Context-Aware Perception: VLM-Augmented All-Weather Detection for Autonomous Driving

Johora Akter Polin
William & Mary
Williamsburg, VA, USA
japolin@wm.edu

Yichen Luo
William & Mary
Williamsburg, VA, USA
yluo11@wm.edu

Sidi Lu
William & Mary
Williamsburg, VA, USA
sidi@wm.edu

Abstract

Adverse weather, such as rain, fog, and snow, remains a major challenge for autonomous vehicles (AVs), degrading sensor reliability and compromising safety. To address this problem, we propose a novel perception pipeline augmented by a powerful vision-language model (VLM) to strengthen detection robustness and contextual understanding across diverse weather scenarios. The pipeline incorporates QwenVL to provide automatic weather labeling, semantically guided data augmentation, and adaptive sensor prioritization through weather-aware reasoning. We evaluate the approach on the BDD100K dataset using YOLOv10-M and RF-DETR, two strong detectors under adverse weather in our baseline comparisons. With VLM integration, both models achieve 8–12% gains in mean Average Precision in rainy and foggy conditions while incurring minimal additional latency. These results indicate that VLM-augmented perception can improve decision reliability and model explainability in autonomous driving. The findings underscore the value of context-aware, interpretable, and computationally efficient perception frameworks for achieving reliable all-weather autonomy.

CCS Concepts

• **Computing methodologies** → **Computer vision**; Scene understanding; • **Computer systems organization** → *Robotics*.

Keywords

Autonomous driving, vision–language model, adverse weather, perception robustness

ACM Reference Format:
Johora Akter Polin, Yichen Luo, and Sidi Lu. 2025. Context-Aware Perception: VLM-Augmented All-Weather Detection for Autonomous Driving. In *The Tenth ACM/IEEE Symposium on Edge Computing (SEC '25), December 3–6, 2025, Arlington, VA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3769102.3774638

1 Introduction

Autonomous vehicles (**AVs**) are a major advance in modern transportation, combining sensor technologies and artificial intelligence to perceive and react to their surroundings in real time. Yet adverse weather such as rain, fog, and snow continues to impair perception by reducing visibility, degrading sensor performance, and increasing uncertainty in safety critical detection on complex roads [16].

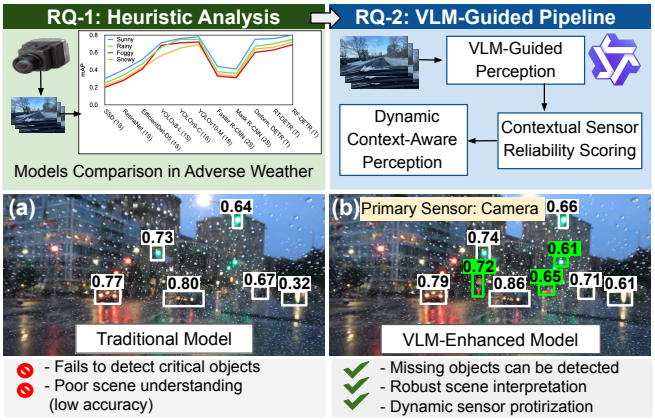

Figure 1: Overview of the proposed framework, illustrating heuristic analysis of adverse weather impact (RQ_1) and the vision–language model (VLM)-guided adaptive perception pipeline (RQ_2). The system integrates contextual sensor reliability scoring and a dynamic reasoning layer. Comparative visualization between (a) traditional and (b) VLM-enhanced models highlights improved scene understanding, object detection accuracy, and adaptive sensor prioritization.

Importance of Sensors for Perception. AVs employ cameras, LiDAR, radar, and ultrasonic sensors. Among these, cameras are typically the primary perception source because their high spatial resolution and rich semantic and color cues enable object detection, lane recognition, and scene understanding [19]. However, cameras are normally sensitive to illumination changes and weather-induced distortions, which degrade performance in low visibility or strong scattering. Improving robustness, therefore, requires adaptive fusion with complementary sensors and weather-aware reasoning, a central direction for a reliable and safe navigation system for AV.

Impact of Adverse Weather. Adverse weather poses critical challenges to perception in autonomous driving, as environmental interference directly impacts sensor reliability and system stability. Rain, fog, and snow cause visual distortions, reduce contrast, and obscure lane markings, hindering cameras from capturing clear images. These factors degrade object detection and scene understanding, increasing the risk of delayed decisions, unsafe maneuvers, and collisions [22, 27]. Recent crash reports indicate that between 2019 and mid-2024, nearly 4,000 AV-related crashes occurred in the U.S., with about 500 resulting in injury or death for adverse weather [8]. Thus, adverse weather undermines perception accuracy and threatens the safety, efficiency, and public trust essential.

Challenges & Motivation. AVs operate in dynamic environments where noise, vibration, and sensor instability degrade perception quality, especially for cameras. Adverse weather such as fog, rain, and snow further reduces visibility and increases uncertainty,

SEC '25, Arlington, VA, USA

ACM ISBN 979-8-4007-2238-7/2025/12
https://doi.org/10.1145/3769102.3774638

undermining reliable decision-making. Because AV perception is time-sensitive, end-to-end latency must remain below 164 ms to react safely to nearby obstacles [32]. These real-time constraints expose the limits of state-of-the-art models under diverse weather.

To address these challenges, we systematically evaluate leading detection frameworks across diverse weather conditions, showed in Figure 1, measuring accuracy, latency, and reliability. Guided by a state-of-the-art vision-language model (**VLM**), we design an adaptive, weather-aware, and efficient perception pipeline that improves robustness with minimal latency overhead even when adverse weather degrades sensor performance and compromises safety, thereby advancing safe autonomous driving.

