Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
5(2): 315–336.
DOI: 10.3934/aci.2025018
Received: 20 October 2025
Revised: 14 December 2025
Accepted: 15 December 2025
Published: 24 December 2025

Research article

Multi-agent visual-language reasoning for comprehensive highway scene understanding

Yunxiang Yang, Ningning Xu and Jidong J. Yang*

Smart Mobility and Infrastructure Lab, College of Engineering, University of Georgia, Athens, GA 30602, USA

* **Correspondence:** Email: Jidong.Yang@uga.edu.

Academic Editor: Pasi Franti

Abstract: This paper introduces a multi-agent framework for comprehensive highway scene understanding, designed around a mixture-of-experts strategy. In this framework, a large generic vision-language model (VLM), such as GPT-4o, is contextualized with domain knowledge to generate task-specific chain-of-thought prompts. These fine-grained prompts are then used to guide a smaller, efficient VLM in reasoning over short videos, along with complementary modalities as applicable. This framework simultaneously addresses multiple critical perception tasks including weather classification, pavement wetness assessment, and traffic congestion detection, which achieve robust multi-task reasoning while balancing accuracy and computational efficiency. To support empirical validation, we curated three specialized datasets aligned with these tasks. Notably, the pavement wetness dataset is multimodal, combining video streams with road weather sensor data, highlighting the benefits of multimodal reasoning. Experimental results consistently demonstrate the strong performance across diverse traffic and environmental conditions. From a deployment perspective, the framework can be readily integrated with existing traffic camera systems and strategically applied to high-risk rural locations, such as sharp curves, flood-prone lowlands, and icy bridges. By continuously monitoring the targeted sites, the system enhances situational awareness and delivers timely alerts, even in resource-constrained environments.

Keywords: vision-language models; chain-of-thought prompt; multimodal foundation models; highway scene understanding; traffic congestion detection; pavement wetness assessment; road weather classification

1. Introduction

Multimodal foundation models, particularly vision-language models (VLMs), have emerged as powerful artificial intelligence (AI) agents capable of understanding and reasoning across diverse data modalities such as images, video, audio, and text [1–3]. These models are typically built on transformer-based architectures that integrate information from different modalities into a shared embedding space, enabling the generation of semantically rich, multimodal representations. A common design is to pair a visual encoder (e.g., Vision Transformer) with a language decoder or a unified encoder-decoder architecture pretrained on large-scale image-text or video-text datasets [4, 5]. This large-scale pretraining endows VLMs with strong generalization capabilities across a wide range of vision-and-language tasks such as image or video captioning, visual question answering, visual entailment, scene retrieval, among others.

Importantly, the adoption of multimodal foundation models marks a paradigm shift in infrastructure monitoring. Traditional systems often rely on dedicated physical sensors such as weather stations, embedded pavement sensors, or radar, which entail significant installation, calibration, and maintenance costs. In contrast, the multimodal models can leverage the existing video camera infrastructure (e.g., CCTV cameras) for robust visual reasoning [6]. These models are not only capable of assessing environmental attributes (e.g., wet pavement, snow accumulation, and visibility reduction) but also recognizing specific hazards such as fallen debris or stalled vehicles. Interactive frameworks like *SeeUnsafe* exemplify this potential by using VLMs to identify safety-critical events in large-scale traffic video data [7].

Moreover, recent research has explored the multi-task learning paradigms that unify diverse downstream tasks within a shared modeling framework [3]. These foundation models learn the transferable representations that span perception, prediction, and decision-making, enabling joint optimization and reducing the need for extensive data annotation. By leveraging shared knowledge across tasks, they offer a holistic and efficient approach to complex transportation scenarios. However, deploying such models in time-critical transportation systems remains challenging due to their large model size and computational overhead. To mitigate these constraints while preserving performance, researchers are exploring strategies such as model distillation, sparse expert routing, and task-specific chain-of-thought (CoT) prompting, each aiming to improve inference efficiency and adaptability in resource-constrained environments. Recent studies use a large teacher VLM or large language model (LLM) to enhance a smaller student model, typically through knowledge distillation or synthetic data generation [8–10]. In these approaches, the teacher's predictions or explanations are used to train a compact model that can later run independently at inference time. Such methods are effective when large labeled or teacher-annotated datasets can be constructed for each target task and when training resources are available. In contrast, our work does not train a new student model; instead, we keep a powerful VLM in the loop to generate domain-informed CoT prompts and then reuse these prompts to guide a smaller, off-the-shelf VLM at inference time. This design avoids task-specific training while still capturing some of the teacher's reasoning pattern, and it is particularly suitable for transportation agencies that may not have the capacity to retrain models but can benefit from orchestrating existing large and small VLMs in a multi-agent manner.

While deep learning has enabled considerable progress in traffic scene understanding, most of the existing approaches remain limited to single-task settings. This study aims to advance multi-task

visual understanding for comprehensive scene interpretation in the transportation domain. Specifically, we focus on both modeling and understanding of the road weather, pavement surface, and traffic conditions, enabling a more holistic and robust perception of real-world transportation environments.

1.1. Road weather understanding

It is critical to accurately assess weather conditions from roadside or in-vehicle perspectives for maintaining traffic safety and ensuring operational resilience. Conventional approaches are grounded in numerical weather prediction (NWP), which relies on data assimilation of satellite, radar, and in situ observations [11–15]. Recent advancements incorporate deep learning for localized weather understanding: [16,17] used convolutional neural networks (CNNs) and conditional generative adversarial networks (CGANs) for weather classification from road images, while Qing et al. [18] and Schmidt et al. [19] utilized long short-term memory (LSTM) and generative adversarial network (GAN) models for short-term forecasting of solar irradiance and cloud patterns. On the language side, foundation models such as ClimateBERT [20] and ClimateGPT [21] have been proposed for climate-focused text understanding.

However, these approaches either focus exclusively on visual inputs or treat weather as a standalone forecasting problem. They often lack real-time, road-level granularity or integration with traffic scene contexts, limiting their utility for real-time hazard detection and warning. In contrast, VLMs offer the ability to infer weather directly from traffic videos and reason about the impact on safety conditions (e.g., reduced visibility and road surface conditions), thus closing the gap between meteorological modeling and transportation decision-making.

1.2. Pavement wetness assessment

Timely detection of pavement wetness is essential for highway safety and operations, as surface water significantly reduces tire–pavement friction and increases the likelihood of hydroplaning, particularly at high speeds. These conditions not only elevate the crash risk but also complicate traffic management and roadway maintenance decisions. Traditional approaches rely on embedded sensors or road weather stations, which are accurate but sparse and expensive to maintain [22]. Modern deep learning systems have applied CNNs and segmentation networks to RGB or infrared images to classify wet, dry, snowy, or icy surfaces [23, 24]. Acoustic sensing systems, such as those proposed by [25], utilize tire-road interaction sounds to estimate surface wetness using support vector machines and logistic regression models.

Recent webcam-based methods leverage pretrained CNNs (e.g., ResNet18) to identify pavement conditions from roadside imagery [22]. Thermal imaging has also been explored to detect sub-surface anomalies and transient wetness features [23]. Hybrid approaches like road Maintenance systems using deep learning and climate adaptation (RMSDC) [26] fuse temporal sensor data using convolutional LSTM (ConvLSTM) for robust interpretable predictions. Despite the progress, these methods are typically static, infrastructure-specific, and short of adaptability across domains. They often require extensive re-labeling or fine-tuning when deployed in new regions or under different weather conditions.

1.3. Congestion analysis

Traffic congestion detection is another domain where deep learning methods have gradually replaced traditional models. Hybrid CNN-LSTM architectures [27] and encoder-based deep networks [28] have been developed to model spatio-temporal traffic dynamics from loop detectors and speed sensors. Vision-based methods have also been applied in enabling real-time congestion classification directly from traffic video feeds [29].

However, most of these models operate as task-specific detectors trained on specific datasets. They lack semantic understanding and struggle with context-sensitive reasoning (e.g., distinguishing construction-induced slowdown from other congestion scenarios). Recent reviews [30] advocate for more explainable and generalizable frameworks. Vision-language models have shown promise in this direction, offering semantic alignment between scene content and user-defined queries, enabling interpretable diagnostics and causality analysis of congestion [1, 7].

