SADWA: Fine-Grained Weather Awareness with Vision-Language Models for Seamless Autonomous Driving in Real Time

JINWOO KIM*
ETRI, KAIST
jwkim81@etri.re.kr, kaist.ac.kr

Hayeon O
ETRI
oph516@etri.re.kr

Youngmin Oh
ETRI
youngmin@etri.re.kr

Kyounghwan An
ETRI
mobileguru@etri.re.kr

Donghwan Lee
KAIST
donghwan@kaist.ac.kr

Abstract

Autonomous driving systems require adaptive driving strategies to ensure safe and efficient operation under diverse weather and road conditions. Traditional weather classification models often categorize images into broad labels such as clear, rainy, or snowy, lacking the finer granularity required for real-world driving scenarios. In this work, we fine-tune a lightweight vision-language model on a dataset that captures nuanced weather conditions, considering factors such as road surface state, precipitation intensity, visibility obstructions, and time of day. Rather than comparing models of different modalities, our focus is to demonstrate that a compact CLIP-based model can efficiently classify 18 strategically defined weather-road interaction classes—providing fast and accurate perception aligned with the demands of autonomous driving. Experimental results show that our approach enables real-time inference (17–19 ms) while maintaining strong classification performance. The proposed method supports practical deployment on resource-constrained platforms and offers a scalable path toward fine-grained weather awareness for autonomous systems, ultimately enhancing safety and decision-making under adverse environmental conditions.

1. Introduction

Autonomous driving systems must rapidly adapt to a diverse range of weather conditions and road environments. Adverse conditions such as rain, snow, fog, and intense glare significantly affect sensor reliability, degrading the accuracy of object detection, lane recognition, and distance estimation [12, 26, 28, 35, 36, 57]. These environmental challenges introduce complex interactions between visibility, road surface state, and illumination effects, which directly influence vehicle control and safety [43, 51].

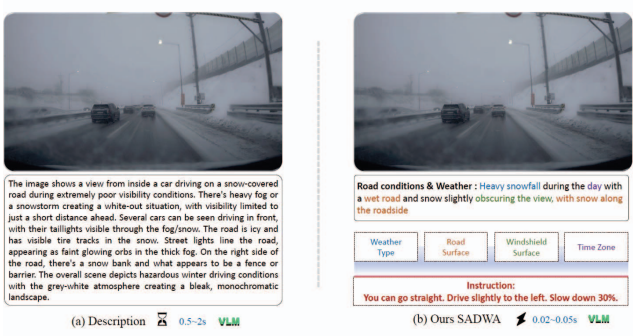

Figure 1. Compared to conventional VLMs that generate extensive environmental descriptions or rely on query-based QA processes, our proposed SADWA framework efficiently extracts and delivers only the key weather-related factors that impact driving strategy. By minimizing unnecessary information, SADWA enables autonomous vehicles to interpret road and weather conditions, ensuring real-time adaptability for safer and more efficient driving.

Beyond conventional object detection and semantic segmentation, it is essential to develop perception models capable of intuitively interpreting the combined impact of weather and road conditions, enabling autonomous vehicles to adapt their driving strategies in real time [21, 58].

Traditional weather classification methods typically use convolutional neural networks (CNNs) with coarse-grained labels such as clear, rainy, or snowy [1, 14, 16, 23, 24, 41]. While effective in controlled scenarios, these methods miss crucial factors, including road surface state (dry, wet, snow-covered), visibility impairments (fog, precipitation occlusions), and illumination effects (glare severity, nighttime reflections) [2, 4, 40]. Such cues are vital for safe driving, influencing braking distance, steering, and lane-keeping [50]. Furthermore, CNN-based models struggle to generalize across diverse weather and adapt to changing conditions, causing performance drops in unseen domains [46, 61].

In contrast, recent advances in Vision-Language Models (VLMs) leverage joint visual-textual embeddings to improve environmental understanding by integrating image features with natural language descriptions [3, 13, 19, 20, 22, 31, 47, 54, 63]. This multimodal approach enables richer perception of driving scenes and reasoning about complex road-weather interactions [17, 39, 61]. However, existing VLM approaches face three limitations: (1) high computational cost restricting real-time use, (2) limited spatial reasoning reducing navigation reliability, and (3) lack of diverse, fine-grained, well-annotated weather datasets [8, 49, 62]. These hinder effectiveness in mission-critical applications needing both high accuracy and low latency [33, 42, 55].

Broadly, VLM-based weather perception follows two directions. The first optimizes real-time performance by reducing computation via compression, distillation, or architecture redesign [9, 33, 33, 42, 44], achieving low latency but sacrificing fine-grained reasoning. The second prioritizes high-precision multi-factor classification, gaining accuracy but at higher latency and memory cost [13, 27, 48]. Bridging these remains an open challenge for deployment.

Existing autonomous driving datasets offer weather diversity but rarely encode fine-grained links between road surface, visibility, and weather effects [18, 25, 34, 35, 43, 62]. For example, a "rainy" label may describe wet roads with clear visibility or heavy rain obscuring lane markings, which differ in driving risk [38, 57]. This gap motivates datasets integrating multiple environmental attributes into a unified taxonomy for safer, adaptive driving [32, 50, 51].

We introduce the SADWA dataset, a fine-grained benchmark capturing complex road-weather interactions via a structured 18-class taxonomy. SADWA encodes combinations of weather type, surface state, and visibility, enabling more nuanced decisions than coarse 3-class schemes. We integrate this dataset with fine-tuned Mobile-CLIP [44], a lightweight VLM optimized for real-time inference. This pairing delivers high accuracy and low latency, deployable on vehicles with limited compute [7, 29]. Data were collected using an automotive-grade camera on an operational autonomous platform, ensuring scalability, consistent annotation, and real-world relevance.

Our main contributions are:

- **SADWA Dataset:** A publicly released fine-grained road-weather dataset with multi-layered taxonomy, filling gaps in current benchmarks.
- **Optimized Lightweight VLM:** Fine-tuned Mobile-CLIP for real-time adverse-weather classification, balancing accuracy and speed for embedded deployment.
- **Adaptive Instruction Generation:** Leveraging spatial reasoning and multimodal alignment to provide interpretable, weather-aware driving guidance.