Specifically, this paper aims to answer three main **research questions (RQs)**: *i*) **RQ$_1$:** To what extent do adverse weather conditions impair the effectiveness of camera-based detection in AVs, and which weather phenomena exert the greatest influence on system performance? *ii*) **RQ$_2$:** How can the decision-making of connected vehicles be improved to ensure reliability and safety under adverse weather conditions?

Contributions: To systematically address these research questions, this work makes the following key contributions:

- **C**$_1$: We design a novel VLM-augmented adaptive weather-aware perception pipeline that integrates semantic weather understanding, automatic annotation, and reliability reasoning. The proposed **QwenVL**-based framework enriches perception through contextual weather semantics, interpretable reasoning, and adaptive sensor prioritization based on the scenarios, ensuring robust performance across diverse weather conditions for AV (described in detail in Section 4).
- **C**$_2$: To address **RQ**$_1$, we conduct a comprehensive evaluation of state-of-the-art camera-based perception models (e.g., YOLOv10, RF-DETR,) under multiple adverse weather scenarios. The analysis quantifies degradation patterns in detection accuracy, latency, reliability and efficiency, identifying fog as the most detrimental weather phenomenon to visual perception (described in Section 3- Table 1).
- **C**$_3$: To address **RQ**$_2$, we demonstrate that integrating VLM-based reasoning with perception models enhances decision-making reliability under degraded visibility. The proposed VLM integration yields 8–10% accuracy improvement in rainy and foggy conditions with only marginal latency overhead, validating the effectiveness of context-aware, language-guided perception for safe and interpretable AV operation (described in detail in Section 4- Table 2).

2 Related Work

This section reviews representative prior work on the effects of weather on perception performance across conditions and identifies research gaps that hinder safe autonomous driving.

(1) Traditional State-of-the-Art Models: Camera-based object detection has advanced through three paradigms: one-stage, two-stage, and transformer-based models. One-stage detectors like SSD, RetinaNet, and YOLO (e.g., YOLOv8-L, YOLO-NAS) provide real-time performance but degrade under adverse weather due to blur, reflections, and low contrast [6, 29]. Two-stage detectors such as Faster R-CNN improve precision through region refinement but suffer higher latency and poor visibility robustness [26, 35]. Transformer-based models like Deformable DETR, RT-DETR, enhance contextual reasoning yet remain sensitive to weather-induced distortions [12]. Overall, these limitations emphasize the need for weather-aware, multimodal perception for reliability.

(2) Impact of Adverse Weather on Perception: AV perception relies on multimodal sensors LiDAR, cameras, and radar each responding differently to adverse weather. Multi-sensor fusion mitigates individual sensor weaknesses and enhances detection reliability, achieving high accuracy despite computational limits affecting real-time performance at higher speeds [5, 15, 22, 24, 36]. Rain degrades LiDAR and camera performance through reflections and scattering, while radar remains more robust [17, 18, 24]. Fog similarly reduces contrast and depth accuracy, impairing LiDAR and camera performance [21]; integrating radar with LiDAR improves stability in low-visibility scenarios [9]. Snow further complicates perception by reducing contrast and adding visual noise [28, 34], making CNN-based detectors struggle to generalize across weather domains. Rain, fog, and snow distort optical signals, necessitating adaptive weather-aware fusion frameworks [28]. Recent research emphasizes radar optical fusion and weather-specific augmentation to enhance robustness [2, 4, 10, 34]. Developing adaptive models that dynamically respond to environmental degradation is essential.

Research Gaps: Ensuring AV safety and reliability in adverse weather remains a major challenge. Despite progress, current perception systems lack robustness to real-world variability rain, fog, and snow reduce contrast, introduce reflections, and obscure objects, compromising navigation and safety. Most models are trained for clear or mildly degraded conditions and fail to generalize across diverse environments. Techniques like real-time fusion, dehazing, and weather-specific augmentation alleviate degradation but add computational cost and latency, limiting real-time deployment. The scarcity of large-scale, weather-diverse datasets further hinders generalization, as synthetic data rarely captures real-world complexity. Bridging these gaps is vital to developing adaptive, intelligent, and efficient systems capable of reliable, low-latency perception across all weather conditions, ensuring safety, scalability, and operational resilience in transportation for autonomous vehicles.

3 Heuristic Analysis on Adverse Weather

In addressing **RQ**$_1$, this section presents how different types of weather impairments affect the effectiveness of camera-based object detection and identifies which adverse conditions exert the most significant impact on perception performance.

3.1 Experiment Setup

3.1.1 Data Collection and Preparation. The framework utilizes the BDD100K dataset [33], comprising over 100,000 high-resolution driving frames across diverse weather, lighting, and geographic conditions. Each frame includes weather labels (W_i), temporal context (t_i), and sensor metadata (s_i) such as exposure, luminance, and GPS. Formally, $\mathcal{D} = (x_i, W_i, t_i, s_i)_{i=1}^{N}$, where x_i denotes the RGB frame. For photometric consistency, images are normalized as $x_i' = (x_i - \mu_x)/\sigma_x$, with μ_x and σ_x representing the dataset mean and standard deviation. This enhances feature stability under illumination variations. Figure 2 illustrates real-world weather scenarios

in different road structures, where green boxes mark detected objects. These examples demonstrate the dataset's diversity across illumination, visibility, and environmental conditions.

Figure 2: Representative camera frames under sunny, rainy, foggy, and snowy conditions, showing visual degradation in contrast, feature clarity, and object visibility. Green boxes mark detected objects, highlighting how weather variations impact model perception.

