

The Blind Spot of Adaptation: Quantifying and Mitigating Forgetting in Fine-tuned Driving Models

Supplementary Material

A. Overview

We begin by outlining the structure of this supplementary material. §B reports more details of our approaches. §C presents additional experimental results of our approach, offering further evidence that the proposed method effectively mitigates catastrophic forgetting. §D includes more detailed information and statistics regarding our dataset. §E provides extended qualitative comparisons and in-depth analyses. Finally, additional examples of catastrophic forgetting in driving models (Figs. 10 to 13), along with well-designed prompts and more samples from our dataset, are provided.

B. More Details of Driving Expert Adapter

Learning Prompt by Prompt Adapter. We maintain a bank of **learnable soft prompts** $\mathbf{P} \in \mathbb{R}^{K \times P \times D}$, where K is the number of prompt groups (bank size), P is the prompt length (number of soft tokens), and D is the model hidden size. Each prompt group $\mathbf{P}_k \in \mathbb{R}^{P \times D}$ can be interpreted as a latent "query family" capturing a core skill set or prior relevant to a subset of autonomous driving. All prompt tokens are optimized while the base VLM remains frozen.

Specifically, given token embeddings of the current input (e.g., question tokens after the text encoder embedding layer), $\mathbf{X} \in \mathbb{R}^{B \times T \times D}$, where B denotes the batch size (index b identifies the b -th sample in the batch) and T denotes the input sequence length in tokens (index t identifies the t -th token), we compute an input semantic summary by masked mean pooling:

$$\mathbf{s}_b = \frac{\sum_{t=1}^T m_{b,t} \mathbf{X}_{b,t}}{\sum_{t=1}^T m_{b,t}}, \quad \mathbf{s} \in \mathbb{R}^{B \times D}, \quad (1)$$

where $m_{b,t} \in \{0, 1\}$ is derived from the attention mask (padding tokens are excluded). We then map \mathbf{s} to prompt-selection logits with a lightweight MLP gate. These logits represent how relevant each prompt group is to the current input.

We convert the logits into a probability simplex \mathbf{w} via softmax. To promote specialization and reduce prompt interference, we optionally enforce sparse routing by retaining only the top- k prompt groups per sample. This procedure ensures that only the most relevant prompt groups receive non-zero probability mass (and thus gradients) for a given input.

The final prompt for each sample is obtained by a

weighted sum over the prompt bank:

$$\mathbf{P}_b^* = \sum_{k=1}^K w_{b,k} \mathbf{P}_k \in \mathbb{R}^{P \times D}. \quad (2)$$

In practice, we prepend \mathbf{P}^* to the original token embeddings (equivalently, we insert P learned *virtual tokens* before the textual tokens) and feed the concatenated sequence into the frozen VLM. Importantly, this adapter operates entirely outside the VLM’s core parameters: only the prompt bank \mathbf{P} and the lightweight gating network are trained, while the VLM backbone weights remain unchanged.

Why it helps? By learning a bank of reusable prompt primitives and routing them based on input-conditioned semantics, the adapter provides a stable task *anchor* that is robust to paraphrases and scene variations in autonomous driving. Sparse top- k routing encourages prompt groups to specialize, reducing cross-task interference and preventing the adapter from collapsing into a single generic template. Consequently, the VLM’s task interpretation becomes more consistent under novel phrasings, generalization improves, and catastrophic forgetting is effectively mitigated without modifying the initial weights.

C. More Details of Main Results & Ablation Study

The results in Tab. 1 provide compelling evidence that our framework remains consistently effective even when applied to larger scale models and another architectures. ImpromptuVLA [1] also fine-tunes Qwen2.5-VL-7B [2] with the larger 7B model demonstrating stronger performance. When scaling up to a larger backbone (Qwen2.5-VL-7B), our approach yields substantial gains across all tasks and still preserves roughly 80% of previously acquired knowledge. This demonstrates that our method maintains robust forgetting mitigation and generalization stability even under significantly increased model capacity. As a stronger base VLM, InternVL3-2B [3] exhibits superior performance over Qwen-based models across all three evaluation tasks, which also contributes to the strong results achieved by the fine-tuned RecogDrive [4]. However, despite its stronger perceptual and reasoning capacity, catastrophic forgetting still remain evident after adaptation to AD datasets. This motivates us to further apply our framework on InternVL3-2B to mitigate such forgetting while preserving its powerful base-model capabilities. On one hand, the Driving Expert Adapter (Ours) framework achieves the highest knowledge retention rate of

82.7%, while also improving performance on the Traffic-QA task by 11.6%. On the other hand, further experimental results, as shown in the Tab. 1, demonstrate that the Prompt Adapter and Task-Adaptive Expert Module still complement each other, enabling the model to acquire specialized driving skills without the destructive side effects of full-parameter updates across different model architectures.

Table 1. Evaluation results of different fine-tuning strategies on InternVL3-2B, Qwen2.5VL-3B and Qwen2.5VL-7B.