2. Related Work

Traditional Weather Classification Approaches. Early weather classification models primarily adopted convolutional neural networks (CNNs) with coarse-grained labels such as clear, rainy, or snowy [1, 14, 16, 23, 24, 41]. While effective in controlled settings, these models overlooked cues like road surface state (dry, wet, snow-covered), visibility impairments (fog, precipitation occlusions), and illumination effects (glare severity, nighttime reflections) [2, 4, 40]. Such factors influence braking distance, lane-keeping stability, and hazard perception [50]. Moreover, CNN-based methods often struggle to generalize to unseen and adapt to rapidly changing conditions [46, 61], limiting robustness in real-world deployments.

Vision-Language Models for Weather Perception. Recent advances in *Vision-Language Models (VLMs)* integrate visual and textual features to enhance contextual understanding of road-weather interactions [3, 13, 19, 20, 47, 54, 63]. This multimodal approach improves scene reasoning and semantic richness compared to vision-only models [17, 39]. However, VLMs face challenges: high computational demand [8, 55, 62], limited spatial reasoning for navigation, and insufficient fine-grained weather datasets [49]. While some works focus on lightweight VLM designs [33, 42, 44] to achieve low-latency inference, they may sacrifice detailed environmental reasoning. Conversely, high-accuracy models [13, 27, 31] often incur latency and memory costs unsuitable for embedded autonomous platforms.

Existing Weather Datasets and Limitations. Several autonomous driving datasets consider weather diversity [25, 35, 43, 62], yet most rely on broad weather labels without capturing road-weather-visibility interdependencies. For example, "rainy" may represent wet roads with clear visibility or heavy rainfall obscuring lane markings [57], which demand different driving strategies. Some benchmarks address adverse conditions [50, 51], but lack structured taxonomies explicitly encoding fine-grained safety-relevant attributes, limiting their utility for adaptive decision-making. Additionally, prior evaluations on CNN-based weather datasets, such as BDD100K [57] and ACDC [35], show strong performance under limited conditions but steep degradation in rare or complex scenarios, further motivating SADWA's fine-grained design.

Real-Time VLM Applications. Deploying VLMs in autonomous driving requires balancing accuracy, interpretability, and efficiency. Recent studies explore architectural optimizations, model compression, and cross-modal distillation [15, 37, 52, 61] to reduce latency while maintaining semantic alignment. These methods demonstrate the feasibility of real-time perception on resource-constrained platforms, providing a foundation for lightweight yet contextually rich road-weather classification models.

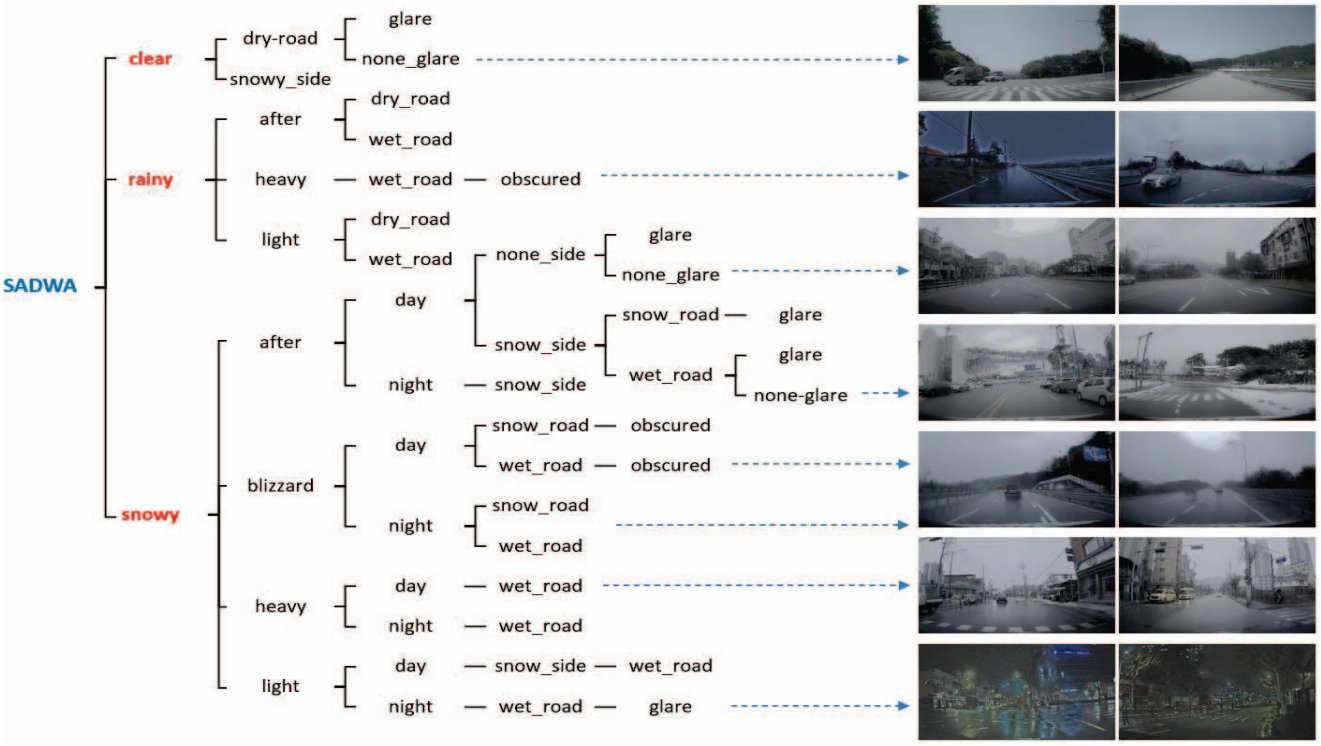

Figure 2. **SADWA dataset classification structure.** The dataset categorizes weather conditions, road surface states, visibility impairments, and external interferences into a hierarchical framework, enabling a comprehensive perception of driving environments.

