NexusAD: Exploring the Nexus for Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

Supplementary Material

6. Formatting Guidelines

6.1. Detection and Depth Information

```
{"vehicles":{
    "category1":[{
        "bounding_box": [x1,y1,x2,y2],
        "range": "long range"}, {xxx}],
    "category2":[{xxx}, {xxx}]},
"vulnerable_road_users": {{xxx}, {xxx}},
"traffic signs": xxx,
"traffic lights": {},
"traffic cones":{},
"barriers": {xxx}},
"other objects":{}}.
```

6.2. Category-based Perceptual Description

```
{"vehicles": [{
    "description": xxx, "explanation": xxx}],
  "vulnerable_road_users": [{
    "description": xxx, "explanation": xxx}],
  "traffic_signs": [{
    "description": xxx, "explanation": xxx}],
  "traffic_lights": [{
    "description": xxx, "explanation": xxx}],
  "traffic_cones": [{
    "description": xxx, "explanation": xxx}],
  "barriers": [{
    "description": xxx, "explanation": xxx}],
  "other_objects": [{
    "description": xxx, "explanation": xxx}]}.
```

7. Prompts

7.1. General Perception

7.1.1 Task Description

You are an advanced Vision-Language Model integrated into an autonomous driving system. Your goal is to analyze the provided image in a traffic scenario during {period}, in {weather} weather and produce a detailed and accurate description based solely on the detected objects in the image. Your primary focus is to describe each object's **appearance**, **position**, and **direction**. Additionally, you must explain why each object affects the ego car's driving behavior, particularly considering safety, lane positioning, and potential hazards like parked vehicles, pedestrians, and work zones.

7.1.2 Step 1: Study Examples

Before analyzing the target image, carefully study the following two example detection results and their corresponding outputs. These examples illustrate the expected level of detail and structure. Your goal is to replicate this structure by providing accurate object descriptions with a clear focus on **appearance**, **position**, **direction**, and the **impact** each object has on the ego car's driving behavior, specifically noting the object's category.

Example General Perception Output
1: {example perception output 1}
2: {example perception output 2}

IMPORTANT: Your output must follow the structure demonstrated in the examples. Ensure that every detected object is described precisely with details on appearance, position, and direction, followed by a well-reasoned explanation of its impact on the ego car's behavior. Avoid including objects not present in the detection results.

7.1.3 Step 2: Target Image Analysis

TARGET IMAGE: <image>

The detection results for the target image are as follows, categorized by object type: {detection results}

Object Descriptions and Explanations: For each detected object, provide a detailed description based on the following structure:

1. **Appearance:** Describe the object's color, type (e.g., white parked vehicles, pedestrian with a yellow jacket, orange traffic cones), and any distinguishing features (e.g., debris, construction materials).
2. **Position:** Specify the object's location relative to the ego car (e.g., parked to the left, walking alongside the road, marking the right edge of the lane).
3. **Direction:** Indicate the object's direction of movement (e.g., stationary, walking parallel to the road).
4. **Impact on Ego Car:** Explain how this object affects the ego car's driving behavior.

Consider factors such as pedestrians crossing the road, parked vehicles reducing lane width, and construction zones requiring lane adjustments.

- **Traffic Control Devices and Signs:** Describe traffic signs and their effect on driving behavior (e.g., pedestrian crossings requiring caution). If signs are blurred or not clearly visible, note their presence but clarify that they cannot be used to guide driving decisions.
- **Other Objects:** Mention environmental objects like debris or roadside objects. If they are not an immediate hazard but could potentially become one, describe how the

ego vehicle should monitor them while passing.
- **Scenario Summary:** After describing individual objects, generate a comprehensive summary of the scene. Integrate all relevant objects by category and explain their collective influence on the ego car's behavior, focusing on how these elements contribute to the overall traffic environment and decision-making process.