1.4. Contributions

In summary, this paper makes the following key contributions:

(1) *Unified VLM-Based Framework for Multi-Task Highway Scene Understanding*. We propose a VLM-driven framework that moves beyond task-specific models by jointly addressing three critical highway perception tasks (weather classification, pavement wetness assessment, and traffic congestion detection) within a single, unified system. This design improves adaptability and reduces the need for frequent retraining when operating across diverse conditions.
(2) *Mixture-of-Agents Reasoning with Domain-Informed CoT Prompts*. We introduce a mixture-of-agents strategy in which a large, general-purpose VLM is used to generate fine-grained, domain-informed CoT prompts tailored to each task. These prompts are then used to guide a smaller, computationally efficient VLM to reason over short video inputs, enabling scalable and edge-friendly multi-task inference without task-specific model redesign.
(3) *Multimodal, Real-World Datasets for Comprehensive Evaluation*. To support rigorous evaluation, we curate three task-aligned datasets collected from real-world deployments. In particular, the pavement wetness dataset integrates traffic video with road weather station data, enabling multimodal reasoning and providing a testbed that, to our knowledge, has not been previously explored using VLM-based approaches.
(4) *Empirical Validation of Collaborative VLM Agents*. Through extensive experiments, we demonstrate that the proposed collaborative VLM framework consistently outperforms simple prompting strategies and is particularly effective in complex and ambiguous scenarios, highlighting the potential of VLM agents as a practical foundation for comprehensive highway scene understanding.

2. Dataset

We collected the publicly accessible traffic video data from the states of Georgia, Virginia, and California. Camera locations were strategically chosen to cover urban, suburban, mountain, and coastal regions, ensuring a diverse set of highway scenes.

2.1. Category definition

For weather classification, we focused on three primary conditions: clear including sunny and cloudy, with no precipitation, rainy, and snowy (refer to Figure 1). For pavement wetness level assessment, we defined seven categories aligned with corresponding weather conditions: dry, rainy fully wet, rainy partially wet, rainy flooded, snowy fully wet, snowy partially wet, and snowy wet with icy warning. A detailed description of each category is provided in Table 1.

Figure 1. Examples from the road weather classification dataset.

Table 1. Pavement wetness level definitions and visual cues.

Category	**Key visual and contextual cues**
Rainy fully wet	Uniformly dark and glossy surface with consistent reflections. Tire sprays are visible across lanes; vehicles often leave moderate water trails.
Rainy partially wet	Mixed appearance with wet patches and dry zones. Water sprays are intermittent or limited to certain lanes. Some vehicles show water trails, others do not.
Rainy flooded	Standing or pooling water is clearly visible. Vehicles generate large water splashes and long, wide spray plumes. Water trails are thick and persistent.
Dry	Light-colored, matte surface with no visible moisture or reflections. No tire sprays or water trails; vehicle movement is clean and uninhibited.
Snowy fully wet	Entire surface is dark and wet from melted snow. Slush may appear near curbs or median dividers. Tire water spray may be visible. No snow patches.
Snowy partially wet	Uneven surface with a mix of wet, dry, or slushy zones. Residual snow or damp spots are visible. Minimal and inconsistent water sprays.
Snowy wet with icy warning	Surface has a faint shine or frosty gloss, suggesting potential black ice. This is often coupled with low temperature and high humidity conditions. Sprays are minimal or absent. Vehicles may move slowly with extra caution.

For congestion detection, we grouped traffic flow conditions into two categories: congested and unobstructed (refer to Figure 2). The detailed descriptions are given in Table 2. The distribution of video clips across weather conditions is presented in Table 3.

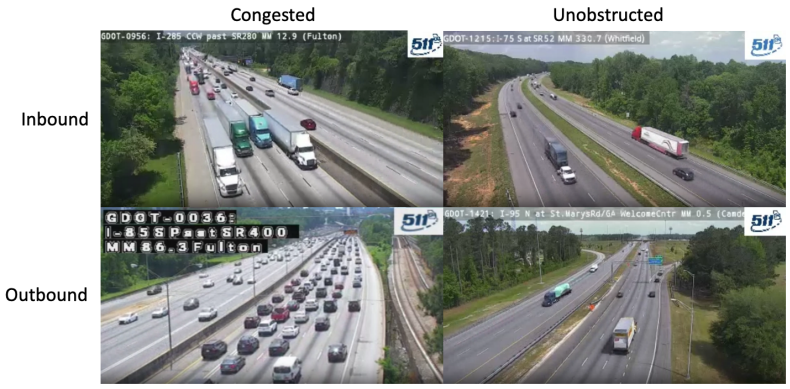


Figure 2. Examples from the congestion analysis dataset.

Table 2. Traffic flow condition definitions and visual cues.

Category	Key visual and contextual cues
Congested	Lanes are visibly full or nearly full of vehicles with minimal open space. Vehicle spacing is tight, making lane changes difficult. Motion is uneven; cars exhibit stop–go patterns, frequent surging and braking, or shock-wave movements. Multiple vehicles display delayed following, indicating disrupted flow.
Unobstructed	Traffic flows smoothly at or near posted speeds. Vehicles are evenly spaced. Motion is steady with little to no deceleration. The road appears open with no apparent disruptions to traffic flow.

Table 3. Summary of the weather video dataset.

Weather condition	Total videos
Clear	66
Snowy	21
Rainy	73
Total	**160**

In addition to the video data, we collected information from the nearest road weather stations, resulting in a multimodal dataset. Depending on the availability of sensor readings at the time of data collection, we distinguished between partial and full multimodal data. Partial multi-modal data includes date/time, current weather, weather precipitation, temperature high/low and elevation. In contrast, full multimodal data provides more detailed environmental context, including date/time, relative humidity, wind speed/direction, air/surface temperature, visibility, dew point temperature,

surface condition, and precipitation. For our downstream tasks, we primarily leveraged this multimodal data for pavement wetness assessment under snowy conditions, where the cross-modal reasoning provides the greatest benefit. A detailed summary of this dataset is shown in Table 4, and some examples are given in Figure 3.

Table 4. Summary of the pavement wetness video dataset.

Category	Total videos	Multi-modal data type
Rainy partially wet	51	46 Full, 5 Partial
Rainy fully wet	73	68 Full, 5 Partial
Rainy flooded	18	Full only
Snow partially wet	5	Partial only
Snow fully wet	9	Partial only
Snow wet with icy warning	21	Partial only
Sunny dry	66	Full only
Total	**243**	

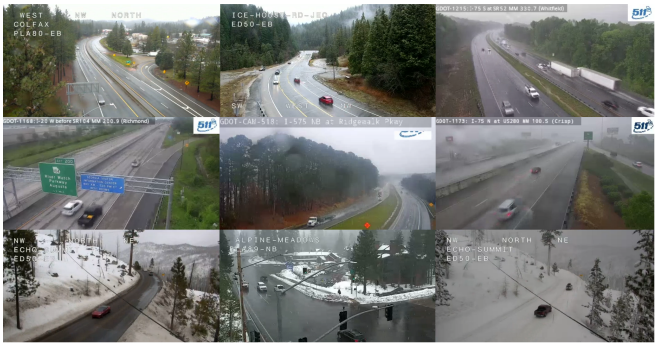


Figure 3. Examples from the pavement wetness assessment dataset. The leftmost column: 1st and 2nd rows—rainy partially wet; 3rd row—snowy partially wet. The middle column: 1st and 2nd rows—rainy fully wet; 3rd row—snowy fully wet. The rightmost column: 1st and 2nd rows—rainy flooded; 3rd row—snowy wet with icy warning.

It is important to note that for congestion analysis, rather than assigning a single label (e.g., congested or unobstructed) to an entire road segment within a video, we explicitly specified the traffic direction of the segment. This distinction accounts for the possibility of direction-dependent traffic patterns. Specifically, inbound refers to vehicles moving toward the traffic camera, while outbound refers to those moving away from it. The resulting dataset is summarized in Table 5.

Table 5. Summary of the traffic congestion video dataset.

Congestion level	Direction	Total videos
Congested	Inbound	10
	Outbound	18
Unobstructed	Inbound	20
	Outbound	16
Total		**64**

3. Methodology

This section introduces our proposed method, which leverages multiple agents to understand traffic scenes including weather, pavement wetness, and congestion conditions. The process begins by extracting sequential frames from a video input to retain temporal dynamics. An initial prompt, incorporating relevant domain knowledge, is constructed and provided to a VLM, referred to as Agent 1. In our experiment, we use GPT-4o [31] in this role. Agent 1 analyzes a scene based on the initial prompt and generates a detailed CoT [4] prompt that systematically addresses multiple aspects of the scene from the surrounding environment to vehicles. Depending on the downstream task, such as pavement wetness assessment under snowy conditions, multimodal data can also be ingested to enhance reasoning. Prompt tuning [5] is also applied to ensure accurate description of the scene is aligned with human observations and domain knowledge. The refined CoT prompt is then passed to Agent 2, which performs inference directly on video inputs and, if applicable, with associated multi-modal data. For Agent 2, we use QWEN 2.5-VL-7B [32], an open-source 7B-scale vision–language model that offers a trade-off between accuracy and computational cost on our target hardware (e.g., a single high-end GPU or embedded edge platforms). QWEN 2.5-VL-7B performs CoT-guided reasoning to generate the final output. This multi-agent framework is illustrated in Figure 4.

