 37 (2024), 23464–23487.

[22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[23] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. arXiv:1804.08771 [cs.CL] https://arxiv.org/abs/1804.08771

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[26] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[27] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[28] Peng Wang, Xiang Wei, Fangxu Hu, and Wenjuan Han. 2024. Transgpt: Multi-modal generative pre-trained transformer for transportation. In *2024 international conference on computational linguistics and Natural Language processing (CLNLP)*. IEEE, 96–100.

[29] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025).

[30] Zihe Wang, Haiyang Yu, Changxin Chen, Zhiyong Cui, Yufeng Bi, Yilong Ren, Zijian Wang, Delan Kong, Jing Tian, Shoutong Yuan, et al. 2025. MoTIF: An end-to-end multimodal road traffic scene understanding foundation model. *Communications in Transportation Research* 5 (2025), 100227.

[31] Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417* (2024).

[32] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. 2025. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 20370–20382.

[33] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[34] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154* (2025).

[35] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106* (2025).

[36] Jianqing Zhang, Xinghao Wu, Yanbing Zhou, Xiaoting Sun, Qiqi Cai, Yang Liu, Yang Hua, Zhenzhe Zheng, Jian Cao, and Qiang Yang. 2025. Htfllib: A comprehensive heterogeneous federated learning library and benchmark. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5900–5911.

[37] Siyao Zhang, Daocheng Fu, Wenzhe Liang, Zhao Zhang, Bin Yu, Pinlong Cai, and Baozhen Yao. 2024. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *Transport Policy* 150 (2024), 95–105.

[38] Tianlong Zhang, Xiaoxi He, Yuxiang Wang, Yi Xu, Rendi Wu, Zhifei Wang, and Yongxin Tong. 2025. FedMetro: Efficient Metro Passenger Flow Prediction via Federated Graph Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 5215–5224.

[39] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[40] Xingchen Zou, Yuhao Yang, Zheng Chen, Xixuan Hao, Yiqi Chen, Chao Huang, and Yuxuan Liang. 2025. Traffic-r1: Reinforced llms bring human-like reasoning to traffic signal control systems. *arXiv preprint arXiv:2508.02344* (2025).

A ExpressMind Setup and Metrics

A.1 Hyperparameter Settings

In this section, we detail the hyperparameter configurations for ExpressMind. To facilitate clear presentation and reproducibility, the specific settings for these phases are reported separately in Table 3.

Table 3: Hyperparameter Settings for Pre-training and GRPO Algorithm

Pre-training Hyperparameters	
Parameter Name	**Value**
Model Architecture	Qwen-14B
DeepSpeed Strategy	ZeRO Stage 3
Precision	bfloat16
Max Sequence Length	8192
Optimizer	AdamW
Weight Decay	0.1
Gradient Clipping	1.0
Learning Rate (lr)	1e-5
Lr Scheduler	Cosine
Warmup Ratio	0.05
Batch Size	256
Epochs	3
GRPO Algorithm Hyperparameters	
Parameter Name	**Value**
Learning Rate (lr)	1e-5
Lr Scheduler	Cosine
Warmup Ratio	0.05
Group Size	16
KL Coefficient	0.04
Clip Range	0.2
Visual Encoder Hyperparameters	
Parameter Name	**Value**
image resolution	224
patch size	14
hidden layers	24
attention heads	16

A.2 GPT-Score

To evaluate the performance of short-answer questions, we employ GPT-Score powered by GPT-4o. Functioning as a virtual domain expert, the model compares the Predicted Answer against the Question and Standard Ground Truth. It assesses factual accuracy and logical completeness to assign a quantitative score ranging from 0 to 100. The system prompt used for this evaluation is defined as Figure 13:

system_prompt: You are a 'Highway Operations and Management Expert' with 20 years of experience. Your task is to evaluate the accuracy of the candidate's (AI model) response, akin to grading a professional examination paper.