3.1.2 Hardware and Software Configuration. Experiments were conducted on a high-performance workstation optimized for machine learning and rendering. The system uses a 15 vCPU AMD EPYC 7543 (32 cores, 2.8 GHz) with dual NVIDIA RTX 3090 GPUs (24 GB each) for efficient deep learning and CUDA processing. It includes 80 GB RAM for large-scale data handling and runs on Ubuntu 18.04 LTS, ensuring reliability, stability and compatibility. This setup delivers the computational power and reliability needed for advanced AV applications for object detection experiments.

3.1.3 Evaluation Metrics. To comprehensively evaluate object detection models, three metrics are employed. These include mean Average Precision (mAP), Latency, and Frames Per Second (FPS). Each metric captures different aspects of model performance, ranging from detection accuracy to real-time feasibility and resilience under adverse weather conditions for both camera-based object detection in diverse scenarios.

(*i*) Mean Average Precision (mAP): Mean Average Precision is the standard metric for evaluating object detection accuracy across multiple classes. It is defined as $\text{mAP} = \frac{1}{N}\sum_{i=1}^{N}\text{AP}_i$, where N is the number of object classes, and AP_i is the Average Precision for class i, computed as the area under the Precision Recall curve. A higher mAP indicates a more accurate result and reliable detection, which is essential for safety-critical applications in autonomous driving.

(*ii*) Latency (L): Latency measures the average inference time (in milliseconds) required to process one input frame from the scenario: $L = \frac{\text{Total Inference Time}}{\text{Number of Frames}}$. Low latency is critical for ensuring timely responses in real-world driving. Excessive latency can delay braking or obstacle avoidance, reducing safety.

(*iii*) Frames Per Second (FPS): FPS represents the processing of the speed of a model and is the reciprocal of latency: $\text{FPS} = \frac{1000}{L}$, where L is latency measured in milliseconds. A higher FPS ensures real-time, stable, efficient operation, and control, which is indispensable for AVs operating in dynamic environments.

3.2 Result Analysis (RQ_1)

The results in Table 1 present object detection performance under diverse weather conditions. Metrics mAP, latency, and FPS reflect accuracy, cost, and efficiency. The analysis of the Table 1 highlights key trends, where 1S, 2S, and T denote one-stage detectors , two-stage detectors, and transformer-based detectors, respectively.

These comparisons provide a foundation for evaluating trade-offs between real-time analysis of performance and perception.

(1) One-Stage Detectors (1S): (1) One-Stage Detectors (1S): Among one-stage baselines in Table 1, EfficientDet-D5 balances precision (mAP ≈ 0.52⟶0.43) and throughput (≈15 FPS). In contrast,

Table 1: Performance of camera-based object detectors under different weather conditions.

Model (Type)	Sunny	Rainy	Foggy	Snowy
Mean Average Precision (mAP)				
SSD (1S) [31]	0.30	0.25	0.20	0.22
RetinaNet (1S) [3]	0.40	0.34	0.28	0.30
EfficientDet-D5 (1S) [14]	0.52	0.47	0.41	0.43
YOLOv8-L (1S) [1]	0.71	0.66	0.68	0.56
YOLOv9-C (1S) [25]	0.76	0.75	0.71	0.65
YOLOv10-M (1S) [30]	0.78	0.75	0.72	0.69
Faster R-CNN (2S) [20]	0.44	0.38	0.33	0.35
Mask R-CNN (2S) [7]	0.41	0.36	0.31	0.33
Deform. DETR (T) [13]	0.75	0.67	0.60	0.64
RT-DETR (T) [23]	0.76	0.70	0.63	0.66
RF-DETR (T) [11]	0.80	0.75	0.69	0.72
Latency (ms)				
SSD (1S)	12	14	16	15
RetinaNet (1S)	142	150	158	155
EfficientDet-D5 (1S)	67	72	78	74
YOLOv8-L (1S)	10	10	10	10
YOLOv9-C (1S)	12	13	13	11
YOLOv10-M (1S)	9	10	11	11
Faster R-CNN (2S)	100	108	115	110
Mask R-CNN (2S)	280	295	310	300
RT-DETR (T)	15	16	18	17
RF-DETR (T)	6	7	8	7.5
Frame Rate (FPS)				
SSD (1S)	83	71	62	66
RetinaNet (1S)	7	7	6	6
EfficientDet-D5 (1S)	15	14	13	14
YOLOv8-L (1S)	100	94	90	85
YOLOv9-C (1S)	100	90	95	80
YOLOv10-M (1S)	111	100	91	83
Faster R-CNN (2S)	10	9	9	9
Mask R-CNN (2S)	4	3	4	4
RF-DETR (T)	71	66	62	60

SSD-VGG and RetinaNet exhibit sharp accuracy drops under fog and snow, showing limited adaptability despite low latency. The *YOLO family* maintains strong resilience and real-time stability across all conditions. Notably, **YOLOv10-M** achieves the highest mAP (0.78) while sustaining over 80 FPS, benefiting from its anchor-free design, adaptive feature extraction, and optimized attention. YOLOv9-C and YOLO-NAS also show robust detection under degraded visibility, confirming the YOLO family's readiness for deployment-critical perception for autonomous vehicle in diverse scenarios.

(2) RCNN Family (2S): RCNN models attain moderate mAP in clear weather but suffer extreme latency (up to 2000 ms/frame, <1 FPS). While Faster R-CNN improves slightly, region-proposal mechanisms remain unsuitable for real-time driving, confirming their inefficiency for dynamic detection.

(3) Transformer-Based Detectors: Transformer architectures exhibit strong adaptability via global attention and semantic embedding. **RF-DETR** delivers the highest mAP (0.80) and lowest latency (6 ms), outperforming RT-DETR and Deformable DETR. Its attention fusion preserves spatial coherence under occlusion and low contrast, underscoring transformers' superiority for weather-aware multimodal perception in the dynamic urban road.

3.3 Key Observations and Discussion

In this section, we present our three critical observations to answer RQ_1, and also discuss the observed trends.