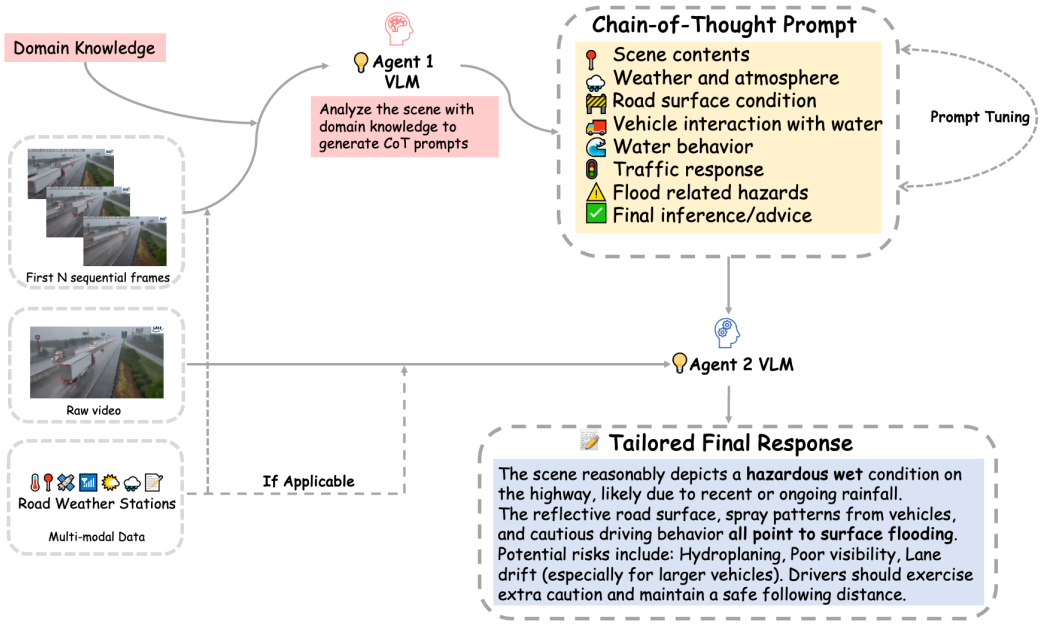


Figure 4. The multi-agent framework for highway scene understanding.

Although we instantiate the framework with QWEN 2.5-VL-7B as the small VLM, the overall approach is not tied to a specific model. The only requirements for Agent 2 are (i) the ability to accept visual inputs (sequential images or short clips) together with textual instructions and (ii) the capability to follow CoT-style prompts. In principle, other lightweight VLMs (e.g., alternative 3–8B multimodal models) could be substituted without modifying the multi-agent architecture; only the prompt templates may require minor retuning. Therefore, the improvements observed for CoT-guided reasoning over simple prompts are expected to be transferable to other small VLMs.

3.1. CoT prompts design and generation

As introduced previously, the definitions used as part of the prompt for Agent 1 (GPT-4o) integrate human observations and transportation domain knowledge tailored to specific types of scenes. For example, in the case of pavement wetness assessment under snowy conditions, these definitions are combined with sequential video frames and multimodal data (e.g., temperature high/low, humidity, wind speed/direction, dew-point temperature), and passed to Agent 1 to analyze the scene and acquire the fine-grained CoT prompt. This process is illustrated in Figure 5.

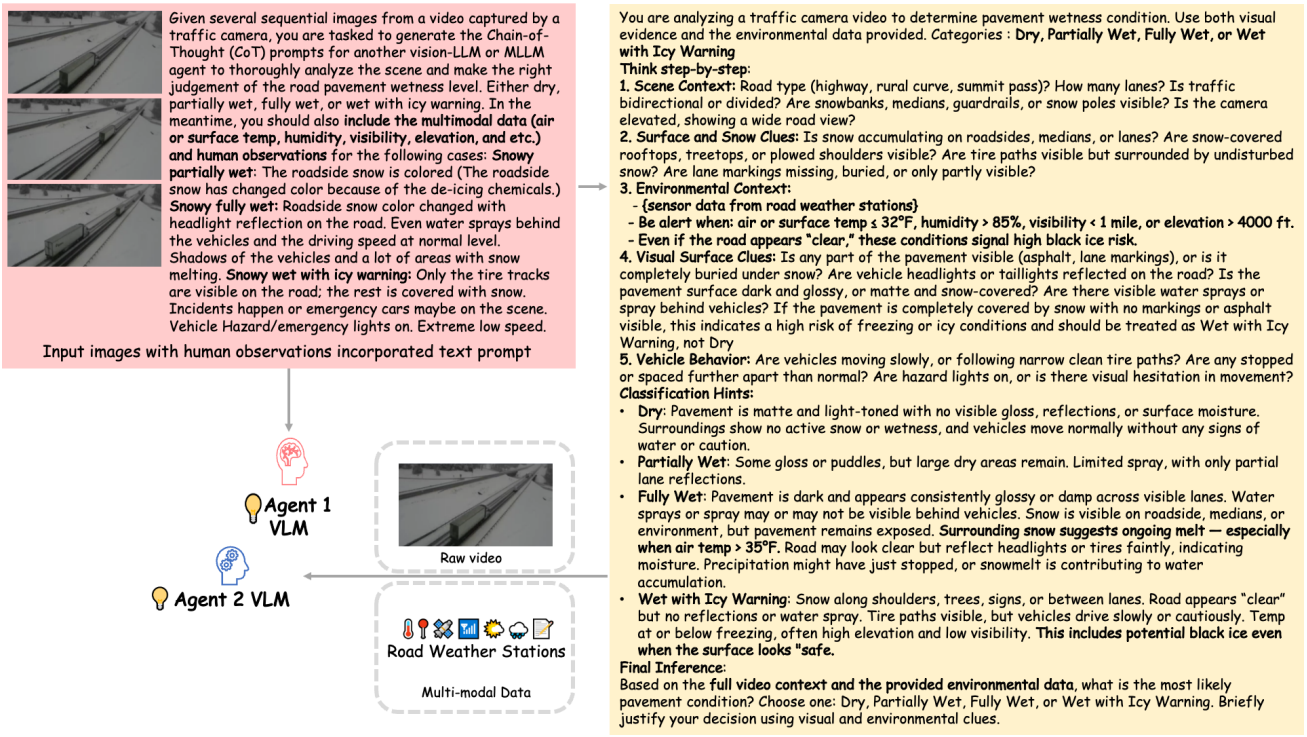


Figure 5. Example of CoT prompt generation for pavement wetness level assessment under snowy conditions.

Following a similar process, we generated CoT prompts for different downstream tasks. Specifically, Figures 6–8 show the generated CoT prompts by Agent 1 (GPT-4o) for road weather understanding, pavement wetness level assessment, and congestion analysis, respectively. To improve classification accuracy for pavement wetness levels, we introduced a threshold for identifying "fully wet" surface condition, which is defined as over 80% of vehicles per frame consistently have water sprays, mist, or strong reflections, which helps the model better distinguish between partially wet and fully wet.

We prioritized the "flooded" condition whenever clear visual cues are present (see Figure 7) and explicitly instructed the model: "If uncertain between fully wet and flooded, always choose flooded to reflect the potential real-world hazard." This safety-oriented directive biases the model toward conservative decision-making, helping ensure reliable detection of flooded scenes.

For congestion analysis, we designed tailored strategies to address the following challenges: (1) For inference efficiency, all video clips have a short length of 4–7 seconds; the acute signs of traffic congestion may not be obvious within a short video clip. (2) Each traffic camera has a different height

and angle, which brings different visual perspectives on traffic flow. (3) During normal peak hours (i.e., without accidents or road closures), the space headway can be short, but vehicles can still move relatively fast. (4) VLMs show a limited ability to accurately assess traffic flow speed or other related dynamic features, especially in such short video clips. Our proposed solution introduces a two-variable gating logic that incorporates *visual pressure* and *flow slow*, with the initial flow impression serving as a soft flag to provide contextual bias (e.g., "The flow appears smooth, but let me verify"). This design reduces over-reliance on visual cues alone for congestion detection. We further define three levels of *visual pressure* (strong, moderate, and weak) based on the number of visual congestion features identified. In parallel, the *flow slow* variable is evaluated as a Boolean flag (true or false) depending on evidence of flow disruption. The CoT prompt implementing this gating logic is illustrated in Figure 8.