You must conduct the scoring strictly according to the following Evaluation Standards:

1. Strict Ground Truth Adherence: Do not merely check for semantic alignment. You must explicitly verify the precision of key data, regulatory articles, and handling procedures against the provided Ground Truth.

2. Key Point Verification: Extract core entities from the Ground Truth (e.g., speed limit values, emergency hotlines like 12122, procedural order like 'turn on hazard lights first') and verify if the candidate's response includes them.

3. Safety Penalty: Any errors involving traffic safety must result in severe penalty points.

4. Scoring: Assign a quantitative score from 0 to 100 based on factual accuracy and logical completeness.

Figure 13: The System Prompt of GPT-Score.

A.3 SFT-sysprompt

During the SFT phase, to ensure consistency in model outputs and adherence to specific interaction protocols, a unified system prompt was integrated at the beginning of each training sample. As illustrated in Figure , this prompt defines the model's core identity, task boundaries, and response style.

SystemPrompt=You are a Senior Engineer and Researcher in Intelligent Transportation Systems (ITS). You specialize in the bridge between classical traffic flow theory and AI-driven autonomous stacks.

1. Expertise Pillars: Standards：xxx,Intelligence: Spatio-temporal forecasting, RL-based decision making, and trajectory prediction.

2. Execution Guidelines: No Fluff: Skip all introductory pleasantries (e.g., "I'd be happy to help"). Start with the answer.

3. Reliability & Safety: Probabilistic Language: Use "95% confidence," "probabilistic bounds," or "asymptotic stability" instead of "perfect" or "absolute."

4. Interaction Templates: Policy Requests: Focus on Liability, Privacy, and Safety Standards.Engineering Requests: Prioritize real-time constraints, redundancy, and hardware deployment.

Figure 14: The System Prompt of SFT.

A.4 Evaluation Using LLM-as-a-Judge

A.4.1 RL Alignment. To evaluate the strategies generated by the LLM for the expressway incident response task, this study utilizes an "LLM-as-a-Judge" framework. Which assesses the strategies based on the following five key dimensions:

- **Safety Compliance:** Evaluates whether the plan prioritizes safety by explicitly including necessary safety distances, on-site protection, and personnel evacuation instructions to fundamentally prevent secondary accident risks.
- **Preventive Insight:** Validates whether the report strictly follows and completes the four-stage cognitive chain, which proceeds from **[Incident Description]** — **[Causal Inference]** — **[Response Strategy Formulation]** — **[Strategy Evaluation]**, ensuring a logically coherent and closed-loop process.
- **Logical Consistency:** Validates whether each cause identified in the Causal Inference section is addressed in the Strategy Formulation section.
- **Actionability:** Evaluates whether the accident response strategy is concise, clear, and non-redundant, ensuring high information density and high executability of the generated strategy for on-site personnel.
- **Cause Depth:** Examines the accuracy and depth of the incident cause analysis, as well as the correct use of professional terminology for expressway incident response, to ensure the foundational knowledge required for effective response planning.

A.4.2 Scene Understanding. To systematically evaluate the MLLMs' scene understanding capabilities on expressway surveillance videos, this study adopts the "LLM-as-a-Judge" evaluation framework and designs the following six core dimensions:

- **Accuracy:** Evaluates the overall correctness of the model's description of incidents, objects, and states in the video, serving as a baseline indicator of comprehensive performance.

- **Level:** Evaluates the model's ability to provide an overall summary and qualitative assessment of the video content, judging whether it can go beyond local details and accurately summarize the core events and overall situation.
- **Precision:** Evaluates the model's accuracy in fine-grained recognition tasks, including precise descriptions of specific targets such as vehicle types, traffic signs, construction facilities, and human behaviors.
- **Space:** Evaluates the model's understanding of static and dynamic spatial relationships between key entities (vehicles, personnel, facilities) in the scene, such as relative positions, lane occupancy, and driving directions.
- **Analysis:** Evaluates the model's ability to infer causality, impact, and potential risks of incidents, such as analyzing accident causes, predicting congestion spread, or assessing the effectiveness of response measures.
- **Time:** Evaluates the model's understanding of event sequencing and dynamic evolution processes, such as judging the order of actions and the continuity of state changes.