★ **Observation$_1$:** *Among adverse weather conditions, fog imposes the most significant degradation on camera-based perception for autonomous vehicle applications* (Table 1).

Discussion: This observation also answers **RQ_1**, fog most severely impacts camera-based perception by scattering and absorbing light,

sharply reducing image contrast and visibility. Suspended droplets blur edges and wash out color gradient features crucial for detectors like R-CNN, making object boundaries indistinct. The resulting scattering also disrupts depth cues and lowers the signal-to-noise ratio, leading to missed or false detections. Thus, fog's optical interference directly weakens the visual features essential for accurate perception in the complex urban road.

★ **Observation$_2$:** *RF-DETR and YOLOv10-M demonstrate superior robustness and contextual reasoning under weather-induced degradation, maintaining high detection accuracy and stable performance in diverse real-world scenarios (Table 1).*

Discussion: Among the evaluated models, *RF-DETR* achieves the highest accuracy, across sunny to foggy conditions with minimal latency and high FPS, confirming the real-time efficiency of transformer-based architectures. One-stage models such as *YOLOv10-M* offer the best speed, accuracy trade-off, with mAP degradation under adverse weather (rain, fog, snow) limited to 10–12%, notably lower than SSD or Faster R-CNN. This stability indicates that newer architectures leveraging transformer attention and advanced feature extraction better withstand environmental noise and visibility loss, delivering superior generalization and efficiency.

★ **Observation$_3$:** *There is a trade-off between accuracy and efficiency across detection architectures* (Table 1).

Discussion: Latency and frame rate differ notably across architectures, revealing a trade-off between detection accuracy and computational efficiency. Two-stage models suffer from high latency and low FPS due to region proposals, limiting real-time applicability. In contrast, one-stage detectors like YOLOv10-M and EfficientDet-D5 offer faster inference with moderate accuracy, while transformer-based models such as RF-DETR balance both, achieving high mAP with acceptable delay. These results highlight how architectural design directly governs deployability in safety-critical AV perception.

Overall, merging **YOLOv10-M** and **RF-DETR** within the proposed VLM-Guided Adaptive Weather-Aware Perception Pipeline for addressing **(RQ$_2$)** ensures both temporal efficiency and all-weather robustness, validating the dual-model training framework for autonomous perception. This integration achieve balanced performance across accuracy, and environmental adaptability.

4 VLM-Guided Adaptive Weather-Aware Perception Pipeline

The proposed pipeline integrates a vision–language model (QwenVL) and a reasoning layer to achieve adaptive perception under varying weather conditions. QwenVL is employed for its ability to jointly interpret visual and textual cues, enabling semantic understanding of weather-related contexts. By generating descriptive weather annotations and guiding realistic data augmentation, it enhances scene comprehension and supports adaptive, context-aware perception across diverse environmental conditions. Figure 3 presents the full pipeline, highlighting from visual input to reliability perception output of the whole work for an autonomous vehicle.

4.1 VLM-Guided Perception Module

4.1.1 Context-Aware Annotation. To achieve semantically consistent and context-aware environmental understanding, this study

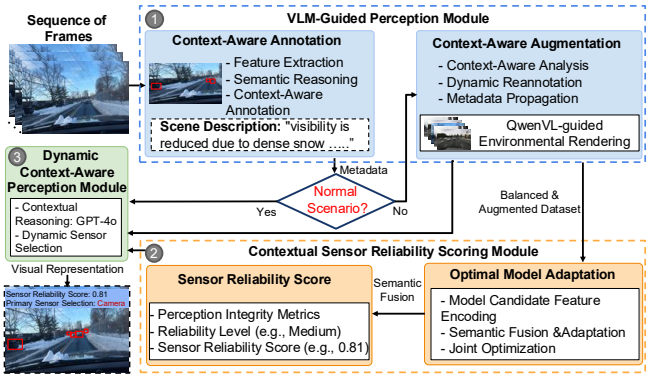

Figure 3: Overall architecture of the proposed VLM-Guided Adaptive Weather-Aware Perception Pipeline. It begins with Context-Aware Annotation, which generates semantic weather descriptions from raw scene inputs. Under normal conditions, the perception process proceeds directly, while in adverse weather, Context-Aware Augmentation leverages QwenVL-guided environmental rendering to generate realistic weather variants while maintaining label and metadata consistency. Optimal Model Adaptation integrates contextual cues into sensor data, and Sensor Reliability Scoring quantifies visual trust under both normal and degraded visibility. Finally, the Dynamic Context-Aware Perception Layer, powered by GPT-4o reasoning, fuses multimodal features to enable adaptive sensor selection and interpretable perception across diverse weather scenarios.

employs the QwenVL model, a vision language transformer pretrained on large-scale data (BDD100K described in Section 3.1.1). QwenVL aligns visual embeddings with linguistic features for multimodal reasoning over weather semantics. Given an input frame x_i, it predicts $y_i = f_{\text{QwenVL}}(x_i) = \{\hat{W}_i, \hat{V}_i, \hat{P}_i\}$, where $\hat{W}_i$ denotes the inferred weather label, $\hat{V}_i$ quantifies visibility, and $\hat{P}_i$ represents the *Perceptual Alignment Confidence (PAC)*. The weather type is determined via $\hat{W}_i = \arg\max_{W \in \mathcal{W}} P(W \mid x_i, \theta)$, where $\mathcal{W} = \{\text{sunny, rainy, foggy, snowy}\}$ and θ are model parameters. PAC is estimated using a cross-modal alignment score $A_i = \sigma(F_v^\top W_a F_l)$ between visual features F_v and linguistic embeddings F_l, with W_a as a learned projection and $\sigma(\cdot)$ as the sigmoid function. The overall confidence is $\hat{P}_i = \lambda_1 A_i + \lambda_2 e^{-\kappa \,\text{Var}(L_i)}$, where $\text{Var}(L_i)$ denotes luminance variance, κ is a scaling factor, and $\lambda_1 + \lambda_2 = 1$. High luminance variance under adverse conditions (e.g., rain or fog) decreases $\hat{P}_i$, indicating reduced perceptual consistency.