You are analyzing a traffic camera video to determine the weather condition. Use both visual evidence and environmental clues to classify the scene. Categories: **Clear (includes sunny or cloudy with no active precipitation), Rainy, Snowy**
Think step-by-step:
1. **Scene Context:**
   - Is the camera elevated?
   - How much of the sky and surroundings are visible?
   - Are buildings, roadsides, or vegetation affected by weather (e.g., snow-covered, wet surfaces)? Does the scene show open highway, urban roads, or rural settings?
2. **Sky and Atmosphere:**
   - What does the sky look like — bright, blue, gray, or dark?
   - Are there visible clouds or signs of overcast conditions?
   - Is there fog, haze, or low visibility in the background?
   - Do shadows or sunlight reflections suggest clear conditions?
3. **Active Precipitation:**
   - Are rain streaks or snowflakes visible falling through the air?
   - Are there droplets on the lens or spray kicked up by vehicles?
   - Do headlights reflect off falling precipitation?
4. **Ground and Surroundings:**
   - Is the pavement dry, wet, glossy, snowy, or slushy?
   - Are rooftops, tree branches, or sidewalks covered with snow or ice?
   - Is snow accumulated on medians or roadsides?
   - Are there puddles, mud, or icy patches?
5. **Vehicle Behavior and Clues:**
   - Are headlights used during daylight, suggesting low visibility?
   - Are windshield wipers visibly in motion?
   - Is there spray or mist behind tires (common in rain or melting snow)?
   - Are vehicles moving more cautiously, as seen in snow or fog?

Classification Hints:
- Clear: Sky is blue, bright, or overcast without rain or snow. Pavement is dry or slightly reflective with no mist or spray. Vehicles move normally without using headlights or wipers.
- Rainy: Active rain visible, or visual cues like mist, tire spray, wipers, or wet pavement. Sky often appears dark gray or hazy. No snow or ice is present.
- Snowy: Falling snow, snow-covered surfaces, slushy or muddy pavement. Snow visible on vehicles, signs, or roadsides. Visibility may be reduced, and vehicles behave cautiously.

Final Inference: Based on the full video context and visible environmental clues, what is the most likely weather condition? Choose one: Clear, Rainy, or Snowy. Briefly justify your decision using specific visual and environmental evidence. Refer to the relevant classification hints.

Figure 6. Generated CoT prompt for road weather understanding.

You are analyzing a traffic camera video to determine the **pavement wetness level**. Use both visual evidence and environmental clues to classify the scene. Categories: **Dry, Partially Wet, Fully Wet, or Flooded.**

Think step-by-step:

1. **Scene Context:**
   - What type of road is shown (e.g., freeway, arterial)?
   - How many lanes are visible in each direction?
   - Are there overhead signs, guardrails, medians, vegetation, or entrance ramps present?
   - Is the camera angle elevated, and how much of the road and surroundings are visible?
2. **Weather and Lighting Conditions:**
   - Does the sky appear overcast, rainy, foggy, or dim?
   - Are there signs of active precipitation, such as rain streaks or visible raindrops?
   - Is visibility reduced in the distance due to weather conditions?
   - What time of day might it be, based on lighting and shadows?
3. **Road Surface Condition:**
   - Is the pavement highly reflective or glossy, suggesting water accumulation?
   - Are lane markings visible or partially obscured by water or glare?
   - Is there visual evidence of sheeting or pooling water on the surface?
   - Are there areas where standing water distorts reflections or textures?
4. **Vehicle and Surface Interaction:**
   - Do tires disturb shallow water or create visible splashes or spray?
   - Are spray trails consistent across multiple vehicles or lanes?
   - Do large trucks generate heavy mist, bow waves, or long water sprays?
   - Do any tires appear partially submerged?
   - Is there evidence of wake-like displacement or turbulent splashes from deeper water?
5. **Driver Behavior:**
   - Are vehicles reducing speed, increasing following distance, or changing lanes to avoid certain areas?
   - Are brake lights or hazard lights frequently activated?
   - Is there any erratic driving or lane drift suggesting poor traction or caution due to road conditions?

Classification Hints:

- **Partially Wet:** Pavement shows a mix of wet and dry patches with irregular surface glossiness. Spray is weak or limited to a few vehicles; reflections are intermittent. Lane markings remain mostly visible.
- **Fully Wet:** Pavement appears uniformly glossy across all visible lanes. Water sprays, mist, or strong reflections are observed consistently **(these visual cues present over 80% of all frames).** Lane markings are partially or fully obscured. Persistent mist or splash patterns are evident across multiple vehicles.
- **Flooded (prioritize classification if any of the following are observed):** Long, wide, and persistent water sprays behind vehicles — especially if spray lingers or extends across the lane. Standing or pooling water visibly present on lanes. Tires appear partially submerged, even briefly. Large trucks create bow waves, heavy splashes, or wake-like movement. Reflections appear distorted or refracted due to uneven water depth. Drivers slow down, activate hazard lights, or change lanes to avoid water.

Final Inference: Follow this strict rule-based decision process: If any Flooded indicators are clearly observed, classify the condition as Flooded, even if direct standing water isn't visible due to camera angle. If there is high risk of surface flooding, classify as Flooded. Only classify as Fully Wet if water is evenly distributed across the lanes, causing strong reflections and mist — but no flooding behavior. Use Partially Wet only when spray and reflections are weak, surface coverage is inconsistent, and no clear hazard is seen. If the pavement shows no signs of moisture, classify as Dry. Important: **If uncertain between Fully Wet and Flooded, always choose Flooded to reflect potential real-world hazard.**

Figure 7. Generated CoT prompt for pavement wetness level assessment under rainy conditions.

```
You are a highway traffic engineer. Your task is to analyze a short traffic camera video (3–5 seconds)
and classify the traffic in the INBOUND direction (vehicles moving toward the camera) as either:
Congested or  Unobstructed.
Congestion must involve both strong visual signs of traffic pressure (tight spacing, full lanes, low
maneuverability) and observable disruption in movement. Do not classify as Congested if traffic is
moving smoothly — even if spacing is tight.

Step 0: Initial Flow Impression (Not Final Decision)
Make a preliminary judgment based on first-glance impression:
-  Does the traffic appear uniform, fluid, and consistent across multiple lanes?
-  Is there no visible braking, hesitation, or pulsing in any direction?
-  If both are true, mark `initial_flow_smooth = True`, but continue to Step 1 and 2 for confirmation.
   If either is unclear, set `initial_flow_smooth = False`

Step 1: Evaluate Visual Congestion Features (Stricter)
Mark whether each of the following is clearly visible:
- Most or all lanes are filled and appear at capacity — minimal gaps, dense pattern
- Vehicle spacing is tight across multiple lanes, nearly bumper-to-bumper for > 1s
- Clusters of vehicles hold shape — packs of cars stay together without internal shifts
- No lane change attempts — vehicles appear boxed in across front/back/sides
- The road is full of vehicles all the way to the distance — no visible gaps or empty road ahead
If all 5 are true → `visual_pressure = Strong`; If 3–4, → `visual_pressure = Moderate`; Fewer than
3 → `visual_pressure = False`

Step 2: Evaluate Flow Disruption
Mark whether 3 or more of the following are true:
- Vehicles clearly move under 30 mph
- Heavy traffic moving and stopping in waves
- Braking or hesitation appears across multiple lanes
- Flow compresses in wave-like pulses along lanes
If 3+ signs set `flow_slow = True`; Otherwise, set `flow_slow = False`

Step 3: Final Classification Logic
-  If `visual_pressure = Strong`, classify as Congested
-  If `visual_pressure = Moderate` and `flow_slow = True`, classify as Congested
-  Else, classify as Unobstructed
> Do not use Step 0 alone for final classification. It only guides your impression — verification
depends on Step 1 and 2.

Final Classification:
Choose one and justify: Congested,  Unobstructed
```

Figure 8. Generated CoT prompt for congestion analysis.

4. Experimental results

We conducted extensive experiments to compare the performance of simple prompts (see Figure 9) versus CoT prompts presented in the preceding section. Exemplar results are shown in Figures 10–16.