Method	KRR \uparrow	SD \uparrow	T-QA \uparrow	NoPR \uparrow
Recogdrive-2B	71%	63.7	42.2	36.1
Base (InternVL3-2B)	-	63.8	31.5	50.8
Base+LoRA	60.6%	61.8	25.1	30.4
Base+PA	77.9%	62.1	38.5	39.6
Base+TAEM	78.3%	62.5	37.4	39.3
Base+TAEM+PA	82.7%	63.7	43.1	42
ImpromptuVLA-3B	68.4%	59.1	33	25.2
Base (Qwen2.5VL-3B)	-	56.6	28.7	36.8
Base+LoRA	64.6%	53.1	22.3	23.8
Base+PA	73.2%	57.3	36.7	26.9
Base+TAEM	74.4%	57.7	35	27.4
Base+TAEM+PA	79%	58.8	41	29.0
ImpromptuVLA-7B	67.5%	60.8	46.3	32.2
Base (Qwen2.5VL-7B)	-	65.1	31	47.7
Base+TAEM+PA	79.8%	61.6	45.3	38.1

D. More Dataset Details

Autonomous driving models must cope with dynamic and complex environments, yet many prior works are trained on single-source datasets and often emphasize the quantity of QA pairs rather than the actual diversity of visual scenes. This narrow data coverage can lead to overfitting and catastrophic forgetting. As a first step toward addressing these limitations, we construct a large-scale, multi-source dataset that integrates 15 heterogeneous data sources, ensuring broad scene diversity and reducing the risk of overfitting and forgetting.

Whereas many works annotate the same underlying datasets (*e.g.*, nuScenes-QA [5], DriveLM [6]), we avoid this redundancy and instead select the following: LingoQA [7], SUTD-TrafficQA [8], CODALM [9], CoVLA [10], WOD-E2E [11], MAPLMQA [12], DriveAction [13], and ImpromptuVLA [1]. The last one includes data derived from eight additional datasets: Navisim [14], Argoverse [15], nuScenes [16], IDD [17], KITTI [18], Mapillary [19], ONCE [20], and Waymo [21].

Data Processing. In accordance with the task requirements, we make several data refinements. For multi-view datasets, we retain only the front-view images to ensure consistency in visual information. For sequential datasets, we select key frames while removing the corresponding redundant

QA pairs. In the case of CoVLA, which involves video annotations, we eliminate redundant frames and annotations to reduce duplication. Additionally, low-quality images in the SUTD-TrafficQA dataset are discarded. Finally, we employ Qwen3-VL-235B-Thinking [22] to annotate the WOD-E2E dataset, following the custom-designed prompt templates, as illustrated in the Fig. 5 below.

Data Statistics. We employ GPT-OSS-120B automatically analyze the QA pairs of each image. Then categorize objects in autonomous driving scenes into five main categories, which is Vehicle, Obstruction, Traffic-Related Element, Human-Related Element and Others. Fig. 7 presents a subset of our statistical results, revealing that most of the annotated objects are vehicles, with Mid-Size Vehicles constituting the majority within the vehicle category. A similar trend is observed within the Traffic-Related Elements, where traffic lights dominate the category. This significant class imbalance is a potential factor contributing to catastrophic forgetting. Additionally, due to the large variety of low-frequency within the remaining three categories, we visualize a part of terms from these categories in a word cloud.

Benchmark. We use Qwen3VL-235B to re-annotate and filter the most challenging scenarios, which were then used for evaluation in the Scene Description and Noteworthy Objects’ Perception tasks, as illustrated in the Figs. 14 to 16 below. Additionally, we integrate test sets from LingoQA [7] and DrivingVQA (derived from a driving theory exam) [23] to assess performance on the Traffic-QA task, with examples shown in the Figs. 17 and 18 below.

E. Qualitative Analysis of other case

Due to space constraints in the main paper, additional and more fine-grained qualitative analyses are provided in this section. In Example 1 (Fig. 1), where the critical objects are small and distant, our approach successfully detects the pedestrians crossing sign (missed by ImpromptuVLA) and even identifies a small dog, which neither ImpromptuVLA nor RecogDrive perceives. This highlights the superior fine-grained perceptual sensitivity of our method, particularly in handling small-pixel, low-salience cues. In Example 2 (Fig. 2), where illumination is severely limited, ImpromptuVLA merely notes the scene as dark, and RecogDrive detects only parked cars. Our approach, however, identifies both multiple vehicles and the traffic control elements ahead, providing the necessary situational awareness in challenging lighting conditions. Likewise, in Example 3 (Fig. 3) and Example 4 (Fig. 4), our approach not only correctly identifies all critical objects in the scene but also produces responses that are noticeably more concise and less redundant. This indicates that our model achieves higher perceptual efficiency, capturing essential visual cues while avoiding unnecessary descriptive overhead. In conclusion, our approach mitigates catastrophic forgetting during training, enabling the model

to acquire driving-specific competencies while preserving and enhancing perceptual capacity. This strengthened perception forms the foundation for more reliable downstream trajectory planning.



ImpromptuVLA:

The image shows a two-lane road with a solid yellow line on the left and a dashed yellow line on the right, indicating a no-passing zone on the left and a passing zone on the right. The road is surrounded by dense greenery, with trees and bushes lining both sides. The sky is clear and bright, suggesting it is daytime. There are **no visible vehicles or pedestrians in the scene**. The road appears to be in good condition, with no visible potholes or debris. The overall driving environment is calm and serene, with **no immediate obstacles or hazards**.