3. Dataset Framework

3.1. SADWA Dataset Structure

The SADWA dataset is designed to provide fine-grained weather perception by integrating both environmental and road surface attributes. Unlike existing datasets that classify weather into broad categories such as clear, rainy, or snowy, SADWA introduces a hierarchical classification structure. The dataset defines 18 distinct weather-road relationship classes, allowing autonomous vehicles to better understand and adapt to various driving conditions. Furthermore, the classification accounts for road surface states under different weather conditions, the extent to which visibility through the windshield is maintained, and the level of road reflection after precipitation. This detailed categorization ensures a more precise understanding of how weather interacts with road conditions, enabling autonomous vehicles to make informed driving decisions.

3.2. Hierarchical Classification Framework

The SADWA dataset structures weather perception into a hierarchical framework, as illustrated in Fig. 2. The classification starts with three primary weather conditions: **clear**, rainy, and snowy. Each category is further subdivided based on road surface conditions and additional environmental factors affecting driving visibility and perception.

Clear Weather Conditions The clear category includes dry-road and snowy-side conditions. The dataset differentiates whether the road is fully dry or has snow accumulated along the roadside. Additionally, we consider glare as a critical factor, as direct sunlight or reflections from wet surfaces can significantly impact perception systems.

Rainy Conditions The rainy category is divided into heavy rain and light rain, where road conditions vary between wet-road and dry-road. The dataset not only classifies precipitation intensity but also considers how rain affects visibility. Water droplets on the windshield and camera lens can create distortion, reducing the effectiveness of object detection algorithms. Additionally, wet roads introduce hydroplaning risks and reflect ambient lights from street lamps and vehicle headlights, potentially misleading perception systems. The dataset labels these effects with obscured when heavy rain significantly limits visibility.

Snowy Conditions The snowy category is further split into after snow, blizzard, heavy snow, and light snow, each of which has distinct road conditions. For example, snow-side refers to accumulated snow beside the road, while snow-road indicates an entirely snow-covered driving surface. Additionally, night-time and glare factors are considered to represent real-world autonomous driving challenges.

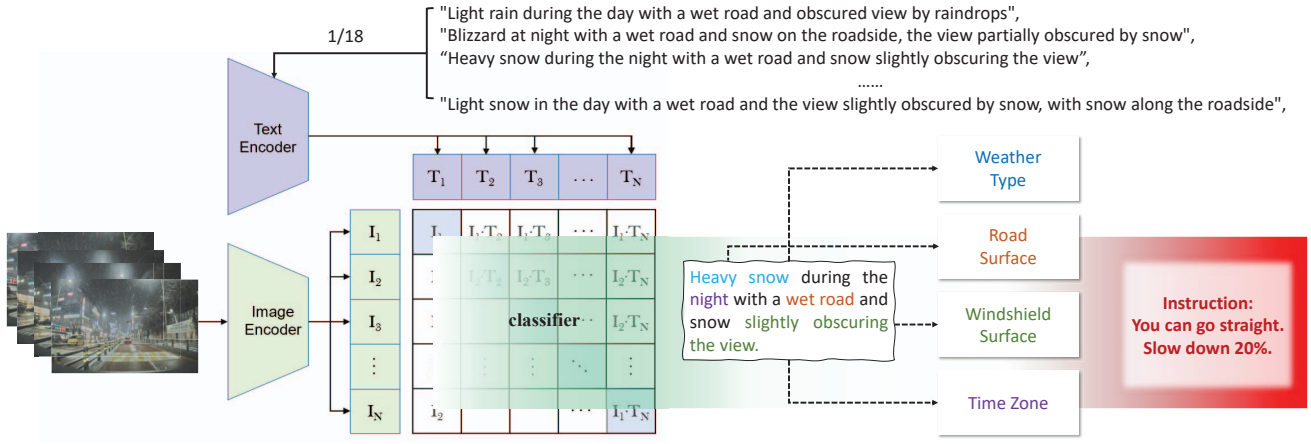

Figure 3. **SADWA dataset classification structure.** The dataset categorizes weather conditions, road surface states, visibility impairments, and external interferences into a hierarchical framework, enabling a comprehensive perception of driving environments.

3.3. Weather and Road Condition Classification

To ensure detailed weather classification, SADWA incorporates multiple attributes that directly impact driving safety. The dataset includes:

- **Weather conditions**: Clear, light rain, heavy rain, light snow, heavy snow, blizzard.
- **Road surface conditions**: Dry road, wet road, snow-covered road.
- **Visibility impairments**: Fog, precipitation-induced occlusions.
- **Light interferences**: Sun glare, nighttime reflections.

Each sample in the dataset is labeled with a combination of these attributes, ensuring that the dataset reflects real-world variations in weather and road conditions.

4. Methodology

4.1. Fine-Tuning VLM Model

Our approach leverages Mobile-CLIP, a lightweight vision-language model, fine-tuned on the SADWA dataset to enable efficient real-time weather classification in Fig. 3. Unlike traditional models that rely purely on visual data, our method integrates both image and text embeddings, enhancing contextual understanding of diverse weather conditions. The SADWA dataset defines 18 predefined weather-road classes, each mapped to corresponding textual descriptions. By aligning visual features with semantically meaningful weather representations, Mobile-CLIP effectively learns fine-grained weather perception, allowing autonomous systems to make more informed driving decisions.

Furthermore, as illustrated in Fig. 2, the SADWA dataset is designed not only for basic weather classification but also for encoding key factors essential to driving strategy determination. Instead of simply distinguishing between general weather conditions, SADWA structures its labels to incorporate predefined keywords relevant to road surface states, precipitation intensity, and visibility impairments. This ensures that the model learns weather attributes in a way that directly contributes to real-time driving adaptability, improving both perception accuracy and vehicle control in complex environmental scenarios.