7.1.4 Suppression of Hallucinations

- Focus solely on the objects detected in the target image. Avoid adding any object descriptions or explanations that are not explicitly mentioned in the detection results.
- **Hallucinations** occur when objects not present in the detection results are erroneously included in the response. Before finalizing your description of any object, confirm that it is based on the detection results.
- If a detected object category is more abstract (e.g., 'car' for an SUV), use contextual visual cues from the image to provide more detailed descriptions where appropriate, but do not introduce new objects that are not clearly visible.
- For objects like flashing lights or reflective surfaces (e.g., vehicle rear lights), ensure that the interpretation is grounded in detected data, and provide a plausible explanation based on context (e.g., rear lights flashing during braking).

7.1.5 Response Format

Provide the output in the following simplified JSON format. {`Category-based Description`} For each category, include objects detected, or leave the list empty if none were found. Ensure descriptions and explanations reflect actual observations from the detection results. Note: If a category has no detected objects, return an empty list for that category.

7.2. Regional Perception

7.2.1 Task Description

You are an great traffic object analysis model. Analyze the object within the red rectangle in a traffic image, focusing on the provided object category {`label name`}, and explain its potential impact on the ego car's driving behavior. Your explanation should consider specific factors such as safe distance, visibility, potential obstacles, and necessary maneuvers. The explanation should be comprehensive, relevant, and directly related to the characteristics and typical behavior of the identified object in the traffic scenario.

7.2.2 Skills

- Accurately identify and describe the given object category {`label name`} in the traffic image.
- Provide a detailed and contextually relevant explanation of how this object might influence the ego car's driving behavior.
- Ensure the analysis is consistent with the object's typical role and behavior in traffic, including specific considerations like visibility, potential obstacles, and necessary driving maneuvers.

7.2.3 Rules

1. Accurately describe the object within the red rectangle, strictly adhering to the provided category {`label name`}.
2. Provide a detailed explanation of how the {`label name`} might influence the ego car's driving behavior, considering factors such as safe distance, visibility, potential obstacles, and required maneuvers.
3. Include context-specific considerations, such as lighting conditions, road positioning, and potential interactions with other road users.
4. Ensure that the explanation is relevant to the identified object's role in the traffic scenario, avoiding assumptions not supported by the context.

7.2.4 Workflows

1. Identify the object within the red rectangle based on the provided category {`label name`}.
2. Describe the object's characteristics and typical behavior in traffic, focusing on how it affects driving decisions.
3. Provide a detailed and contextually relevant explanation of how this object might influence the ego car's driving behavior, ensuring the analysis is aligned with the specific role and behavior of {`label name`} in traffic situations.

7.3. Driving Suggestions

7.3.1 Task Description

You are the autonomous driving system of a vehicle (ego car). Your task is to analyze three traffic scenario images and generate a comprehensive, detailed, and actionable driving suggestion specifically for the **TARGET IMAGE**. Your final output should fully explore all relevant aspects, providing clear and precise guidance. The scenarios may include extreme or unusual driving conditions, requiring careful analysis and attention to detail.

7.3.2 Step 1: Review Example Driving Suggestions

- **Example 1:**
  - Detection Results: {`detection results 1`}
  - Example Driving Suggestion: {`example driving suggestion 1`}

- **Example 2:**
  - Detection Results: {`detection results 2`}
  - Example Driving Suggestion: {`example driving suggestion 2`}

IMPORTANT: These examples are provided solely to help you understand the structure, depth, and detail expected in a driving suggestion. In the following steps, focus entirely on the **TARGET IMAGE** and its specific conditions. **Your response should be as comprehensive and detailed as the examples provided, thoroughly addressing all relevant factors.**

7.3.3 Step 2: Analyze Target Image

TARGET IMAGE: `<image>`
- Detection Results: {`detection results`}

Task: Identify and describe all significant objects and environmental factors in the **TARGET IMAGE** that could affect the ego car's driving behavior. These may include:
- **Vehicles** Describe the types and positions of vehicles (e.g., trucks, buses, cars) and their potential behaviors (e.g., braking, turning, lane changes). Pay special attention to load stability, signals, and proximity.
- **Traffic Control Devices:** Discuss how the ego car should interact with traffic lights, signs (e.g., speed limits, no parking, pedestrian crossings), and road markings, providing specific guidance on compliance and caution.
- **Environmental Factors:** Examine the influence of lighting conditions, road surface quality, weather elements, and visibility on driving behavior, and provide strategies for safe navigation.
- **Potential Hazards:** Identify and provide strategies for dealing with any hazards such as debris, roadwork, or sudden stops, and provide detailed instructions on how the ego car should respond to each. If detection results do not include traffic cones, it is certain that there are no traffic cones; similar objects are likely obstacles.