Simple Prompt for Road Weather Classification
Given the video from the traffic camera, please make a judgement of the weather condition in the scene, then pick one from the following categories as your answer and justify: clear, rainy, or snowy.

Simple Prompt for Pavement Wetness Assessment without Multi-modal Data
Given the video from the traffic camera, please make a judgement of the road pavement wetness level, then pick one from the following categories as your answer and justify: dry, partially wet, fully wet, flooded or wet with icy warning.

Simple Prompt for Pavement Wetness Assessment with Multi-modal Data
Given the video from the traffic camera and the following weather conditions: **{env_summary}**, please make a judgment of the road pavement wetness level. Choose one from the following mutually exclusive categories as your answer:
- Dry
- Partially Wet
- Fully Wet
- Wet with Icy Warning

You may consider visual features such as reflections, spray, snow coverage, tire paths, and vehicle behavior,
along with temperature, humidity, elevation, and other environmental cues when making your decision.

Simple Prompt for Congestion Level Classification
You are analyzing a traffic camera video to determine the traffic flow speed in the **{outbound/inbound}** direction:
Return one of the following categories with high confidence:
- Congested
- Unobstructed

Figure 9. Simple prompts.

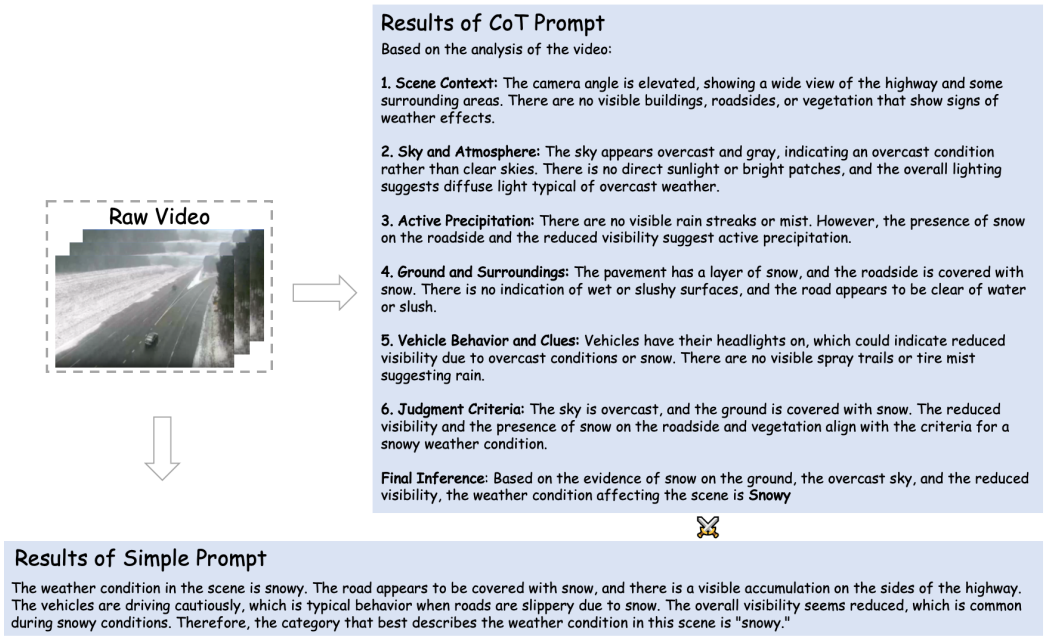


Figure 10. Results of road weather classification via CoT prompt.

4.1. Road weather classification

As shown in Table 6, road weather classification is relatively straightforward using either simple or CoT prompts, as VLMs can effectively leverage the abundant visual cues present in most conditions. For clear and snowy weather, performance differed only slightly: the model achieved over 90% accuracy for clear conditions and perfect accuracy for snowy conditions, where the visual evidence

of snow is highly distinctive. In contrast, the performance declined under rainy conditions, with some rainy scenes misclassified as clear when the visual cues were insufficiently pronounced.

Table 6. Accuracy comparison between simple and CoT prompts for weather understanding using QWEN 2.5-VL 7B model.

Weather Condition	Simple Prompt (%)	CoT Prompt (%)
Clear	92.42	95.45 (↑**3.03**)
Snowy	100.00	100.00 (—)
Rainy	73.24	71.23 (↓**2.01**)

4.2. Pavement wetness level classification

Prompting the model via CoT or simple prompt can result in significantly different results and accuracies. A simple case in Figure 11 shows the prediction results under a flooded condition. The simple prompt appears to make a judgment based on the first glance without careful attention. In contrast, The CoT prompt instructs the model to look through several aspects, from the scene to road surface, weather, vehicle and driver behaviors, particularly to inspect the existence of standing water. Eventually the model tags the video as flooded with the evidence of persistent tire splash-and-spray effects behind vehicles.

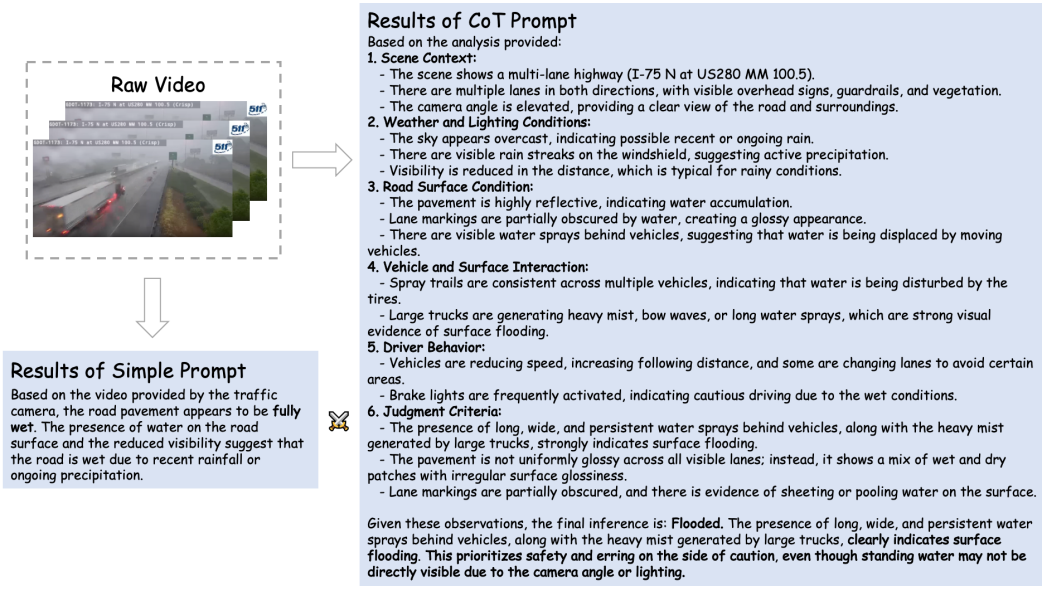


Figure 11. Results of pavement wetness level classification via CoT prompt and simple prompt under flooded condition.

Snowy partially wet condition presents a particularly challenging and ambiguous case because it can easily be misinterpreted as fully wet or icy, even by human observers without careful inspection. This is where the CoT approach outperforms a simple prompt. As shown in Figure 12, the model guided by a well-crafted CoT prompt arrived at the correct assessment by systematically examining multiple wetness indicators and ruling out conditions that did not meet the criteria.

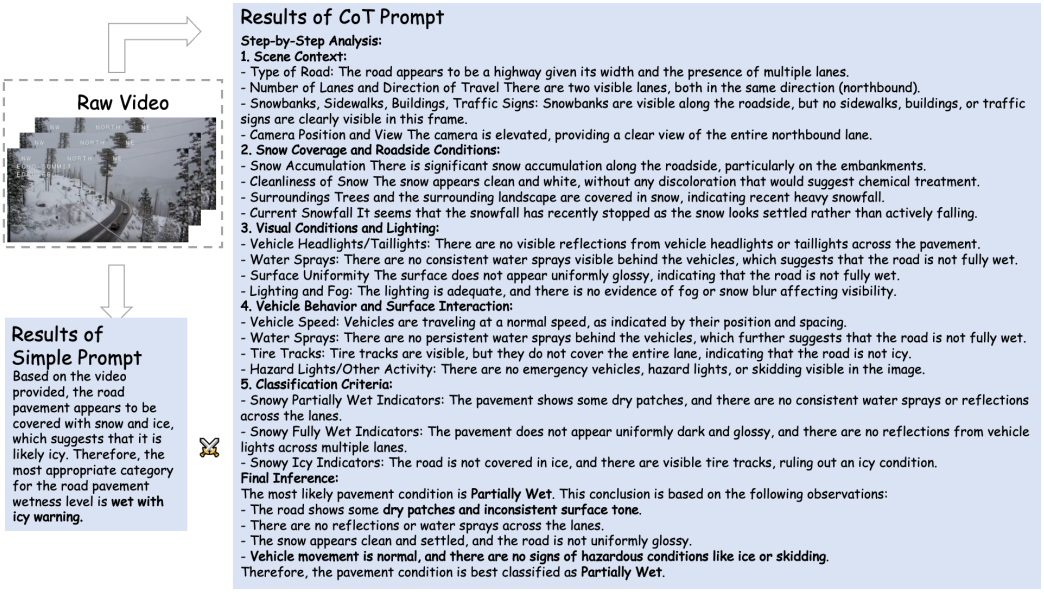


Figure 12. Results of pavement wetness level classification via CoT prompt and simple prompt under snowy partially wet condition.