These six dimensions are evaluated by the LLM acting as a judge according to structured instructions, and their scores collectively constitute a systematic and comprehensive evaluation of the model's multimodal scene understanding capabilities.

B Implementation Details of Reward Functions

In this section, we provide the granular implementation details of the Structure-Knowledge-Semantics reward mechanism, including the construction of the expert vocabulary, the specific configurations of the embedding models, and the hyperparameter settings used in our experiments.

B.1 Structural Integrity Constraints (R_{struct})

To enable precise parsing of the model's CoT, we defined four special control tokens corresponding to the standard traffic incident disposal workflow. The Structural Integrity Reward R_{struct} performs strict string matching to verify the existence and order of these tokens.

Table 4: Definition of Reasoning Stage Delimiters

Stage Index (k)	Stage Name	Control Token (S_k)
1	Perception	`[Incident Description]`
2	Analysis	`[Causal Inference]`
3	Decision	`[Response Strategy Formulation]`
4	Reflection	`[Strategy Evaluation]`

The index function $\text{idx}(S_k)$ returns the character position of the first occurrence of token S_k in the generated string O. If a token is missing, $\text{idx}(S_k) = \infty$. The sequence check $\mathbb{I}(\text{idx}(S_1) < \cdots < \text{idx}(S_4))$ ensures the reasoning flow is logically valid.

B.2 Domain Knowledge Alignment (R_{know})

The domain alignment reward relies on a curated **Stage-Specific Expert Vocabulary ($\mathcal{V}_k$)**.

Vocabulary Construction. We constructed $\mathcal{V}_k$ by mining high-frequency professional terms from a corpus of 500+ real-world expressway traffic emergency plans and national standard documents (e.g., GB/T 29100-2012). We utilized TF-IDF to extract keywords and manually filtered them to ensure relevance to each specific reasoning stage. Examples are shown in Table 5.

Table 5: Examples of Expert Vocabulary $\mathcal{V}_k$ for Each Stage

Stage	Representative Keywords (Translated)
S_1: Perception	Multi-vehicle pileup, Hazardous chemical leakage, Occupying emergency lane, Visibility range, Traffic volume saturation, Fire spreading
S_2: Analysis	Secondary accident risk, Chain reaction, Brake failure, Fatigue driving, Lane capacity reduction, Danger radius
S_3: Decision	Remote diversion, Upstream interception, Green wave control, Air-ground coordination, Break-bulk transport, Gating control
S_4: Evaluation	Residual congestion, Rescue efficiency, Public sentiment monitoring, Secondary damage assessment

Perplexity Computation. To calculate the PPL penalty term $\text{PPL}(O)$, we utilize a frozen version of the Supervised Fine-Tuned (SFT) model as the reference. The threshold τ_{ppl} is dynamically set to the 95th percentile of the PPL distribution observed on the validation set, preventing the model from generating incoherent keyword lists.

B.3 Semantic Consistency (R_{sem})

The semantic consistency reward evaluates the strategic quality of the generated response.

- **Embedding Model (ϕ):** We employ **BGE-M3** (BAAI General Embedding), a state-of-the-art multilingual embedding model, to map text into dense vectors. We specifically use the `[CLS]` token embedding.
- **Focus Scope:** We extract only the content within the S_2 (Analysis) and S_3 (Decision) segments for embedding, denoted as $S_2 \oplus S_3$. This isolates the core logic from the generic description (S_1) or formatting text.
- **Reference Set ($\mathcal{D}_{ref}$):** For each query in the training batch, we retrieve top-K ($K = 5$) most similar historical expert cases from our vector database to form the local reference set $\mathcal{D}_{ref}$. This ensures the reward is calculated against contextually relevant ground truth.