Recogdrive:

The driving scene is set on a two-lane road flanked by lush greenery and tall trees, suggesting a suburban or rural area. The road appears to be in good condition, with clear lane markings and no visible potholes or debris. There are no other vehicles in the immediate vicinity, indicating light traffic. A **pedestrian crossing sign** is visible on the right side of the road, but no pedestrians are currently present. The sky is clear, and the lighting suggests it is either early morning or late afternoon, with the sun casting long shadows. Overall, the scene is calm and conducive to a smooth drive, with **no immediate hazards or obstructions**.

Ours:

The scene depicts a multi-lane road bordered by dense greenery, with **a dog** positioned in the middle of the road ahead. **Pedestrian crossing signs** are also visible along the roadside, indicating areas where pedestrians may cross. The road is marked with lanes and includes a **'PED' designation**, signifying pedestrian-friendly zones. Given the presence of the dog on the road, which could potentially obstruct traffic or cause an accident, and the visibility of pedestrian crossing signs, it is advisable to decelerate and stop to ensure safety and avoid any potential hazards.

Figure 1. Qualitative Comparison Example 1.



ImpromptuVLA:

The image shows a nighttime driving scene with **limited visibility due to darkness**. The road appears to be a two-lane highway with a solid white line on the left and a dashed white line on the right, indicating the division between lanes. There are no visible road signs or other vehicles in the immediate vicinity. The headlights of the car illuminate the road ahead, revealing a clear path with **no immediate obstacles**. The surrounding area is **dark**, with no significant light sources other than the car's headlights. The overall scene suggests a quiet, possibly rural or less-traveled road at night.

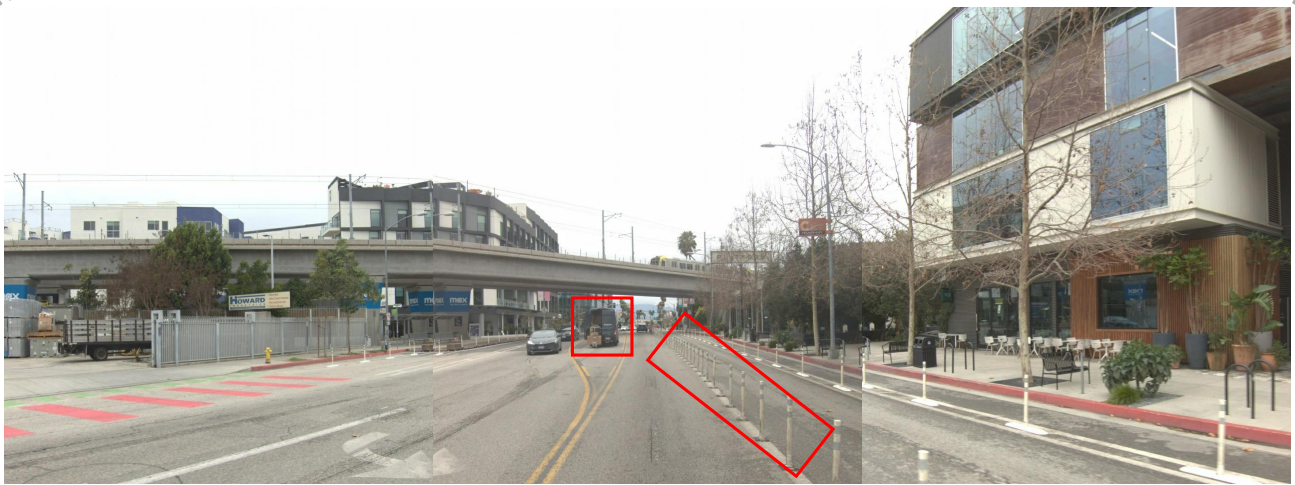
Recogdrive:

The scene is set at night, with limited visibility due to the darkness. The road ahead appears to be a two-lane street with **no visible lane markings**, suggesting a less-traveled or residential area. On the left side of the road, there are **parked cars** along the curb, indicating a residential or commercial area with buildings set back from the street. The right side of the road also has buildings, but they are less distinct due to the darkness. There are no moving vehicles or pedestrians in the immediate vicinity, and **the road ahead seems clear**. However, the lack of streetlights or other sources of illumination makes it difficult to see further down the road. The distant glow of lights suggests that there may be a more populated area or intersection ahead. Overall, the driving conditions require extra caution due to the low visibility and the potential for unexpected obstacles or changes in the road conditions.

Ours:

The scene is a nighttime driving environment on a residential street with low visibility. The road is illuminated by distant streetlights, and there are buildings on both sides with minimal lighting. **A parked vehicle** is visible on the left side of the road, and there are **traffic cones and a dirt pile in the middle of the road**, indicating potential construction or an obstacle. The road ahead appears clear but dark. The recommended actions are to keep the lane and decelerate due to the reduced visibility and the presence of obstacles on the road.

Figure 2. Qualitative Comparison Example 2.