Additionally, QwenVL produces textual captions $\text{Caption}_i = \text{Decoder}_\phi(\text{Encoder}_\psi(x_i))$ that describe the scene context (e.g., *"dense fog reduces lane visibility"*), providing interpretable cues for dataset enrichment. This dual encoding converts raw frames into multimodal representations, coupling semantic confidence with linguistic awareness, forming the robust basis for adaptive, weather-aware perception systems.

Perception workflow in normal scenario: Under sunny conditions, the Vision Language Model offers minimal accuracy gains since baseline detectors like YOLOv10-M and RF-DETR already perform optimally with clear visuals (Figure 3). Well-defined edges, textures, and colors enable traditional models to operate effectively without additional semantic reasoning. Thus, the VLM's strength lies primarily in compensating for perception loss under adverse weather, where ambiguity and degradation are more significant.

4.1.2 Context-Aware Augmentation. To mitigate dataset imbalance and enhance representation of adverse weather, the proposed work QwenVL-Based Weather Augmentation Layer synthetically generates realistic environmental effects under semantic control. It enriches data diversity by simulating weather distortions such as rain, fog, and low illumination, thereby improving model robustness to unseen scenarios. Given an image frame x_i with weather label W_i, QwenVL extracts contextual tokens $y_i = f_{\text{QwenVL}}(x_i) = \{\text{scene type, visibility, illumination, surface texture}\}$ to guide sampling of augmentation parameters $\mathcal{A}_i \sim p(\mathcal{A} \mid y_i, W_i)$, where $\mathcal{A}_i = \{\rho_{\text{fog}}, \rho_{\text{rain}}, \rho_{\text{snow}}, \delta_{\text{illum}}\}$ controls fog density, rain intensity, snow coverage, and illumination change. The image is transformed as $x_i' = g(x_i, \mathcal{A}_i) = x_i \odot T_{\text{vis}}(\rho_{\text{fog}}, \delta_{\text{illum}}) + R_{\text{spec}}(\rho_{\text{rain}}) + S_{\text{mask}}(\rho_{\text{snow}})$, where T_{vis}, R_{spec}, and S_{mask} model transmittance attenuation, rain streaks, and snow occlusion, respectively. Visibility degradation follows $I'(x) = I(x)e^{-\beta d(x)} + A(1-e^{-\beta d(x)})$, with β from ρ_{fog}, $d(x)$ as scene depth, and A as atmospheric light.

QwenVL filters inconsistent outputs carefully to retain semantically coherent samples, forming the augmented dataset $\mathcal{D}' = \mathcal{D} \cup \{(x_i', W_i, t_i, s_i)\}$ with balanced weather coverage $p(W_i) \approx 1/|\mathcal{W}|$. By merging semantic reasoning with physics-based rendering, this layer bridges realism and diversity, yielding a multimodal and weather-invariant training corpus.

4.2 Contextual Sensor Reliability Scoring

4.2.1 Optimal Model Adaptation. The enriched dataset $\mathcal{D}'$, generated from the QwenVL-based Weather Augmentation Layer, trains two complementary models: YOLOv10 for real-time detection and RF-DETR for weather-resilient spatial reasoning balancing *speed* and *contextual depth* across weather domains. YOLOv10 acts as a fast, anchor-free detector optimized for localization, with loss $\mathcal{L}_{\text{YOLO}} = \mathbb{E}(x,y) \sim \mathcal{D}'[\lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}]$. Uniform sampling ($p(W_i) = 1/|\mathcal{W}|$) mitigates clear-weather bias.

RF-DETR uses transformer-based reasoning to handle occlusion and low visibility, embedding weather cues as $\mathbf{E}i = \text{Embed}(x_i') + \psi(W_i, \text{visibility}i)$, with loss $\mathcal{L}_{\text{RF-DETR}} = \mathbb{E}(x,y) \sim \mathcal{D}'[\lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} + \lambda_{\text{match}}\mathcal{L}_{\text{match}}]$. $\mathcal{L}_{\text{YOLO}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} + \lambda_{\text{obj}}\mathcal{L}_{\text{obj}}$. The joint objective $\mathcal{L}_{\text{total}} = \eta_1\mathcal{L}_{\text{YOLO}} + \eta_2\mathcal{L}_{\text{RF-DETR}}$ balances both models ensures a balance between real-time agility and spatial reasoning. Feature fusion $F\text{fusion} = \Phi(F_{\text{YOLO}}, F_{\text{RF-DETR}})$ yields a multimodal representation, ensuring robust, weather-adaptive perception under illumination loss and environmental degradation.

4.2.2 Sensor Reliability Score (SRS). It quantifies the trustworthiness of camera-based perception under varying lighting and weather conditions. It integrates three principal factors:

(i) Model-based confidence (C_m): For each frame f_i, the mean detection confidence is given by $C_m(f_i) = \frac{1}{N_i}\sum_{j=1}^{N_i} c_{ij}$, where c_{ij} is the confidence of the j^{th} object and N_i is the total detections.

(ii) Image quality (Q_i): Visual clarity is assessed through sharpness (S), brightness (B), and contrast (K), normalized as $Q_i(f_i) = \frac{1}{3}\left(\frac{S(f_i)}{S_{\max}} + \frac{B(f_i)}{B_{\max}} + \frac{K(f_i)}{K_{\max}}\right)$, where each term lies in $[0,1]$ after adverse scenarios dataset normalization.