The training process fine-tunes Mobile-CLIP on the SADWA dataset by learning visual and textual representations of weather conditions. The input image sequence is processed through a vision encoder to extract key environmental attributes while predefined weather descriptions are embedded using a text encoder to structure weather-related knowledge. Leveraging contrastive learning, the model aligns semantically similar image-text pairs, improving its ability to distinguish fine-grained weather-road interactions essential for adaptive driving strategies.

4.2. Keyword Embedding and Instruction

Once the model classifies an input scene into one of the 18 predefined weather categories, it further embeds critical weather-related keywords to refine decision-making. As illustrated in Fig. 3, the extracted textual description is structured into four key perception attributes:

- **Weather Type**: Identifies environmental conditions (e.g., heavy snow, light rain).
- **Road Surface**: Determines surface conditions (e.g., dry road, wet road, snow-covered road).
- **Windshield Surface**: Assesses visibility obstructions caused by raindrops, snow, or fog.
- **Time Zone**: Incorporates temporal information (daytime or nighttime) to adjust lighting-based perception.

These structured embeddings serve as input for the final Instruction Generation Module, which translates weather perception into actionable driving guidance.

Figure 4. The data collection vehicle and the autonomous vehicle equipped with SADWA.

4.3. Driving Instructions

The final instruction module generates adaptive driving recommendations, ensuring safe vehicle operation under adverse weather conditions. Driving guidance is determined based on:

- **Drivable Space**: Identifies whether the detected conditions allow safe forward driving.
- **Planning Adjustment**: Advises lateral biasing (left or right) in cases of wet or icy road surfaces.
- **Speed Regulation for degradation driving**: Enforces legal speed reductions based on traffic regulations:
  - **Rainy Conditions**: 20% speed reduction.
  - **Snow/Foggy Conditions**: 30–50% speed reduction.

This structured methodology ensures that weather-aware perception directly influences autonomous vehicle behavior, enhancing safety and real-time adaptability.

5. Experiments

5.1. SADWA Dataset

Dataset Acqusition. The SADWA dataset was collected using an autonomous vehicle in Fig. 4 equipped with automotive-grade cameras, capturing diverse driving environments under various weather conditions. As illustrated in Fig. 2, the dataset includes data from regions with heavy snowfall and rainfall, ensuring robust coverage of seasonal variations affecting road conditions. The camera system records Full HD (1920×1080) video with a 74-degree field of view and HDR, maintaining high-quality imagery across different lighting conditions.

Training Dataset. The dataset consists of 18 weather-road interaction classes, categorized by precipitation intensity, road surface state, and visibility impairments. The training set contains 3,757 images, while the validation and test sets include 940 and 800 images, respectively, ensuring a well-balanced evaluation framework. Each instance is annotated with structured captions describing environmental conditions and their impact on driving strategies.

Annotation and Data Collection. The dataset was collected using an automotive-grade camera mounted on an autonomous vehicle platform. Data was gathered in diverse environments, including urban roads, highways, and rural areas, to ensure broad applicability. To enhance annotation quality, we employed a hybrid labeling approach:

- **Manual verification**: We reviewed and refined labels to correct inconsistencies with sequence images for a long-term period while driving.
- **Automated tagging**: Weather conditions were initially identified using sensor metadata and real-time weather reports while driving with our autonomous vehicle.

This two-step process ensures high annotation accuracy, making SADWA a reliable dataset for weather-aware autonomous driving applications.

SADWA Captioning in detail Environmental attributes are categorized into four key elements: weather conditions, road surface states, visibility levels, and time of day, each color-coded for clarity. Table 1 presents the structured captioning framework and driving instructions used in the SADWA dataset. These structured labels enable autonomous systems to interpret and adapt to diverse driving scenarios efficiently. Speed adjustments are determined based on traffic regulations: no slowdown is required in clear conditions, whereas extreme conditions such as blizzards or severe visibility impairment necessitate up to a 50% reduction. Light rain or wet road conditions require a 20% reduction, while snow accumulation and partially obscured visibility call for a 30% reduction to ensure driving stability.

Fig. 5 shows examples of structured captioning and instruction generation applied in the our SADWA dataset. Each image is annotated with weather-road interaction descriptions that facilitate real-time adaptive decision-making for autonomous vehicles. This structured approach ensures consistent perception, allowing vision-language models to capture and utilize essential environmental information. By maintaining reliable annotation consistency, the SADWA dataset enhances autonomous navigation robustness in adverse weather conditions, improving real-world deployment viability.