ImpromptuVLA:

The image shows a daytime urban street scene with clear weather. The road is a multi-lane highway with a **yellow dividing line**, indicating a no-passing zone. There are **no visible vehicles** in the immediate lane ahead, suggesting light traffic. The road curves gently to the right. On the left side, there is a sidewalk with a few pedestrians and a truck parked near the curb. The right side features a modern building with large glass windows and a sidewalk with outdoor seating. The sky is overcast, and the overall lighting is even, with no harsh shadows. The scene is calm, with no immediate hazards or traffic signals visible.

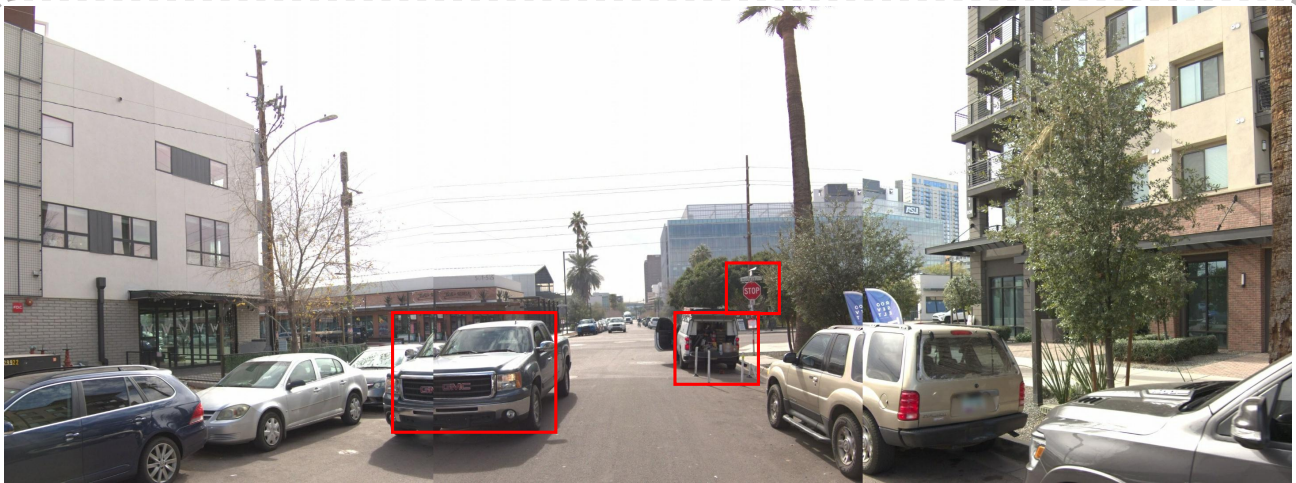
Recogdrive:

The driving scene is set in an urban environment with a clear view of the road ahead. The weather appears overcast, with no direct sunlight, suggesting it might be early morning or late afternoon. The road is relatively empty, with **only a few vehicles visible in the distance**, indicating light traffic conditions. On the right side of the road, there are modern buildings with glass facades, and a sidewalk lined with trees and outdoor seating areas, suggesting a commercial or mixed-use area. On the left side, there is a pedestrian bridge crossing over the road, connecting two buildings, and a fenced-off area, possibly under construction or a parking lot. The road markings are clearly visible, with a **double yellow line** indicating a no-passing zone. There are **no immediate obstacles or hazards** on the road, and the overall scene suggests a calm and safe driving environment. However, drivers should remain cautious and attentive to any potential changes in traffic conditions or pedestrian activity.

Ours:

The image depicts an urban street scene with a two-lane road marked by a **yellow center line and bollards separating the lanes**. An elevated overpass with a light rail train spans the road in the distance. The traffic includes **a black car in the opposite lane and a larger vehicle ahead**, which is likely a bus or truck. The road appears clear with no immediate obstacles, but there are buildings with commercial establishments lining both sides, including a modern building with outdoor seating on the right. The sky is overcast, indicating cloudy weather conditions. The recommended actions are to keep the current lane and follow the lead vehicle ahead.

Figure 3. Qualitative Comparison Example 3.



ImpromptuVLA:

The image shows a city street scene with a clear view of the road ahead. The street is lined with parked cars on both sides, and there are no visible pedestrians or moving vehicles in the immediate vicinity. The road appears to be in good condition, with no visible potholes or debris. **The traffic light ahead is red**, indicating that you should stop and wait for the light to turn green before proceeding. The buildings on either side of the street are modern and well-maintained, with large windows and flat roofs. The sky is overcast, suggesting that it may be a cloudy day. The overall scene is calm and orderly, with no signs of congestion or chaos.

Recogdrive:

The driving scene is set in an urban area with a mix of commercial and residential buildings. The road is lined with parked cars on both sides, **including a variety of vehicles such as sedans, SUVs, and a pickup truck**. There is a **stop sign** visible on the right side of the road, indicating an upcoming intersection where the vehicle must come to a complete stop. The road appears to be relatively clear of moving traffic ahead, with a few cars visible in the distance. Palm trees and other greenery are interspersed among the buildings, adding a touch of nature to the urban landscape. The sky is overcast, suggesting it might be a cloudy day. Overall, the scene is calm and orderly, with no immediate hazards or obstacles in the immediate path of the vehicle.