(iii) Weather influence (W_w): Each weather type is assigned an empirical reliability value $W_w \in [0,1]$ to account for visual degradation from phenomena such as rain, fog, or snow.

Reliability fusion and classification: The overall reliability is computed as $SRS(f_i) = \alpha C_m(f_i) + \beta Q_i(f_i) + \gamma W_w$, where $\alpha+\beta+\gamma = 1$ and $(\alpha, \beta, \gamma) = (0.6, 0.2, 0.2)$ ensure robustness across conditions. Based on the resulting score, reliability is categorized as *High* ($SRS \geq 0.8$), *Medium* ($0.5 \leq SRS < 0.8$), *Low* ($0.3 \leq SRS < 0.5$), and *Very Low* ($SRS < 0.3$). These levels guide adaptive sensor weighting within the Weather-Aware Reasoning Layer, enabling consistent perception performance across dynamic environmental conditions. Overall, SRS provides a unified, interpretable, and weather-aware measure of camera reliability for AVs.

4.3 Dynamic Context-Aware Perception

The reasoning layer unifies weather awareness, reliability estimation, and sensor prioritization within an adaptive perception framework. Using **GPT-4o**, the system performs multimodal reasoning to dynamically assign sensor weights based on QwenVL-derived weather semantics and frame-level reliability cues. Given estimated weather W_i, camera reliability R_c, and temporal context t_i, the decision is defined as $\text{Decision} = f_{\text{LLM}}(W_i, R_c, t_i) \rightarrow [w_c, w_r, w_l]$, where $[w_c, w_r, w_l]$ are normalized fusion weights for camera, radar, and LiDAR, and the adaptive output is $O_t = \sum_{s\in c,r,l} w_s F_s$.For each frame x_i, the system produces a decision tuple $z_i = \text{frame}_i, W_i, R_i, D_i, E_i$, where R_i is the reliability score, D_i the selected sensor, and E_i the explanatory statement. Reliability is modeled as $R_i = \alpha V_i + \beta(1-B_i) + \gamma E_i^{(\text{stability})}$ with $(\alpha, \beta, \gamma) = (0.4, 0.3, 0.3)$ and $\alpha+\beta+\gamma = 1$. A camera is used as the primary source when $R_i \geq \tau_W$ ($\tau_W = 0.80$), ensuring robust SNR under clear conditions; otherwise, radar or LiDAR are prioritized. The LLM also generates interpretable statements such as *"High visibility ensures strong optical reliability ($R_i = 0.88$)" or "Fog reduces contrast ($R_i = 0.45$); Other sensor prioritization"*.Overall, this fusion of quantitative reliability and qualitative reasoning enables transparent, weather-adaptive perception, effectively completing the pipeline with enhanced robustness, scalability, explainability, and operational consistency across all weather conditions.

4.4 Result Analysis (RQ$_2$)

The results presented in Table 2 demonstrate that integrating the VLM significantly enhances detection robustness while maintaining feasible real-time performance varying weather.

Table 2: Performance comparison of YOLOv10-M, RF-DETR, and VLM-Enhanced models under diverse scenarios.

Scenario	YOLOv10-M			RF-DETR			VLM-Enhanced		
	mAP	L(ms)	FPS	mAP	L(ms)	FPS	mAP	L(ms)	FPS
Sunny	0.78	9	111	0.80	6	71	0.78	9	105
Rainy	0.75	10	100	0.77	7	62	0.86	12	91
Foggy	0.72	11	91	0.69	8	76	0.78	15	83
Snowy	0.69	11	83	0.72	7	60	0.80	14	77

Note: Here, mAP: Mean Average Precision, L: Latency (ms), FPS: Frames per Second is shown. VLM-Enhanced models integrate weather reasoning for robust perception.

The integration of the VLM greatly enhances perception robustness, contextual awareness, and efficiency. For *YOLOv10-M*, VLM-guided reasoning improves accuracy by up to 11% under low visibility, from 0.75 to 0.86 in rain and 0.72 to 0.78 in fog, while maintaining real-time performance with only a slight latency rise (10 ms to 12–15 ms) and stable throughput above 80 FPS. Similarly, for *RF-DETR*, accuracy increases from 0.77 to 0.86 in rain and from 0.69 to 0.78 in fog, sustaining 60–70 FPS. By embedding weather

semantics and reliability-driven attention, the VLM enhances spatial reasoning, mitigates feature ambiguity, and ensures consistent detection, serving as an adaptive intelligence layer that boosts both accuracy, efficiency, reliability, and real-time stability.

Figure 4 illustrates the architecture and operational flow of the Weather-Aware Reasoning Layer, the final result within the proposed VLM-based perception framework.

Figure 4: Illustration of the visual representation of Dynamic Context-Aware Perception Module. The module integrates visual reliability, weather semantics, and multimodal feature fusion to generate adaptive sensor selection and interpretable output.

Overall, VLM integration demonstrates consistent accuracy gains across challenging weather conditions with minimal computational cost, further validating its effectiveness in achieving weather-resilient, interpretable, and real-time perception performance.

4.5 Key Observations and Discussion

This section summarizes key observations addressing $\mathbf{RQ}_2$, emphasizing key performance trends, model behavior, and the significant impact of VLM-guided reasoning on perception robustness under varying weather conditions.

★ **Observation$_1$:** *QwenVL-enabled module boost in weather resilient perception-based accuracy* (Table 2).
Discussion: The integration of the QwenVL Vision Language Model notably enhances contextual understanding of weather semantics. By jointly encoding visual and linguistic cues, QwenVL improves scene interpretation beyond pixel-level perception, enabling more accurate inference of visibility, illumination, and environmental context. This multimodal reasoning yields an 8–12% accuracy gain under adverse conditions through more reliable weather classification and informed feature augmentation. Moreover, QwenVL's interpretable textual outputs (Figure 4) provide human-understandable explanations, supporting transparency, accountability, and explainability in autonomous perception systems.