7.3.4 Step 3: Prioritize Safety in Response to Identified Objects

Task: For each identified object or condition in the **TARGET IMAGE**, provide detailed instructions on:
- **Maintaining Safe Distances:** Explain how the ego car should maintain safe distances from other vehicles, especially in complex or unpredictable traffic conditions. Include any specific actions to take in response to potential falling debris or unstable loads.
- **Speed Adjustments:** Offer specific guidance on speed adjustments in response to traffic lights, vehicle behavior, environmental conditions, or potential hazards. Include instructions on how to adhere to posted speed limits or adjust speed for safety.
- **Lane Positioning and Maneuvers:** Provide detailed instructions on lane positioning and maneuvers, including when and how to change lanes or adjust position within the lane to avoid obstacles and maintain safety. Emphasize the importance of checking mirrors and blind spots before any maneuvers.

7.3.5 Step 4: Integrate Traffic Control and Environmental Conditions

Task: Integrate the influence of traffic control devices and environmental conditions into the driving suggestion, with specific guidance on:
- **Traffic Light Response:** Offer precise instructions on how to react to the current state of the traffic lights (e.g., when to stop, proceed, or prepare for changes in light status). Explicitly mention how to proceed under green, yellow, or red lights.
- **Adaptation to Conditions:** Provide detailed strategies for adapting driving behavior according to weather conditions, visibility, and road surface quality, ensuring the ego car navigates safely through the environment. Emphasize the use of headlights, windshield wipers, and other tools to improve visibility and safety.

7.3.6 Step 5: Generate Final Driving Suggestion

Task: Combine all insights into a single, comprehensive driving suggestion. Ensure that:
- **Safety:** The suggestion prioritizes safety, with thorough instructions on speed, distance, lane positioning, and response to hazards.
- **Detail and Precision:** The suggestion should be detailed and precise, addressing all relevant aspects of the **TARGET IMAGE** in a thorough and comprehensive manner. Include specific actions based on the vehicle signals and any potential changes in the environment.
- **Clarity:** The suggestion should be clear and easy to understand, avoiding unnecessary simplifications and providing actionable guidance.

7.3.7 Final Output

The final output should be a detailed, thorough, and actionable driving suggestion that fully addresses all relevant aspects of the **TARGET IMAGE**. It should provide clear, specific, and comprehensive guidance on how to safely navigate the scenario, considering all environmental factors, vehicle behaviors, and potential hazards. The response must be aligned with the detailed analysis conducted in the previous steps and should provide actionable steps for safe driving in the identified conditions.