Ours:

The image depicts an urban street scene with a mix of residential and commercial buildings on both sides. There are multiple vehicles parked along the curbs, and a few cars are driving on the road. A red octagonal **stop sign** is visible ahead at an intersection, indicating that the vehicle should come to a complete stop. The road appears to be a two-lane street with a **GMC truck in the left lane** and a **parked SUV on the right**. The presence of the stop sign necessitates that the driver stops before proceeding through the intersection.

Figure 4. Qualitative Comparison Example 4.

Suppose you are driving and I provide you with images captured by the car's front camera. Generate a description of the driving scene that includes the following key factors for driving planning:

- **The positions and movements of vehicles and pedestrians.**
- **The prevailing weather conditions.**
- **The time of day, distinguishing between daylight and nighttime.**
- **Road conditions, indicating whether the surface is smooth or if there are obstacles present.**
- **The status of traffic lights, specifying whether they are red or green, as they influence decision-making.**

The description should be concise and provide an accurate understanding of the driving environment to facilitate informed decision-making.

Figure 5. The prompt used to label WOD-E2E Dataset

You are an information extraction assistant for autonomous driving.

Your input is:

- a QUESTION about a driving scene, and
- an ANSWER given by a model for that scene.

Carefully read both the question and the answer, and extract ALL explicit, useful information entities mentioned in them.

Treat as "information entities" any short phrase that encodes concrete, reusable information about the scene or the driving task, including but not limited to:

- **Dynamic agents:** vehicles, pedestrians, cyclists, animals, emergency vehicles, etc.
- **Static elements:** road type (highway, intersection, roundabout), lanes, sidewalks, crosswalks, traffic islands, barriers, buildings, vegetation, etc.
- **Traffic infrastructure:** traffic lights, signs, road markings, speed limits, lane arrows, stop lines, yield signs, etc.
- **Environmental conditions:** time of day, weather, visibility, lighting, road surface condition (wet, icy, damaged, with puddles, etc.).

Rules:

- Use only information explicitly stated or clearly implied by the question or the answer.
- Do NOT invent entities that are not supported by the text.
- Be as fine-grained as necessary: separate different entities (e.g., "red traffic light", "pedestrian crossing the road", "wet road surface" should be three separate items).

Output format:

- Return ONLY a valid Python list of strings.
- Each list element should be a concise English phrase describing exactly ONE information entity.
- Do not include any explanations, comments, or extra text outside the Python list.

Now process the following input and output the Python list:

QUESTION: {QUESTION}
ANSWER: {ANSWER}

Figure 6. The prompt used to extract information from QA pairs

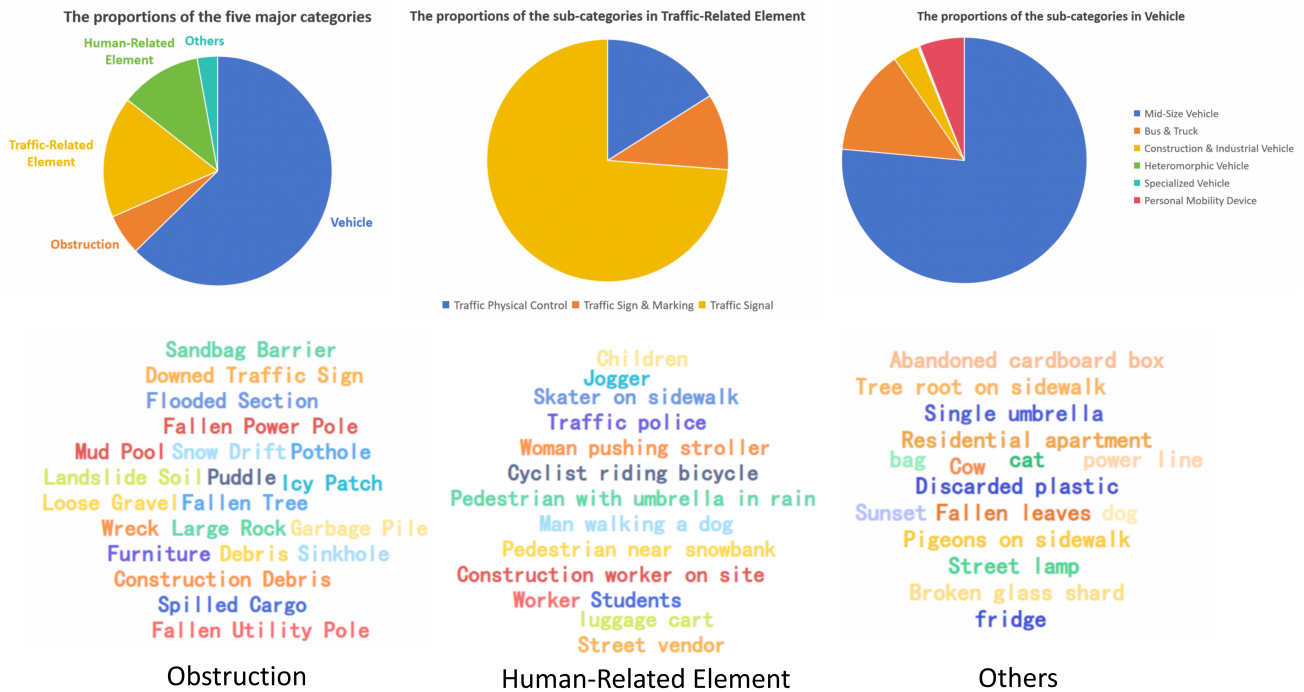


Figure 7. Partial statistical results

Scene Description Judge

Suppose you are driving, the student gives a description of the driving scene.