★ **Observation$_2$:** *Context-aware data augmentation significantly boosts model resilience and overall performance.*
Discussion: The proposed *Context-Aware Augmentation* synthetically generates diverse weather scenarios, such as fog and rain, guided by semantic context. It improves class balance and exposes models to rare or extreme conditions often absent in real datasets. Models trained on the augmented dataset $\mathcal{D}'$ achieve 10–11% higher mAP (Table 2) in fog and rain due to realistic, semantically guided distortions (e.g., fog density, rain streaks, illumination shifts, low brightness), enabling stronger feature invariance, generalization, and enhanced resilience under complex adverse weather scenarios.

★ **Observation$_3$:** *Automated context labeling boosts performance by improving annotation efficiency while minimizing human bias.*
Discussion: QwenVL enables automatic, scalable labeling of diverse weather, lighting, and visibility directly from raw visual inputs, eliminating manual annotation and reducing preparation time and bias. Its multimodal reasoning generates structured JSON-based labels (e.g., *weather: "foggy", visibility: "low"*) (Figure 4), ensuring consistent and interpretable metadata across diverse datasets. This automated pipeline enriches data diversity, supports adaptive retraining, and significantly improves perception performance.

★ **Observation$_4$:** *Context-Aware and reliable perception through dynamic scenario-driven sensor prioritization and decision-making.*
Discussion: QwenVL enables dynamic sensor adaptation based on real-time weather interpretation. Under adverse conditions like *dense fog* or *heavy rain*, the system intelligently reduces camera reliance and prioritizes other sensors inputs. This adaptive reweighting effectively mitigates degraded visual cues, enhancing reliability and ensuring consistent, safe, and reliable all-weather perception.

★ **Observation$_5$:** *Minimal performance gain observed under normal weather scenarios with VLM integration pipeline.*
Discussion: Under clear weather, the Vision Language Model offers minimal accuracy gain since baseline detectors like YOLOv10-M and RF-DETR already perform optimally with well-defined edges, textures, and colors. With little visual ambiguity or environmental degradation to correct, the VLM provides limited benefit, its primary strength emerging clearly under adverse weather conditions.

5 Conclusion

This study tackles the challenge of ensuring reliable AV perception under adverse weather that degrades sensor performance and threatens safety. Through a systematic evaluation of detection models, fog is identified as the most detrimental condition, while RF-DETR and YOLOv10-M offer the best balance of accuracy and efficiency. To mitigate weather-induced degradation, this work proposes a VLM-guided adaptive perception pipeline is proposed, integrating QwenVL for semantic weather reasoning, automated labeling, and augmentation. Results show up to 10-12% accuracy gains in low-visibility scenarios with minimal latency. The Context-Aware Reasoning Module further enables dynamic sensor prioritization and interpretable decision-making, enhancing transparency and trust. Overall, this framework establishes a robust foundation for scalable, explainable, and weather-resilient perception, enabling future intelligent multimodal all-weather autonomy.

Acknowledgement

This work is supported in part by the National Science Foundation (NSF) grant CNS-2348151.

References

[1] Zakia Afrin, Fariya Tabassum, Hafsa Binte Kibria, MD Rafi Imam, and Md Rokibul Hasan. 2023. Yolov8 based object detection for self-driving cars. In *2023 26th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 1–6.

[2] QA Al-Haija, M Gharaibeh, and A Odeh. 2022. Detection in Adverse Weather Conditions for Autonomous Vehicles via Deep Learning. AI 2022, 3, 303-317.

[3] Mohanad N Alhasanat, Moath H Alsafasfeh, Abdullah E Alhasanat, and Saud G Althunibat. 2021. Retinanet-based approach for object detection and distance estimation in an image. *International Journal on Communications Antenna and Propagation (IRECAP)* 11, 1 (2021), 1–9.

[4] Tim Brophy, Darragh Mullins, Ashkan Parsi, Jonathan Horgan, Enda Ward, Patrick Denny, Ciarán Eising, Brian Deegan, Martin Glavin, and Edward Jones. 2023. A Review of the Impact of Rain on Camera-Based Perception in Automated Driving Systems. *IEEE Access* (2023).

[5] Harrison Delecki, Masha Itkina, Bernard Lange, Ransalu Senanayake, and Mykel J Kochenderfer. 2022. How do we fail? stress testing perception in autonomous vehicles. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5139–5146.

[6] Tausif Diwan, G Anirudh, and Jitendra V Tembhurne. 2023. Object detection using YOLO: challenges, architectural successors, datasets and applications. *multimedia Tools and Applications* 82, 6 (2023), 9243–9275.

[7] Shuqi Fang, Bin Zhang, and Jingyu Hu. 2023. Improved mask R-CNN multi-target detection and segmentation for autonomous driving in complex scenes. *Sensors* 23, 8 (2023), 3853.

[8] Craft Law Firm. 2024. Autonomous Vehicle Accidents: NHTSA Crash Data (2019–2024). https://www.craftlawfirm.com/autonomous-vehicle-accidents-2019-2024-crash-data/. Accessed: 2025-10-02.

[9] Ivan Fursa, Elias Fandi, Valentina Musat, Jacob Culley, Enric Gil, Izzeddin Teeti, Louise Bilous, Isaac Vander Sluis, Alexander Rast, and Andrew Bradley. 2021. Worsening perception: Real-time degradation of autonomous vehicle perception performance for simulation of adverse weather conditions. *arXiv preprint arXiv:2103.02760* (2021).