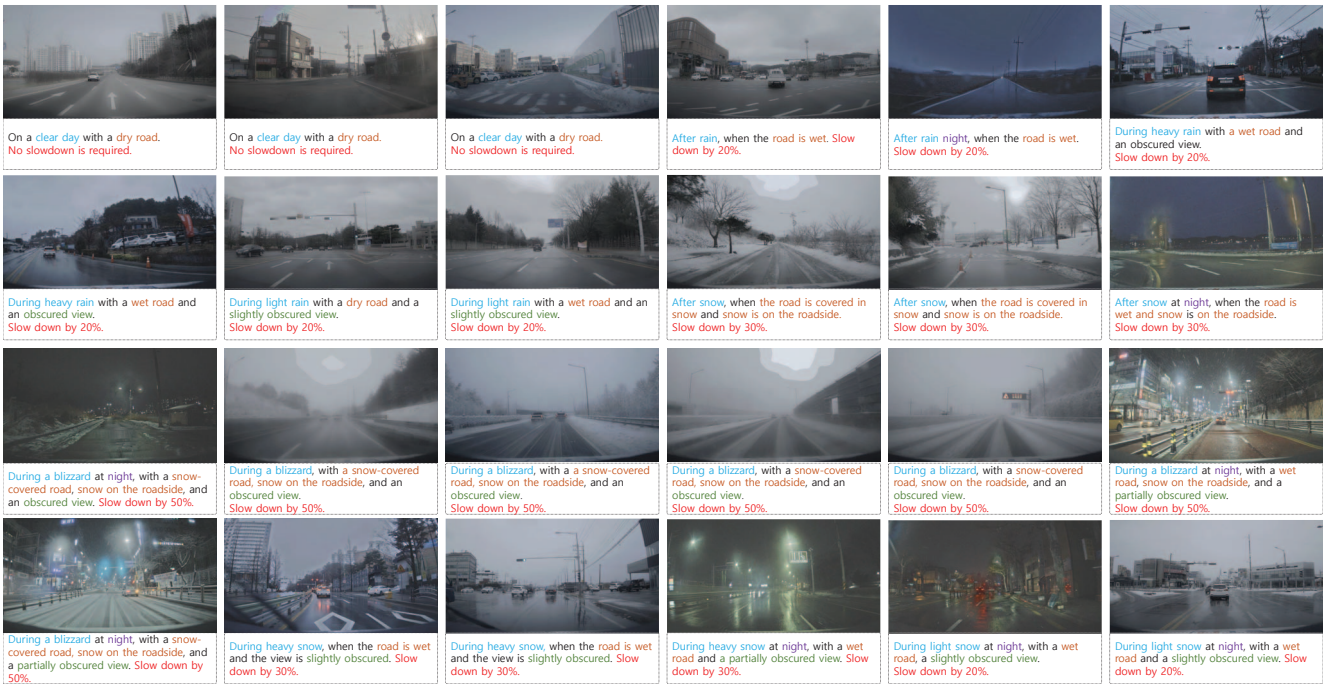

Figure 5. The data collection vehicle and the autonomous vehicle equipped with SADWA.

Category	Types of Driving Condition	Driving Instruction
Weather Condition	Clear, Rainy, Snowy, Blizzard, Heavy Snow, Light Snow	No slowdown required (clear conditions)
Road Surface	Dry Road, Wet Road, Snow-Covered Road, Snow-Covered Road side	Slow down by 20% (light rain, wet roads)
Visibility Status	No Obstruction, Slightly Obscured, Partially Obscured	Slow down by 30% (snow accumulation, nighttime conditions)
Time of Day	Daytime, Nighttime	Slow down by 50% (blizzard, extreme visibility loss)

Table 1. Structured Captioning and Driving Instructions in the SADWA Dataset. Categories are color-coded for clarity, linking weather-road interactions to adaptive driving strategies.

5.2. Evaluation for SADWA

Evaluation Metrics. To assess the performance of our CLIP-based models trained on the dataset, we compare them with BLEU-4 [53], ROUGE-L [30], and CIDEr [45] as key evaluation benchmarks. Additionally, we analyze the effectiveness of the VL-iRoSA CDWSP training approach using SPICE [5], BERTScore [60], mAP [6, 59] to validate its performance further.

Evaluation on SADWA Dataset Our primary objective was to compare the performance of fine-tuned CLIP-based models against pre-trained VLMs such as InternVL2 and Gemma3-12B in classifying images into 18 predefined weather categories. The pre-trained VLMs used textual prompts for classification, while our MobileCLIP-S1 and S2 models were trained directly on the SADWA dataset without text prompts. This experimental setup highlights the benefits of direct fine-tuning over prompt-based classification. The results demonstrate that MobileCLIP-S1 and S2 outperform InternVL2-1B and InternVL2-8B across key metrics, including BLEU-4, CIDEr, ROUGE-L, METEOR, and SPICE. Notably, MobileCLIP-S1 achieves the highest ROUGE-L (0.8134) and SPICE (0.8373) scores, reflecting superior captioning accuracy and semantic alignment. In contrast, InternVL2 models relying on text prompts exhibit lower performance, with InternVL2-1B achieving only 0.3629 in BLEU-4 and 2.1812 in CIDEr. These findings confirm that task-specific fine-tuning leads to better performance than using general-purpose VLMs.

Inference Speed and Practicality A critical advantage of our approach is inference efficiency (see Table 2). While large-scale models such as InternVL2-8B and Gemma3-12B require 1332ms and 3551ms per image, respectively, our MobileCLIP-S1 and S2 achieve inference speeds of 17.76ms and 19.09ms. This represents a speed-up of over 100 times compared to InternVL2-8B and nearly 200 times compared to Gemma3-12B, highlighting the feasibility of our fine-tuned models for real-time autonomous driving applications.