4.3. Congestion level classification

Figure 13 illustrates that even when the initial flow appears smooth, the model can still identify congestion by leveraging the values of *visual pressure* and *flow slow*. This two-variable gating logic, *combined* with an initial soft flag, significantly enhances accuracy by striking a balance between the VLM model's strengths, such as identifying visual congestion features, and its limitations, particularly in estimating traffic flow speed. In contrast, the simple prompt relies on the initial impression of traffic flow without further verification, leading to an incorrect judgment.

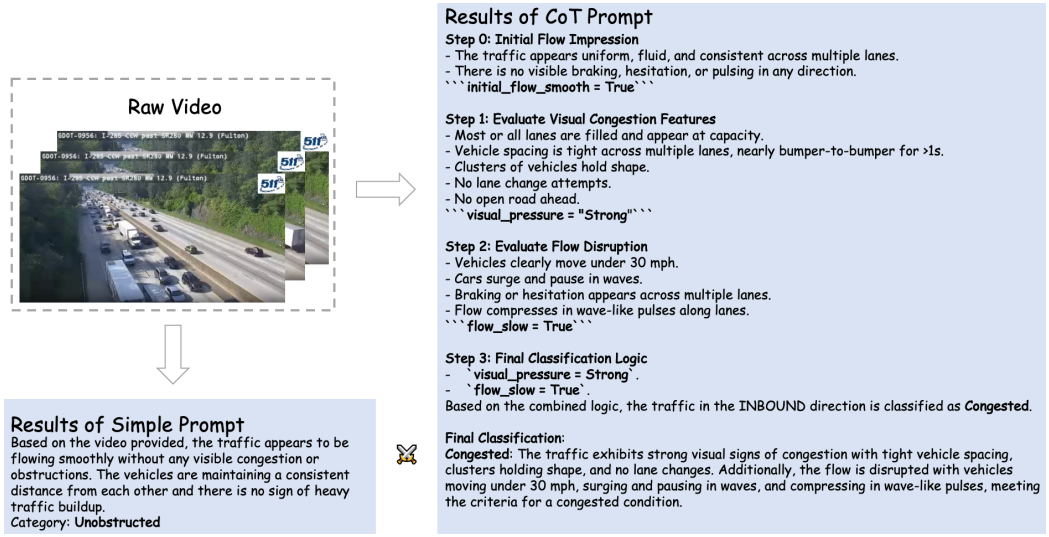


Figure 13. Results of congestion level classification via CoT prompt and simple prompt under congested condition.

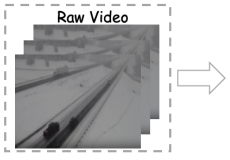
Table 7 presents improved performance for different directions and cases after implementing our carefully designed CoT prompt.

Table 7. Accuracy comparison between simple and CoT prompts for congestion analysis using QWEN 2.5-VL 7B model.

Congestion level	Direction	Simple prompt (%)	CoT prompt (%)
Congested	Inbound	40.00	100.00 (↑**60.00**)
	Outbound	72.22	100.00 (↑**27.78**)
Unobstructed	Inbound	60.00	80.00 (↑**20.00**)
	Outbound	81.25	100.00 (↑**18.75**)

4.4. Multimodal reasoning

We utilized the multimodal pavement condition dataset to highlight the benefits of multimodal reasoning. As shown in Figures 14 and 15, while the model can perform basic scene analysis based on snow coverage and other apparent visual cues, the inclusion of additional modalities substantially improves its ability to generate more comprehensive and accurate assessments. By incorporating data such as temperature, dew point, and humidity, the model is able to successfully identify icy conditions, a capability particularly valuable for transportation safety applications.



Results of CoT Prompt with Video Alone

Step-by-Step Analysis:
1. Scene Context:
- Type of Road: The image shows a multi-lane highway.
- Number of Lanes: There are at least three visible lanes, with two lanes in each direction.
- Direction of Travel: Vehicles are traveling in both directions.
- Roadside Conditions: Snowbanks are visible along the sides of the road, and there are no clear sidewalks, buildings, or traffic signs in the frame.

2. Snow Coverage and Roadside Conditions:
- Snow Accumulation: There is a significant amount of snow covering the roadside and median.
- Cleanliness of Snow: The snow appears relatively clean and white, without signs of chemical treatment or significant melting.
- Surroundings: Rooftops and trees in the background are also covered in snow, indicating recent snowfall.
- Current Snowfall: It does not appear that snowfall is currently active, as the snow looks undisturbed and settled.

3. Visual Conditions and Lighting:
- Vehicle Headlights/Taillights: There are no visible reflections of headlights or taillights across the pavement.
- Water Sprays: There are no distinct water sprays visible behind the vehicles.
- Pavement Surface & Lighting Conditions : The pavement appears predominantly white/snow-covered, with no uniform glossy film or dark wet patches visible. The lighting seems adequate, allowing for clear visibility of the road conditions.

4. Vehicle Behavior and Surface Interaction:
- Vehicle Speeds & Tire Tracks & Hazard Lights : Vehicles appear to be traveling at a normal speed for snowy conditions, which suggests that the road is not fully icy. Tire tracks are visible, but they are primarily indentations in the snow/ice layer, indicating that the road is not uniformly wet. Hazard Lights: There are no visible hazard lights activated by vehicles.
- Overall Movement: Vehicle movement is not excessively cautious or slow, which further supports the idea that the road is not fully icy.

5. Classification Criteria:
- Snowy Partially Wet Indicators: There are no clear visual evidence of wet patches or glossy surfaces across the entire road. The presence of tire tracks and the lack of water sprays suggest that the road is not fully wet.
- Snowy Fully Wet Indicators: The road does not exhibit a uniform dark and glossy appearance, nor are there consistent water sprays behind vehicles.
- Snowy Icy Indicators: The road is predominantly snow-covered, with minimal bare pavement visible. However, the tire tracks are primarily indentations in the snow/ice layer, and there is no significant evidence of liquid water or widespread gloss.

Final Inference:
The most likely pavement condition is **Partially Wet**. This is because while there is snow coverage, the road does not exhibit the characteristics of a fully wet or icy surface. The presence of tire tracks and the lack of water sprays indicate that the road is not uniformly wet, and the overall appearance suggests a mix of snow and some areas where the road is partially exposed. Therefore, the road is best classified as **Partially Wet**.

Figure 14. Pavement wetness level classification via CoT prompt without multimodal data under snowy fully wet condition.