[10] Kshitiz Garg and Shree K Nayar. 2007. Vision and rain. *International Journal of Computer Vision* 75 (2007), 3–27.

[11] Yupei Guo, Yota Yamamoto, Hideki Yaginuma, and Yukinobu Taniguchi. 2025. Vehicle detection in CCTV with global-guided self-attention and convolution. *Complex & Intelligent Systems* 11, 10 (2025), 458.

[12] Weijie He, Yuwei Zhang, Ting Xu, Tai An, Yingbin Liang, and Bo Zhang. 2025. Object detection for medical image analysis: Insights from the RT-DETR model. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence*. 415–420.

[13] Zhi-peng JIANG, Zi-quan WANG, Yong-sheng ZHANG, Ying YU, Bin-bin CHENG, Long-hai ZHAO, and Meng-wei ZHANG. 2024. A vehicle object detection algorithm in UAV video stream based on improved Deformable DETR. *Computer Engineering & Science* 46, 01 (2024), 91.

[14] N Kandavel, S Vinod, B Shalini, R Pavithra, S Thangam, et al. 2025. Comparative Analysis of YOLOv8 and EfficientDet for Object Detection in Autonomous Vehicles. In *2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*. IEEE, 1–6.

[15] Matti Kutila, Pasi Pyykönen, Maria Jokela, Tobias Gruber, Mario Bijelic, and Werner Ritter. 2020. Benchmarking automotive LiDAR performance in arctic conditions. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 1–8.

[16] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V Albrecht. 2024. Explainable AI for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[17] Sidi Lu and Weisong Shi. 2023. Vehicle computing: Vision and challenges. *Journal of Information and Intelligence* 1, 1 (2023), 23–35.

[18] Yichen Luo, Daoxuan Xu, Gang Zhou, Yifan Sun, and Sidi Lu. 2024. Impact of Raindrops on Camera-Based Detection in Software-Defined Vehicles. In *2024 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)*. 193–205. doi:10.1109/MOST60774.2024.00028

[19] Cade Metz and Neal E Boudette. 2021. Inside Tesla as Elon Musk Pushed an Unflinching Vision for Self-Driving Cars. *International New York Times* (2021), NA–NA.

[20] Tanzim Mostafa, Sartaj Jamal Chowdhury, Md Khalilur Rhaman, and Md Golam Rabiul Alam. 2022. Occluded object detection for autonomous vehicles employing YOLOv5, YOLOX and Faster R-CNN. In *2022 IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, 0405–0410.

[21] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. 2022. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems* 35 (2022), 3819–3829.

[22] Edoardo Palladin, Roland Dietze, Praveen Narayanan, Mario Bijelic, and Felix Heide. 2025. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *European Conference on Computer Vision*. Springer, 484–503.

[23] Shiva Shankar Reddy, Midhunchakkaravarthy Janarthanan, and Inam Ullah Khan. 2025. RT-DETR with Attention-Free Mechanism: A Step towards Scalable and Generalizable Traffic Sign Recognition. *SGS-Engineering & Sciences* 1, 2 (2025).

[24] Linda Rutten. 2023. Deep Learning for Weather Condition Adaptation in Autonomous Vehicles. *Journal of Artificial Intelligence Research and Applications* 3, 1 (2023), 274–306.

[25] Prerna Saini, Anusha Dixit, and Deepak Kumar Sharma. 2025. Enhancing Object Detection in Adverse Weather for Autonomous Driving with YOLOv9. *International Energy Journal* 25 (2025).

[26] Fatih Sezgin, Daniel Vriesman, Dagmar Steinhauser, Robert Lugner, and Thomas Brandmeier. 2023. Safe autonomous driving in adverse weather: Sensor evaluation and performance monitoring. In *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1–6.

[27] Teena Sharma, Benoit Debaque, Nicolas Duclos, Abdellah Chehri, Bruno Kinder, and Paul Fortier. 2022. Deep learning-based object detection and scene perception under bad weather conditions. *Electronics* 11, 4 (2022), 563.

[28] Noor Ul Ain Tahir, Zuping Zhang, Muhammad Asim, Junhong Chen, and Mohammed ELAffendi. 2024. Object detection in autonomous vehicles under adverse weather: a review of traditional and deep learning approaches. *Algorithms* 17, 3 (2024), 103.

[29] Ajantha Vijayakumar and Subramaniyaswamy Vairavasundaram. 2024. Yolo-based object detection models: A review and its applications. *Multimedia Tools and Applications* 83, 35 (2024), 83535–83574.

[30] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. 2024. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems* 37 (2024), 107984–108011.

[31] Fanchang Yang, Lidong Huang, Xuewen Tan, and Yan Yuan. 2024. FasterNet-SSD: A small object detection method based on SSD model. *Signal, Image and Video Processing* 18, 1 (2024), 173–180.

[32] Bo Yu, Wei Hu, Leimeng Xu, Jie Tang, Shaoshan Liu, and Yuhao Zhu. 2020. Building the computing system for autonomous micromobility vehicles: Design constraints and architectural optimizations. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 1067–1081.

[33] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2636–2645.

[34] Shizhe Zang, Ming Ding, David Smith, Paul Tyler, Thierry Rakotoarivelo, and Mohamed Ali Kaafar. 2019. The impact of adverse weather conditions on autonomous vehicles: How rain, snow, fog, and hail affect the performance of a self-driving car. *IEEE vehicular technology magazine* 14, 2 (2019), 103–111.

[35] Biwei Zhang, Murat Simsek, Michel Kulhandjian, and Burak Kantarci. 2024. Enhancing the Safety of Autonomous Vehicles in Adverse Weather by Deep Learning-Based Object Detection. *Electronics* 13, 9 (2024), 1765.

[36] Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. 2023. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing* 196 (2023), 146–177.