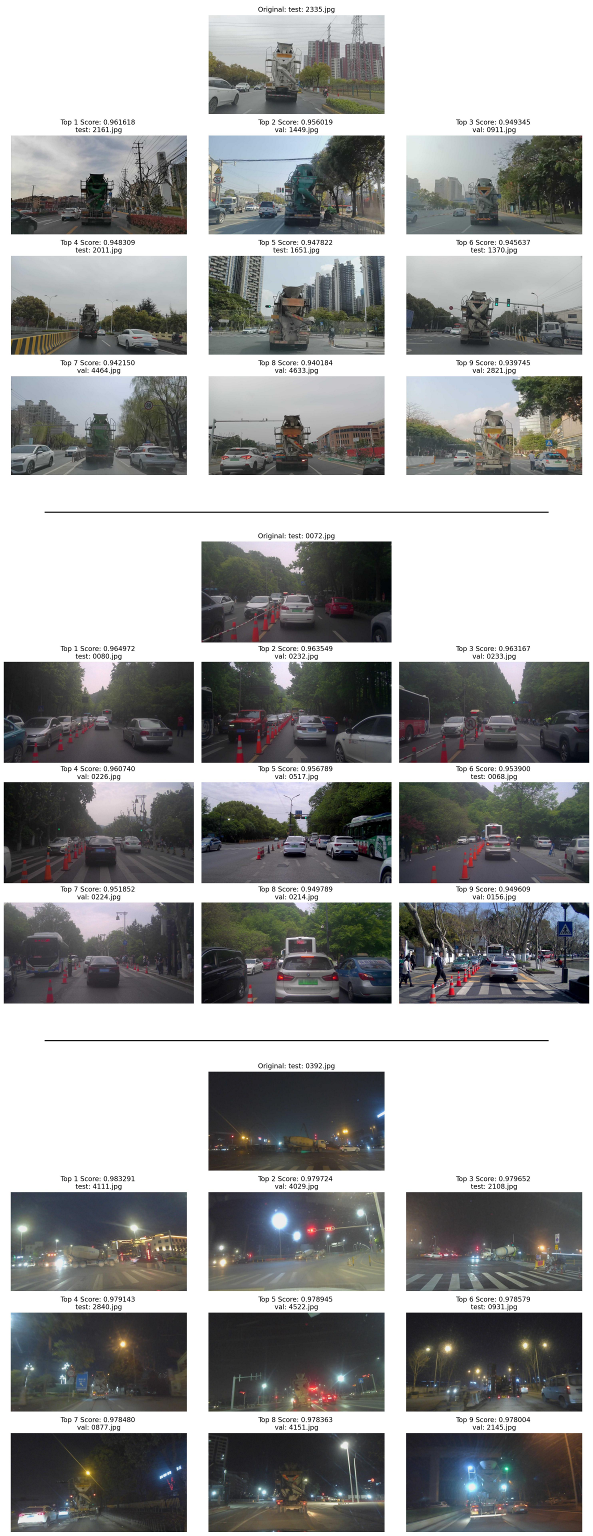

Figure 2. Several sample pairs for multiple searches. The top row displays the query sample, and the subsequent rows show the top-ranked retrieval results.

8. Retrieval Visualization Analysis

Figure 2 illustrates the effectiveness of our scene-aware retrieval-augmented enhancement. Our approach not only identifies visually similar samples but also selects those with comparable scene contexts. This method significantly improves driving perception and suggestion accuracy by leveraging the most relevant contextual information.

9. Casual Musings

Due to the absence of ground truth values for the test set, we conducted local validation using a mini subset of the CODA-LM dataset, following official prompts and procedures. To manage costs, we utilized the ChatGPT-4O-mini model. Variations in evaluation models and subsets may lead to discrepancies between our local results and leaderboard scores. Therefore, the subsequent results should be considered preliminary.

Table 2. Comparison of scores between directly inputting depth maps and using linguistic forms to provide depth information

Depth Map	Language	Score
✗	✗	4.10
✗	✓	4.36
✓	✗	4.58

As shown in Table 2, directly inputting the depth map into vision-language model may achieve better results. However, considering the number of tokens spent on the image, the huge cost of fine-tuning, and the illusion problem caused by multiple inputs, we still use linguistic form of depth information to guide the model's perception and comprehension in this competition.

Table 3. Score comparison between coherent text and structured json formats.

Coherent Text	Structured JSON	Score
✓	✗	5.98
✗	✓	6.40 (+0.42)

From the results in Table 3, it seems that ChatGPT prefers structured output and gives higher scores when rating. This may be a small trick to increase scores. In this competition, we ultimately submitted using the same coherent natural language format as the dataset annotation.

Finally, Direct Preference Optimization (DPO) is a method for aligning language models with human preferences. Although we didn't apply DPO in this competition due to time constraints, we are excited about its potential and consider it a promising direction for future research.