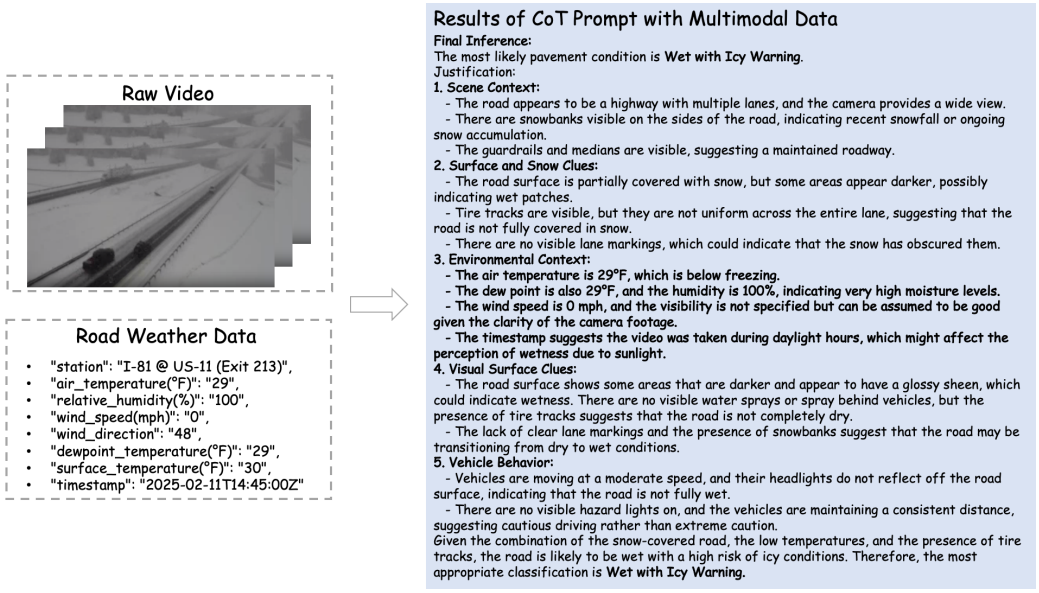


Figure 15. Pavement wetness level classification via CoT prompt with multimodal data under snowy fully wet condition.

We also evaluated the performance of multimodal reasoning on additional icy scenarios using the simple prompt. Surprisingly, even when the model could accurately infer conditions using only video data, the simple prompt sometimes produced incorrect results (see Figure 16). This suggests that while simple prompts may suffice in cases with strong and unambiguous visual cues, they fail when handling multimodal inputs, particularly when visual information is incomplete or ambiguous. Our results indicate that CoT prompting is essential for robust performance with multimodal data (as shown in Table 8); otherwise, the model may struggle when processing environmental parameters without sufficient visual context. For instance, the model misclassified "snowy fully wet" scenes as "partially wet," and similar mistakes were observed for rainy conditions. These errors may stem from the model's over-reliance on explicit visual cues (e.g., reflections, water spray, and road gloss) that are often subtle, inconsistent, or missing due to environmental and data limitations. In snowy scenes, slush and snow accumulation can obscure pavement texture, leading to misclassification of fully wet surfaces. Likewise, icy conditions can visually resemble fully wet roads, increasing the likelihood of false positives. In rainy scenarios, diminished sprays or weak headlight reflections may also cause false negatives. These findings highlight the challenges of relying solely on visual cues and underscore the importance of complementary modalities and thought-provoking prompt design for robust multimodal reasoning.

These challenges are further exacerbated by two key factors: (1) the short duration of the video clips, ranging from only 4 to 7 seconds, which limits the availability of temporal cues such as sustained water trails or subtle vehicle motion dynamics, and (2) the varying video resolution, with many clips being of lowresolution, hinders the model's ability to detect fine-grained visual features essential for accurately reasoning about pavement surface conditions.

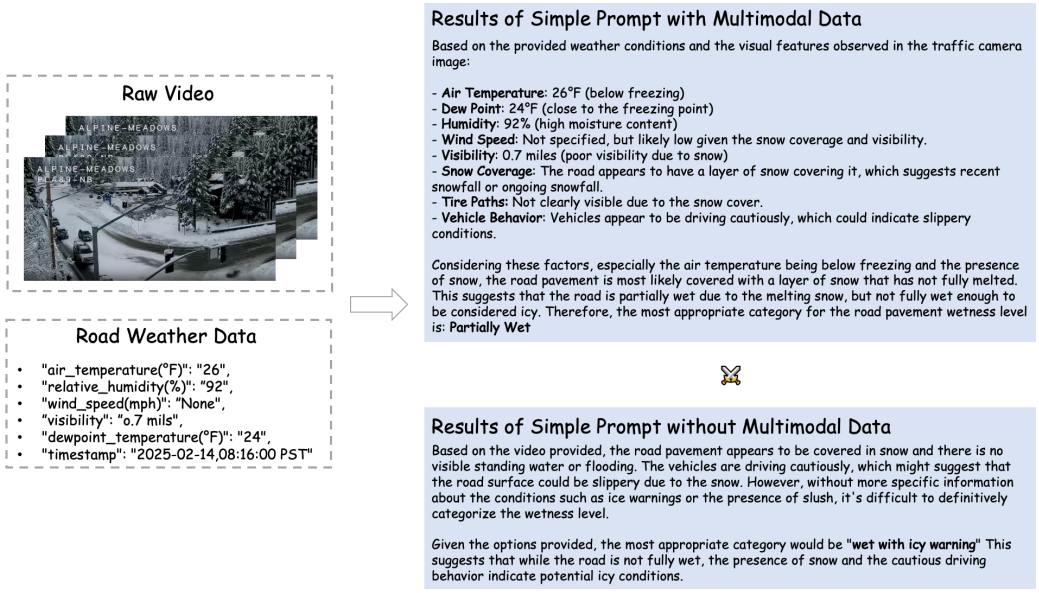


Figure 16. Comparison of pavement wetness level classification via simple prompt with/without multi-modal data under snowy wet with icy warning condition.

Table 8. Accuracy comparison between simple and CoT prompts across input modalities (Video alone versus multimodal data) using Agent 2 (QWEN 2.5-VL-7B model).

Input Type	Condition	Simple prompt (%)	CoT prompt (%)
Video alone	Rainy partially wet	37.25	58.82 (↑**21.57**)
	Rainy fully wet	23.29	57.33 (↑**34.04**)
	Rainy flooded	0.00	57.89 (↑**57.89**)
	Snowy partially wet	20.00	100.00 (↑**80.00**)
	Snowy fully wet	55.56	0.00 (↓**55.56**)
	Snowy wet with icy warning	76.19	14.29 (↓**61.90**)
	Sunny dry	83.33	95.45 (↑**12.12**)
Multimodal (Video and Sensor Data)	Snowy fully wet	10.00	100.00 (↑**90.00**)
	Snowy wet with icy warning	14.29	100.00 (↑**85.71**)

5. Conclusions and future work

In this work, we proposed a general multi-agent framework for comprehensive highway scene understanding. The framework leverages a large VLM to generate CoT prompts enriched with domain knowledge, which are then used to guide a smaller, efficient VLM in reasoning over video inputs, with complementary modalities as applicable. This design enables robust performance across multiple core perception tasks including weather classification, pavement wetness assessment, and traffic congestion detection.

To evaluate the effectiveness of the proposed framework, we curated three datasets. For the pavement wetness assessment task in particular, we constructed a multimodal dataset to demonstrate the benefits of multimodal reasoning. By leveraging carefully designed CoT prompts, the framework

achieves significantly improved reasoning performance and substantial gains in overall accuracy. This zero-shot, multi-agent approach leverages domain knowledge through a large VLM and unlocks the potential of small VLMs, offering a scalable and cost-effective solution for diverse transportation applications. Our framework can be readily integrated with the abundant network of existing traffic cameras, enabling large-scale deployment. In rural areas, where traditional sensor coverage is sparse, our method supports strategic monitoring by focusing on high-risk locations such as sharp curves, flood-prone lowlands, or icy bridges. By continuously analyzing scene conditions at these targeted sites, the system enhances situational awareness and provides timely alerts even in disconnected environments. Additionally, the ability to automatically detect congestion and road hazards allows transportation agencies to efficiently screen regional or statewide traffic camera feeds and quickly identify problem areas without intensive manual review.

Nonetheless, there remains a substantial room for improvement. Another direction is to distill or design a more compact VLM tailored to the target tasks. Such a lightweight model would be well-suited for edge deployment, facilitating the integration of advanced AI capabilities into existing traffic camera networks and enabling scalable and real-time intelligent scene understanding.

Despite these promising results, we acknowledge this study has several limitations. First, the datasets used here are relatively small, especially for congestion analysis, and cover a limited set of geographical regions and camera configurations. Second, the video clips are short (4–7 seconds) and most are lowresolution, which restricts the temporal cues and fine-grained visual details available to the VLMs. This constraint is representative of current practice in many deployments but may underestimate the potential of the proposed framework under higher-quality data. Third, we instantiate the small VLM (QWEN 2.5-VL-7B) and focus on accuracy as the primary metric in a fixed-split evaluation. A broader comparison across alternative lightweight VLMs, richer metric design, and more extensive statistical analyses is needed for assessing its reliability in real-world settings.