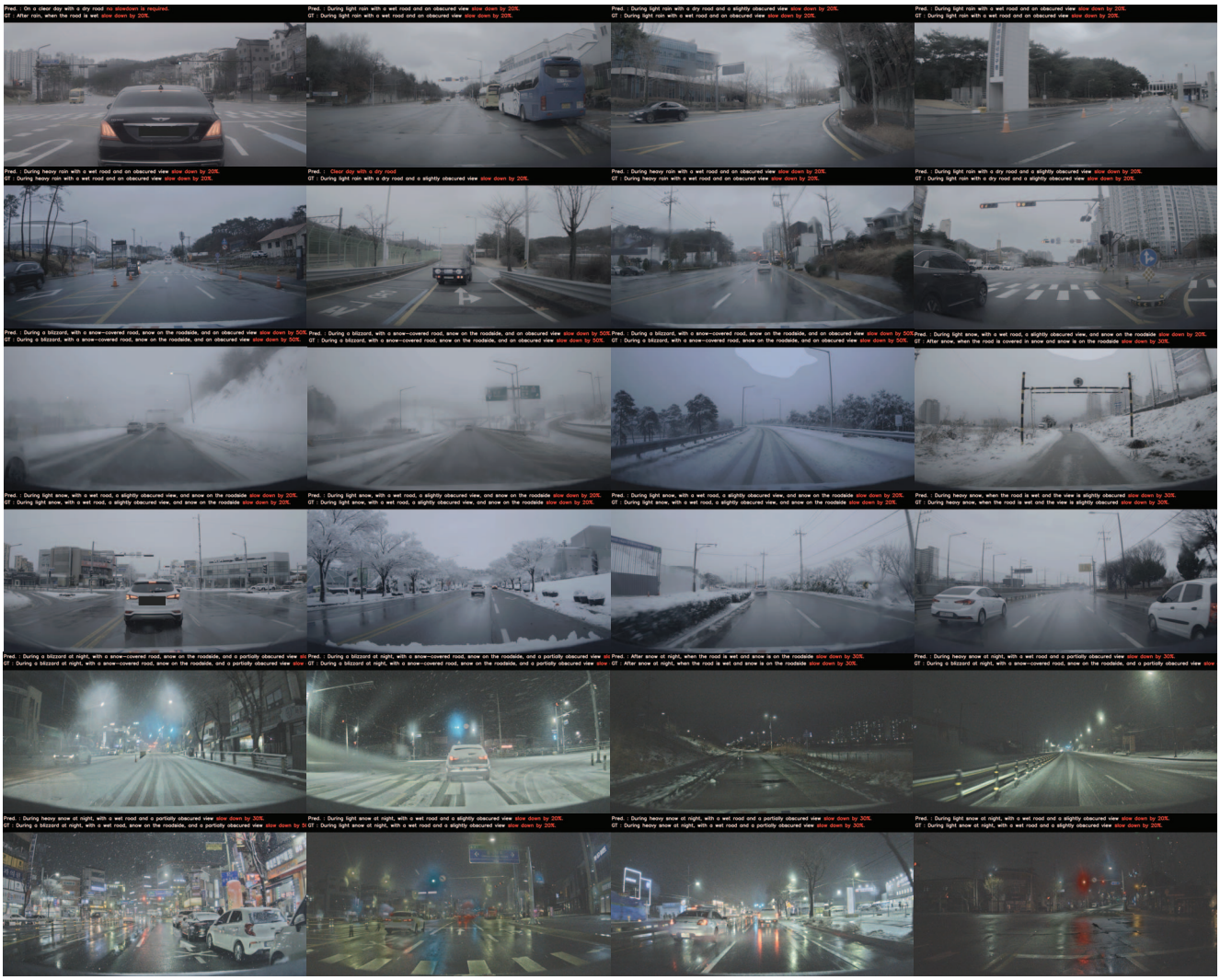

Figure 6. The data collection vehicle and the autonomous vehicle equipped with SADWA.

Method	Text prompt	SADWA Dataset					
		BLEU-4 ↑	CIDEr ↑	ROUGE_L ↑	METHOR ↑	SPICE ↑	Inference (ms/img)
InternVL2-1B [10]	✓	36.29	2.1812	56.51	33.15	52.06	520.2
InternVL2-8B [10]	✓	48.30	3.4835	66.46	38.44	64.10	1332
Gemma 3-12B [11]	✓	55.60	3.9964	72.60	41.58	70.81	3551
CLIP-KD [56]	×	61.34	5.3572	72.57	45.24	76.34	**14.20**
MobileCLIP-S1 [44]	×	**75.36**	**6.5877**	81.34	**53.56**	**83.73**	17.76
MobileCLIP-S2 [44]	×	74.74	6.4225	**81.47**	52.71	83.50	19.09

Table 2. Experimental results on SADWA test set.

Additionally, MobileCLIP-S1 and S2 significantly outperform Gemma3-12B in CIDEr (6.59 vs. 4.00) and ROUGE-L (81.47 vs. 72.60), clearly demonstrating the effectiveness of lightweight, task-specific models over large, generalist VLMs. In summary, the results strongly validate our fine-tuned CLIP models, clearly showing their ability to surpass larger VLMs while consistently maintaining real-time inference. Future work will further refine SADWA models, expand dataset diversity, and incorporate additional multimodal inputs for improved real-world performance.

Efficiency and Deployment Considerations As shown in Table 2, the inference speed limitations of InternVL and Gemma VLMs make them impractical for real-time autonomous driving instruction generation. More critically, autonomous systems require lightweight, embedded AI solutions. Large VLMs demand significant GPU VRAM, limiting their deployment in resource-constrained environments. While high-capacity VLMs offer broader adaptability, our approach prioritizes efficiency, ensuring that autonomous vehicles obtain essential speed adjustment instructions under adverse weather conditions based on regulatory guidelines. MobileCLIP significantly reduces inference time and memory consumption, making it a practical alternative for integrating real-time, weather-aware decision-making in autonomous driving systems. Fig 6 shows inference results on unseen test samples excluded from training. Despite the small dataset, our model performs robustly and aligns well with the structured framework, offering accurate driving guidance under adverse weather. The generated instructions provide detailed environmental understanding and speed regulation based on legal standards. Our model was trained using automotive-grade camera data from a real autonomous vehicle, supporting real-world applicability. Future work includes expanding the dataset and fine-tuning with larger VLMs to automate data generation and improve generalization across diverse weather conditions.

5.3. Ablation study

To evaluate the benefit of multimodal models over traditional vision-only architectures, we compare CLIP-based approaches with a CNN baseline. ResNet-50 shows the lowest performance due to its vision-only architecture, which cannot capture the nuanced, multi-attribute structure of the 18-class SADWA dataset. In contrast, MobileCLIP effectively leverages multimodal alignment between visual and textual semantics, enabling more accurate classification of detailed road-weather interactions. Fine-tuning further boosts performance across all models, with MobileCLIP-S1 improving from 19.33 mAP and 31.79% Top-1 accuracy to 67.78 mAP and 95.21%. Between the two MobileCLIP variants, S1 achieves slightly higher accuracy, while S2 provides a better balance between accuracy and parameter count.

A coarse 3-class taxonomy (clear, rain, snow) simplifies classification but omits key attributes like glare severity, surface wetness, and snow accumulation, which directly influence driving strategies. These fine-grained cues are essential for safe decision-making, making the 18-class design central. Consistent gains in MobileCLIP and CLIP-KD confirm transferability across CLIP architectures with identical training settings.

Model	# Params	mAP ↑	Top-1 Acc (%) ↑
MobileCLIP-S1	85M	19.33	31.79
MobileCLIP-S2	0.1B	20.64	30.23
CLIP-KD	0.3B	18.75	27.67
ResNet-50	25.6M	14.92	18.33
MobileCLIP-S1*	85M	67.78	**95.21**
MobileCLIP-S2*	0.1B	**66.36**	96.89
CLIP-KD*	0.3B	69.65	88.75
ResNet-50*	25.6M	38.31	43.22

Table 3. Comparison of CLIP-based and CNN-based models before and after fine-tuning on the SADWA dataset.