Please evaluate the student's answer on a scale from 0 to 100 according to your deeply understanding in "Reference answer".

The higher score reflects precise alignment with the "Reference answer", well-supported reasoning, detailed description of the environment and accurate identification of road condition and objects.

Deduct points for minor inaccuracies, omissions, or lack of clarity. Distribute the Total Score across the following criteria:

1. Accuracy and Alignment (50 points):

Assign up to 20 points based on how the Student's answer correctly aligns with the description of the environment, road condition and objects referred in the Reference answer.

2. Conciseness and Clarity (30 points):

Punish those inaccurate, ambiguous and irrelevant Student's answer that have nothing to do with the facts.

3. Bonus (20 points):

If the Student's answer are correctly aligned in concise language, an additional bonus of 0-20 points will be given depending on the Clarity and Correctness level.

Reference answer: {ref_answer}

Student's answer: {stu_answer}

Figure 8. The prompt of Scene Description Judge

Traffic-QA Judge

You are compassionate but fair and strict AI referee to judge **True or False** of student's answer. Suppose you are driving, the student gives an answer of a question related to the driving scene. Please judge the student's answer according to your deeply understanding in "Reference answer".

Question: {question}

Reference answer: {ref_answer}

Student's answer: {stu_answer}

Noteworthy objects Perception Judge

Task: Tell me whether the following object is present or implied in the described scene.

Object to check: {object}

Scene description: {scene_description}

Instruction: Does the object "{object}" appear explicitly or is it implicitly implied based on the description?

Answer with one of the following:

Yes, explicitly present

Yes, implicitly implied

No, not present or implied

Figure 9. The prompt of Traffic-QA and Noteworthy objects Perception Judge

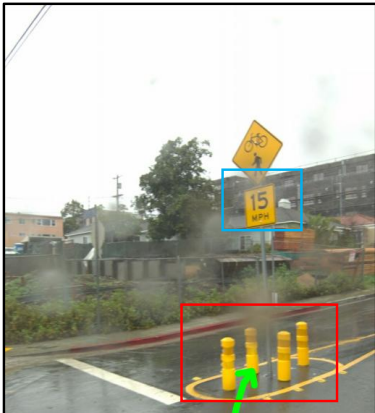




	<p>InternVL3-8B:  </p> <p>Based on the image, here are the important objects to consider for driving decisions:</p> <ol style="list-style-type: none">1. Speed Limit Sign (15 MPH): Indicates the maximum speed you should maintain in this area.2. Bicycle and Pedestrian Sign: Suggests the presence of cyclists and pedestrians in the area, requiring extra caution.3. Yellow Bollards: Mark the edge of the road or a designated area, possibly for pedestrian safety.4. Weather Conditions: It appears to be raining, which may affect visibility and traction.5. Road Surface: The road is wet, which can increase stopping distances.6. Surrounding Vegetation and Fencing: May indicate construction or restricted areas ahead.7. Sidewalk and Red Curb: Suggests no parking or stopping zones, important for pedestrian safety.
	<p>Finetuned ReCogDrive VLM:  </p> <p>The objects that are important for planning the driving decision include the speed limit sign indicating 15 MPH, the pedestrian crossing sign, the parked cars on both sides of the street, and the wet road conditions due to rain. These factors suggest the need for reduced speed, increased vigilance for pedestrians, and careful maneuvering to avoid potential hazards such as opening car doors or pedestrians stepping out from between parked vehicles.</p>

Figure 10. Another example of catastrophic forgetting in Recogdrive.



WiseAD: The traffic is a **motorbike, a cyclist, a truck, and a bus** in the opposite lane.

Base VLM(MobileVLM V2 1.7B):In the image, **a herd of cattle** is crossing a road in front of a car. The car is driving on the right side of the road, and the cattle are on the left side of the road. The road is surrounded by a hillside and a mountain range in the background. The sky is cloudy and the weather appears to be overcast. The image captures a moment of life on the road, with the cattle moving in harmony with the flow of traffic.

Figure 11. **A example of catastrophic forgetting in WiseAD.** The fine-tuned model incorrectly recognizes cattle as vehicles, illustrating a severe case of catastrophic forgetting.



Mini-InternVL2-4B-BDD: The car is driving down the street.

Mini-InternVL2-4B-DriveLM: There is one moving car in front of the ego car, one parked car to the front left of the ego car.