Future work can therefore focus on several directions. On the data side, we plan to expand the datasets to include additional states, roadway types, and seasonal patterns, as well as longer video segments that capture incident evolution and recovery. This will enable more comprehensive training and evaluation, including cross-location generalization and domain adaptation studies. On the modeling side, another avenue is to distill or design more compact VLMs tailored to highway scene understanding with explicit safety-aware objectives so that real-time inference can be reliably supported on edge devices. It will also be valuable to benchmark the proposed framework against a wider range of task-specific baselines and alternative small VLMs and to investigate streaming or event-triggered processing strategies for continuous camera feeds. Finally, future deployment-oriented work should consider formal risk analysis, human-in-the-loop monitoring, and integration with existing traffic management workflows to better understand how multi-agent VLM reasoning can support human-AI collaborative operational decision-making at scale.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported by the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program, through the Center for Regional and Rural Connected Communities (CR2C2) under Grant No. 69A3552348304.

Conflict of interest

Jidong J. Yang is an editorial board member for Applied Computing and Intelligence and was not involved in the editorial review and the decision to publish this article.

References

1. A. Keskar, S. Perisetla, R. Greer, Evaluating multimodal vision-language model prompting strategies for visual question answering in road scene understanding, *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2025, 1027–1036. https://doi.org/10.1109/WACVW65960.2025.00115
2. S. Park, C. Cui, Y. Ma, A. Moradipari, R. Gupta, K. Han, et al., Nuplanqa: a large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models, arXiv: 2503.12772. https://doi.org/10.48550/arXiv.2503.12772
3. S. Luo, W. Chen, W. Tian, R. Liu, L. Hou, X. Zhang, et al., Delving into multi-modal multi-task foundation models for road scene understanding: from learning paradigm perspectives, *IEEE Transactions on Intelligent Vehicles*, **9** (2024), 8040–8063. https://doi.org/10.1109/TIV.2024.3406372
4. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, et al., Chain-of-thought prompting elicits reasoning in large language models, *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, 24824–24837.
5. H. Yang, J. Lin, A. Yang, P. Wang, C. Zhou, H. Yang, Prompt tuning for generative multimodal pretrained models, arXiv: 2208.02532. https://doi.org/10.48550/arXiv.2208.02532
6. H. Gao, L. Zhang, Y. Zhao, Z. Yang, J. Cao, Application of vision-language model to pedestrians behavior and scene understanding in autonomous driving, arXiv: 2501.06680. https://doi.org/10.48550/arXiv.2501.06680
7. R. Zhang, B. Wang, J. Zhang, Z. Bian, C. Feng, K. Ozbay, When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis, *Accident Anal. Prev.*, **219** (2025), 108077. https://doi.org/10.1016/j.aap.2025.108077
8. Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, Z. Liu, Compressing visual-linguistic model via knowledge distillation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 1428–1438. https://doi.org/10.1109/ICCV48922.2021.00146
9. Y. Liu, C. Wu, S. Tseng, V. Lal, X. He, N. Duan, Kd-vlp: improving end-to-end vision-and-language pretraining with object knowledge distillation, *Proceedings of Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, 1589–1600. https://doi.org/10.18653/v1/2022.findings-naacl.119

10. Y. Yang, N. Xu, J. Yang, Structured prompting and collaborative multi-agent knowledge distillation for traffic video interpretation and risk inference, *Computers*, **14** (2025), 490. https://doi.org/10.3390/computers14110490
11. P. Lynch, The origins of computer weather prediction and climate modeling, *J. Comput. Phys.*, **227** (2008), 3431–3444. https://doi.org/10.1016/j.jcp.2007.02.034
12. F. Zhang, M. Zhang, J. Hansen, Coupling ensemble kalman filter with four-dimensional variational data assimilation, *Adv. Atmos. Sci.*, **26** (2009), 1–8. https://doi.org/10.1007/s00376-009-0001-8
13. P. Bechtold, N. Semane, P. Lopez, J. Chaboureau, A. Beljaars, N. Bormann, Representing equilibrium and nonequilibrium convection in large-scale models, *J. Atmos. Sci.*, **71** (2014), 734–753. https://doi.org/10.1175/JAS-D-13-0163.1
14. A. Geer, F. Baordo, N. Bormann, P. Chambon, S. English, M. Kazumori, et al., The growing impact of satellite observations sensitive to humidity, cloud and precipitation, *Q. J. Roy. Meteor. Soc.*, **143** (2017), 3189–3206. https://doi.org/10.1002/qj.3172
15. P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction, *Nature*, **525** (2015), 47–55. https://doi.org/10.1038/nature14956
16. Y. Shi, Y. Li, J. Liu, X. Liu, Y. Murphey, Weather recognition based on edge deterioration and convolutional neural networks, *Proceedings of 24th International Conference on Pattern Recognition (ICPR)*, 2018, 2438–2443. https://doi.org/10.1109/ICPR.2018.8546085
17. H. Zhen, Y. Shi, J. Yang, J. M. Vehni, Co-supervised learning paradigm with conditional generative adversarial networks for sample-efficient classification, *Appl. Comput. Intel.*, **3** (2023), 13–26. https://doi.org/10.3934/aci.2023002
18. X. Qing, Y. Niu, Hourly day-ahead solar irradiance prediction using weather forecasts by lstm, *Energy*, **148** (2018), 461–468. https://doi.org/10.1016/j.energy.2018.01.177
19. V. Schmidt, M. Alghali, K. Sankaran, T. Yuan, Y. Bengio, Modeling cloud reflectance fields using conditional generative adversarial networks, arXiv: 2002.07579. https://doi.org/10.48550/arXiv.2002.07579
20. N. Webersinke, M. Kraus, J. Bingler, M. Leippold, Climatebert: a pretrained language model for climate-related text, arXiv: 2110.12010. https://doi.org/10.48550/arXiv.2110.12010
21. D. Thulke, Y. Gao, P. Pelser, R. Brune, R. Jalota, F. Fok, et al., Climategpt: towards ai synthesizing interdisciplinary research on climate change. arXiv: 2401.09646. https://doi.org/10.48550/arXiv.2401.09646
22. M. Khan, M. Ahmed, Weather and surface condition detection based on road-side webcams: application of pre-trained convolutional neural network, *International Journal of Transportation Science and Technology*, **11** (2022), 468–483. https://doi.org/10.1016/j.ijtst.2021.06.003
23. S. Chandra, K. AlMansoor, C. Chen, Y. Shi, H. Seo, Deep learning based infrared thermal image analysis of complex pavement defect conditions considering seasonal effect, *Sensors*, **22** (2022), 9365. https://doi.org/10.3390/s22239365
24. Z. Wang, S. Wang, L. Yan, Y. Yuan, Road surface state recognition based on semantic segmentation, *Journal of Highway and Transportation Research and Development*, **15** (2021), 88–94. https://doi.org/10.1061/JHTRCQ.0000779

25. M. Kalliris, S. Kanarachos, R. Kotsakis, O. Haas, M. Blundell, Machine learning algorithms for wet road surface detection using acoustic measurements, *Proceedings of IEEE International Conference on Mechatronics (ICM)*, 2019, 265–270. https://doi.org/10.1109/ICMECH.2019.8722834
26. H. Elwahsh, A. Allakany, M. Alsabaan, M. Ibrahem, E. El-Shafeiy, A deep learning technique to improve road maintenance systems based on climate change, *Appl. Sci.*, **13** (2023), 8899. https://doi.org/10.3390/app13158899
27. A. Mihaita, H. Li, M. Rizoiu, Traffic congestion anomaly detection and prediction using deep learning, arXiv: 2006.13215. https://doi.org/10.48550/arXiv.2006.13215
28. Y. Liu, Z. Cai, H. Dou, Highway traffic congestion detection and evaluation based on deep learning techniques, *Soft Comput.*, **27** (2023), 12249–12265. https://doi.org/10.1007/s00500-023-08821-6
29. P. Chakraborty, Y. Adu-Gyamfi, S. Poddar, V. Ahsani, A. Sharma, S. Sarkar, Traffic congestion detection from camera images using deep convolution neural networks, *Transport. Res. Rec.*, **2672** (2018), 222–231. https://doi.org/10.1177/0361198118777631
30. T. Azfar, J. Li, H. Yu, R. Cheu, Y. Lv, R. Ke, Deep learning-based computer vision methods for complex traffic environments perception: a review, *Data Sci. Transp.*, **6** (2024), 1. https://doi.org/10.1007/s42421-023-00086-7
31. OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, et al., Gpt-4 technical report, arXiv: 2303.08774. https://doi.org/10.48550/arXiv.2303.08774
32. S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, et al., Qwen2.5-vl technical report, arXiv: 2502.13923. https://doi.org/10.48550/arXiv.2502.13923

AIMS Press