C Benchmark

This research constructs a comprehensive evaluation benchmark for expressway operation, comprising three distinct dimensions:

The **Textual Knowledge QA** benchmark consists of 20,000 pairs of QA data. It focuses on the core dimensions of expressway operation, such as functional positioning and land-use control mechanisms. Additionally, it incorporates expert knowledge for emergency driving, providing precise instructions for handling vehicle instability, such as correcting rear-wheel skids on hazardous road surfaces. An example of this dataset's structure and content is illustrated in Figure 15.

The **Incident CoT** benchmark includes 300 samples designed to evaluate complex decision-making and causal reasoning. Each entry follows a structured logical flow. This simulates the chain of command in a smart expressway operarion center, moving from initial accident reporting to liability determination and the evaluation

Text Konwledge QA
Question1: What are the key points of urban rail transit planning compilation?
Answer1: Urban rail transit planning focuses on six core dimensions: research scope and timeframe, functional positioning, and network/station layout. It further addresses intermodal integration with other transit systems, prioritized research areas, and land-use control mechanisms to ensure the feasibility and implementation of the network.
Question2: How should you handle a situation where the rear wheels of a vehicle skid to the right on a muddy road?
Answer2: To correct a rear-wheel skid to the right, steer to the right. Avoid sudden braking or accelerating; instead, gently turn the wheel in the direction of the skid while lightly tapping the brakes. This helps the vehicle regain traction and realign. Remember the rule: steer with a rear skid and against a front skid to stay on path.

Figure 15: Example of Textual Knowledge QA.

of response strategies like lane closures. The structured reasoning process is shown in Figure 16.

Incident CoT Context
1. Incident Description:
At [Time], a rear-end collision involving ... occurred on the [position], ...
2. Causal Inference:
Upon analysis, [xxx] is determined to bear primary liability for the accident, ...
3. Response Strategy Formulation:
First, establish traffic control, ...
4. Strategy Evaluation:
The decision to close Lane 1 is appropriate.

Figure 16: Example of the Incident CoT Structured Data.

The **Express-VQA** is a multi-modal benchmark designed for expressway scene understanding, consisting of 670 real-world video segments captured from expressway surveillance systems. It evaluates the model's, particularly the LLM's, ability to understand and reason about expressway scenes, specifically in identifying and analyzing six typical expressway incidents: traffic accidents, congestion, road construction, abnormal parking, pedestrian intrusion, and debris clearance. Visual examples of these categories and the corresponding annotation style are provided in Figure 17.

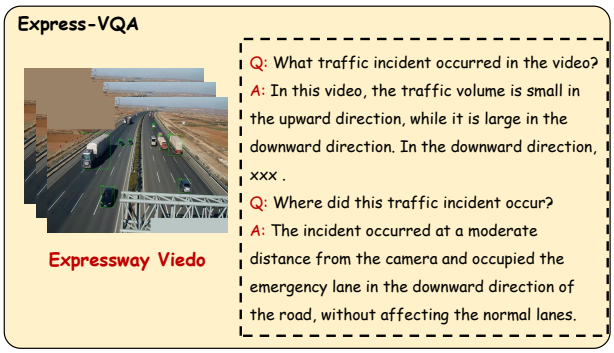


Figure 17: Visualization of the Traffic Incident VQA data.

D Application

The ExpressMind-VL intelligent operation system has been deployed in practical applications for multi-task scenarios on expressways. The visualization system, as shown in the figure 18, includes functions such as releasing warning information, summarizing traffic conditions, describing video events, and generating handling recommendations. We have deployed the system on the expressways in Shandong and Zhejiang provinces. In the intelligent management of Shandong expressways, ExpressMind-VL classifies the types and severity levels of traffic surveillance videos, and generates analytical reports along with handling strategies for traffic incidents. For the intelligent management of Guangdong expressways, ExpressMind-VL detects traffic events based on real-time video streams and produces structured textual descriptions for comprehension. The code, data, benchmark and the demonstration of the system application are available at: https://wanderhee.github.io/ExpressMind/.

Figure 18: Application of EpxressMind-VL.