Base(InternVL2-4B): The image depicts a driving scene on a multi-lane road with a clear blue sky overhead. In the foreground, there is a white SUV with a roof rack, positioned on the left side of the road. The traffic light ahead shows a red signal, indicating that vehicles in this lane must stop. To the right of the SUV, there is a green traffic light, allowing vehicles in that lane to proceed. The road is flanked by lush greenery on both sides, with trees and shrubs providing a natural backdrop. In the distance, modern high-rise buildings with reflective glass facades can be seen, suggesting an urban setting. The road appears to be well-maintained, with clear lane markings, and there are no visible pedestrians or other vehicles in the immediate vicinity. The overall atmosphere is calm and orderly, with ample daylight illuminating the scene.

Figure 12. A example of catastrophic forgetting in Mini-Intern. After fine-tuning, the model perceives significantly fewer scene elements and exhibits mechanically repetitive outputs, indicating a substantial loss of perceptual richness.



ImpromptuVLA-7B: The image shows a residential street with a clear view of the road ahead. The street is lined with parked cars on both sides, and there are trees and houses visible. The road appears to be in good condition, with no visible potholes or debris. The sky is clear and blue, indicating good weather conditions. There are no pedestrians or other vehicles in the immediate vicinity. The stop sign is visible on the right side of the road, indicating that you should come to a complete stop before proceeding. The overall scene suggests a calm and quiet neighborhood, with no immediate hazards or obstacles.

Base(QwenVL2.5-7B): The ego vehicle is at a residential intersection with clear weather. The scene includes a **stop sign**, a street sign for 'California', parked vehicles along the road, trees lining the streets, and pedestrians (**three people and a dog**) crossing the crosswalk ahead. The road ahead is straight with multiple parked cars and houses on both sides

Figure 13. A example of catastrophic forgetting in ImpromptuVLA. The fine-tuned model loses fine-grained perceptual capability.



Scene_Description: The scene captures a daytime driving environment on a two-lane road under an overcast sky, with diffused light creating a muted atmosphere. The asphalt road, marked by a faded yellow dividing line, stretches ahead with a silver sedan approaching in the opposite lane on the left, while on the right shoulder, a green utility vehicle lies overturned, its rear end crushed and debris scattered across the pavement. The dense, neatly trimmed hedges and tall, slender trees line the roadside, interspersed with utility poles and streetlights. In the distance, another vehicle is visible on the horizon, and the road is flanked by well-maintained greenery, including flowering bushes on the left median, while the hazy sky and lack of direct sunlight contribute to a somber mood, emphasizing the disruption caused by the accident amid an otherwise calm, suburban-style thoroughfare.

Noteworthy_Objects: [silver sedan in the opposite lane on the left, overturned utility vehicle]

Figure 14. The example of Scene Description and Noteworthy objects Perception



Scene_Description: The scene is a nighttime urban road captured from a car's front camera, with the sky completely dark and the road illuminated by the vehicle's headlights. A person riding a bicycle is visible ahead in the middle of the two - lane street, wearing light - colored clothing that stands out against the dark surroundings. The road has clearly marked white lane lines, and to the left, residential houses with one displaying a lit exterior light are accompanied by trees and a sidewalk. On the right, a large, leafy tree overhangs the street next to a fence and a sidewalk. In the distance, scattered streetlights and the faint glow of other vehicles' headlights provide minimal illumination, while silhouettes of utility poles and palm trees are visible against the dark sky, creating a quiet, suburban nighttime atmosphere with the cyclist as a notable presence in the road ahead.

Noteworthy_Objects: [a person riding a bicycle]

Figure 15. The example of Scene Description and Noteworthy objects Perception



Scene_Description: The scene depicts a rainy urban street viewed from a car's front camera, with the wet pavement reflecting the lights of vehicles and surroundings, creating a glossy sheen. A large yellow utility truck with bright amber warning lights dominates the center of the road, likely causing a traffic slowdown; several cars, including a white SUV and a silver sedan, are queued behind it. To the right, blue and black recycling bins sit near the curb beside a white minivan, while on the left, a person in an orange raincoat appears to be directing traffic or assisting near the truck. Residential houses with sloped roofs and palm trees line the street, with utility poles and crisscrossing power lines visible against an overcast, gray sky. The overall atmosphere is damp and muted, with raindrops likely on the windshield, enhancing the reflective quality of the road and contributing to the sluggish, cautious mood of urban driving in poor weather.