6. Discussion

Data imbalance remains a challenge, as rare conditions like blizzards and dense fog are underrepresented, potentially affecting robustness. The SADWA-trained model may also struggle to generalize across unseen road or lighting environments. Nevertheless, our 18-class taxonomy offers a strategic structure capturing critical weather-road semantics—such as glare, wetness, and visibility loss—crucial for autonomous driving. This allows lightweight models to learn efficiently and issue actionable guidance. Our results show that task-specific design can outweigh scale or volume. Expanding the dataset with diverse, real-world scenarios can further enhance generalization and robustness.

7. Conclusion

We presented a fine-tuned, lightweight CLIP-based model trained on SADWA for real-time weather classification. Our model extracts key features from structured class descriptions to generate interpretable driving instructions, improving decision efficiency. Unlike generalist VLMs, our method leverages compact vision-language alignment to precisely classify road environments affected by weather conditions in real time. This enables more strategic driving decisions without relying on large-scale models. Results show that lightweight, task-aligned models can rival larger ones when paired with structured, domain-specific taxonomies. Future work will expand SADWA, integrate multimodal sensors, and develop adaptive learning to improve real-world robustness.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00236245, Development of Perception/Planning AI SW for Seamless Autonomous Driving in Adverse Weather/Unstructured Environment).

References

[1] V. Afxentiou and T. Vladimirova. Evaluation of cnn-based approaches to adverse weather image classification for autonomous driving systems. *IEEE Open Journal of Intelligent Transportation Systems*, 2025. 1, 2

[2] M. Z. Alam, Z. Kaleem, and S. Kelouwani. Glare Mitigation for Enhanced Autonomous Vehicle Perception. *IEEE Transactions on Intelligent Vehicles*, 2024. 1, 2

[3] Jean-Baptiste Alayrac, Jeff Donahue, Paul Luc, et al. Flamingo: A Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022. 2

[4] N. Aloufi, A. Alnori, and A. Basuhail. Enhancing Autonomous Vehicle Perception in Adverse Weather: A Multi-Objectives Model for Integrated Weather Classification and Object Detection. *Electronics*, 2024. 1, 2

[5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016. 6

[6] X. Cao, T. Zhou, Y. Ma, W. Ye, and C. Cui. MAPLM: A Real-World Large-Scale Vision-Language Benchmark for Map and Traffic Scene Understanding. In *CVPR*, 2024. 6

[7] David Chen, Emma Wilson, and Mark Taylor. Multi-modal Perception in Autonomous Driving: A Comprehensive Survey. *IEEE TPAMI*, 45(9):9876–9895, 2023. 2

[8] David Chen, Emma Wilson, and Mark Taylor. Real-time Vision-Language Models: Challenges and Solutions. *IEEE TPAMI*, 46(1):234–253, 2024. 2

[9] J. Chen and S. Lu. An Advanced Driving Agent with the Multimodal Large Language Model for Autonomous Vehicles. In *IEEE International Conference on Mobility*, 2024. 2

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *CVPR*, 2024. 7

[11] Google DeepMind. Welcome Gemma 3: Google's All New Multimodal, Multilingual, Long Context Open LLM. Hugging Face Blog, 2025. Published: March 12, 2025. 7

[12] A. Elgazwy, K. Elgazzar, and A. Khamis. Predicting pedestrian crossing intentions in adverse weather with self-attention models. *IEEE Transactions on Intelligent Vehicles*, 2025. 1

[13] T. Feng, W. Wang, and Y. Yang. A survey of world models for autonomous driving. *IEEE TPAMI*, 2025. 2

[14] Q. Gao, H. Hu, and W. Liu. Traffic Sign Detection under Adverse Environmental Conditions Based on CNN. *IEEE Access*, 2024. 1, 2

[15] W. Huang, Z. Zhai, Y. Shen, S. Cao, F. Zhao, and X. Xu. Dynamic-LLava: Efficient Multimodal Large Language Models via Dynamic Vision-Language Context Sparsification. *arXiv preprint arXiv:2412.00876*, 2024. 2

[16] B. Jayaprakash, M. Eagon, and L. Zhan. De-Snowing Algorithm for Long-Wavelength LiDAR. In *IEEE Intelligent Vehicles Symposium (IV)*, 2024. 1, 2

[17] M. Jiang, T. Qin, and M. Yang. Rl-ogm-parking: Lidar ogm-based hybrid reinforcement learning planner for autonomous parking. In *ICRA (accepted)*, 2025. 2

[18] Minsoo Kim, Junyoung Park, and Sungho Lee. Hazy-da: A benchmark for hazy driving scene understanding in autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[19] Junnan Li, R. R. Selvaraju, A. Gotmare, et al. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 2021. 2

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *CVPR*, 2022. 2

[21] Ming Li, Yan Zhao, Jie Xu, and Hao Wang. Evaluation of Safety Cognition Capability in Vision-Language Models for Autonomous Driving. In *ECCV*, 2024. 1

[22] Wei Liu, Yue Zhang, and Fangyu Chen. Vision-language models: A comprehensive survey on multimodal learning for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *CVPR*, 2022. 1, 2

[24] N. Luo, J. Zhong, Y. Lin, X. Dai, and X. Lai. Research on optimization strategy of yolov8 target detection model based on scene adaptation. In *IEEE Intelligent Vehicles Symposium (IV)*, 2024. 1, 2

[25] B. Ma. Design of automatic driving and parking system for new energy vehicles based on artificial intelligence. In *International Conference on Frontier Computing*, 2025. 2

[26] C. Meng, K. Song, H. Xie, and W. Xing. Adaptive Localization Algorithm for Autonomous Mining Trucks with Motion Distortion Removal. *SAE Technical Papers*, 2025. 1