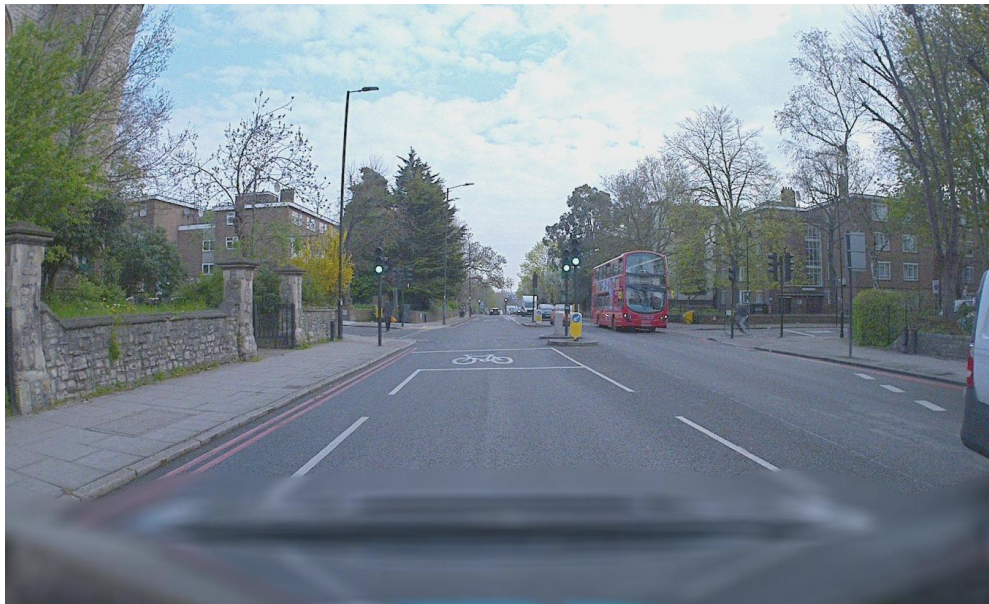
Noteworthy_Objects: [rainy urban street, yellow utility truck, recycling bins, a person in an orange raincoat]

Figure 16. The example of Scene Description and Noteworthy objects Perception



Question: Are there any pedestrians crossing the road? If yes, how many?

Answer: Yes, there is one jaywalker crossing the road right in front of us and then there is a pedestrian further away, but we shouldn't worry about him yet.



Question: Are you permitted to make a right turn at the upcoming junction?

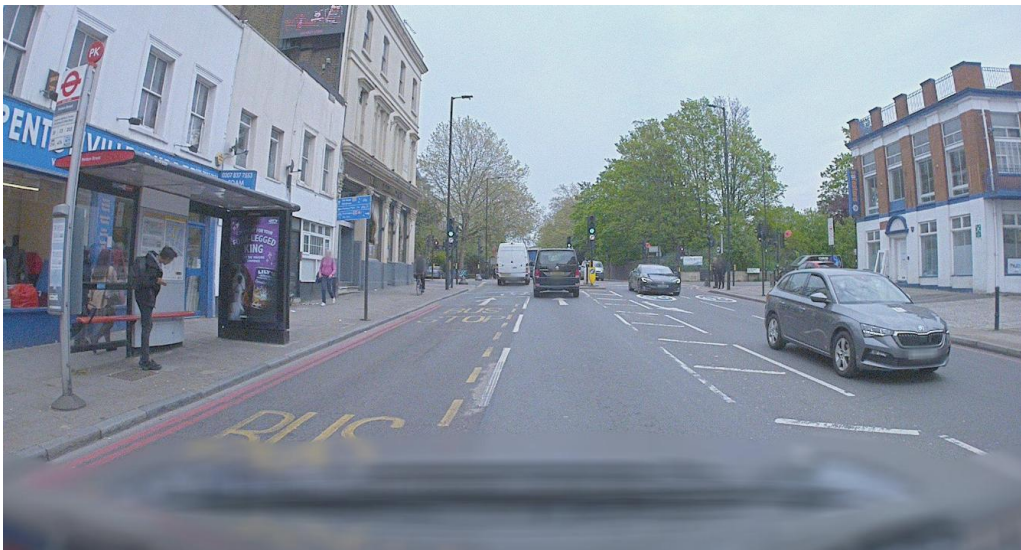
Answer: Yes

Figure 17. The example of Traffic-QA



Question: What action do you plan to take next and why?

Answer: Move into the right lane and accelerate to safely overtake the cyclist with plenty of space.



Question: Is it safe to enter the intersection at this time?

Answer: It is safe to enter the intersection because the traffic light is green.

Figure 18. The example of Traffic-QA



Figure 19. More examples

References

- [1] Haohan Chi, Huan ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, Leichen Wang, Xingtao Hu, Hao Sun, Hang Zhao, and Hao Zhao. Impromptu vla: Open weights and open data for driving vision-language-action models, 2025. 1, 2
- [2] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1
- [3] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1
- [4] Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv:2506.08052*, 2025. 1
- [5] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023. 2
- [6] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer, 2024. 2
- [7] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024. 2
- [8] Li Xu, He Huang, and Jun Liu. Sutr-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events, 2021. 2
- [9] Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7817–7826. IEEE, 2025. 2
- [10] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1933–1943. IEEE, 2025. 2
- [11] Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Drago Anguelov. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios, 2025. 2
- [12] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, et al. Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21819–21830, 2024. 2
- [13] Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, Peng Jia, et al. Driveaction: A benchmark for exploring human-like driving decisions in vla models. *arXiv preprint arXiv:2506.05667*, 2025. 2
- [14] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024. 2
- [15] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 2
- [16] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. 2
- [17] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1743–1751. IEEE, 2019. 2
- [18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2
- [19] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Buló, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 2
- [20] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 2
- [21] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2
- [22] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. 2
- [23] Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. Retrieval-based inter-

leaved visual chain-of-thought in real-world driving scenarios,
2025. [2](#)