[27] Z. Meng, Y. Zhang, Z. Zheng, Z. Zhao, and J. Ma. Agentalign: Misalignment-adapted multi-agent perception for resilient inter-agent sensor correlations. In *CoRL (accepted)*, 2024. 2

[28] E. Oyedokun and B. William. Thorough Analysis of Object Detection for Autonomous Vehicles. *Preprints*, 2025. 1

[29] Emma Park, Robert Wilson, and Sarah Davis. Edge Case Understanding in Dynamic Environments. In *ICCV*, pages 3456–3465, 2023. 2

[30] J. Park and S. Choi. Bridging Vision and Language: Modeling Causality and Temporality in Video Narratives. *arXiv preprint*, 2024. 6

[31] Michael Park, Emma Davis, and Robert Wilson. Vision-Language Models for Motion Planning: A Survey. *IEEE TPAMI*, 45(11):12345–12364, 2023. 2

[32] Sarah Park, David Kim, and Jennifer Lee. Cornerdrive: Visual recognition of anomalous driving scenarios. In *CVPR*, 2024. 2

[33] Sarah Park, Emma Wilson, and Robert Davis. Lightweight Vision-Language Models for Real-time Understanding. In *CVPR*, 2024. 2

[34] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Foggy cityscapes: Semantic understanding in fog. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 2

[35] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. In *ICCV*, 2021. 1, 2

[36] Z. Shao, H. Wang, Y. Cai, and L. Chen. Ua-fusion: Uncertainty-aware multimodal data fusion framework for 3d object detection of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 1

[37] A. Sharshar, L. U. Khan, W. Ullah, and M. Guizani. Vision-Language Models for Edge Networks: A Comprehensive Survey. *arXiv preprint arXiv:2502.07855*, 2025. 2

[38] Pei Sun et al. The waymo open dataset: Panoramic video perception for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[39] Rui Sun, Zhen Qian, Wei Zhou, and Tianyu Lin. AutoTrust: Benchmarking Trustworthiness in Large Vision-Language Models for Autonomous Driving. In *NeurIPS*, 2024. 2

[40] N. Tahir, Z. Zhang, M. Asim, J. Chen, and M. ELAffendi. Object Detection in Autonomous Vehicles under Adverse Weather: A Review of Traditional and Deep Learning Approaches. *Algorithms*, 2024. 1, 2

[41] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *ICML*, 2019. 1, 2

[42] Mark Taylor, Sarah Wilson, and David Chen. Efficient Vision-Language Processing for Real-time Applications. *IEEE Robotics and Automation Letters*, 8(4):4567–4578, 2023. 2

[43] Abhinav Valada et al. DADA-seg: Learning to Segment Adverse Driving Scenes. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1, 2

[44] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training. In *CVPR*, pages 15963–15974, 2024. 2, 7

[45] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015. 6

[46] J. Wang, S. Wang, M. Wu, H. Yang, Y. Cao, and S. Hu. Multi-scale adaptive detail enhancement dehazing network for autonomous driving perception images. *Pattern Analysis and Applications*, 2025. 1, 2

[47] M. Wang, H. Pi, R. Li, Y. Qin, Z. Tang, and K. Li. VLScene: Vision-Language Guidance Distillation for Camera-Based 3D Semantic Scene Completion. *arXiv preprint arXiv:2503.06219*, 2025. 2

[48] Emma Wilson, Robert Davis, and David Chen. Foundation models for autonomous driving: A survey and perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[49] Emma Wilson, Robert Davis, and David Chen. Out-of-Distribution Detection in Autonomous Driving: A Comprehensive Survey. *IEEE TPAMI*, 46(2):2345–2364, 2024. 2

[50] Sarah Wilson, Emma Park, and David Chen. Sensor Fusion with Language Models for Autonomous Driving. *IEEE Robotics and Automation Letters*, 8(4):2234–2245, 2023. 1, 2

[51] Sarah Wilson, David Chen, and Emma Taylor. Uncertainty Estimation in Vision-Language Models for Driving. *IEEE Robotics and Automation Letters*, 9(1):345–356, 2024. 1, 2

[52] Chunwei Wu, Zhuofan Liu, Yiwen Wang, et al. EMMA: End-to-End Multimodal Autonomous Driving with Vision and Language. In *CVPR*, 2024. 2

[53] D. Xiao, M. Dianati, and P. Jennings. HazardVLM: A Video Language Model for Real-Time Hazard Description in Automated Driving Systems. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 6

[54] Z. Xiong, Y. Wang, W. Yu, A. J. Stewart, and J. Zhao. GeoLangBind: Unifying Earth Observation with Agglomerative Vision-Language Foundation Models. *arXiv preprint arXiv:2503.06312*, 2025. 2

[55] M. Xu, D. Cai, W. Yin, S. Wang, X. Jin, and X. Liu. Resource-Efficient Algorithms and Systems of Foundation Models: A Survey. *ACM Computing Surveys*, 2025. 2

[56] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *CVPR*, pages 15952–15962, 2024. 7

[57] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, et al. BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling. In *CVPR*, 2020. 1, 2

[58] Kai Zhang, Hong Liu, Jie Wang, and Xi Chen. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. In *ICCV*, 2024. 1

[59] P. Zhang, X. Li, X. Hu, J. Yang, and L. Zhang. VinVL: Revisiting Visual Representations in Vision-Language Models. In *CVPR*, 2021. 6

[60] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *ICLR*, 2019. 6

[61] Wenxuan Zhang, Peter Anderson, Jiaya Jia, and Lei Zhang. Scene Understanding with Large Vision-Language Models: Benchmarks and Analysis. *IEEE TPAMI*, 46(2):1097–1112, 2024. 1, 2

[62] L. Zhao, L. Li, Z. Tan, A. Hawbani, and Q. He. Multi-agent deep reinforcement learning-based cooperative perception and computation in vec. *IEEE Internet of Things Journal*, 2025. 2

[63] W. Zhou, M. Tao, C. Zhao, H. Guo, and H. Dong. PhysVLM: Enabling Visual Language Models to Understand Robotic Physical Reachability. *arXiv preprint arXiv:2503.08481*, 2025. 2