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

Track 1 | Corner Case Scene Understanding

Trustworthy Visual Intelligence Group

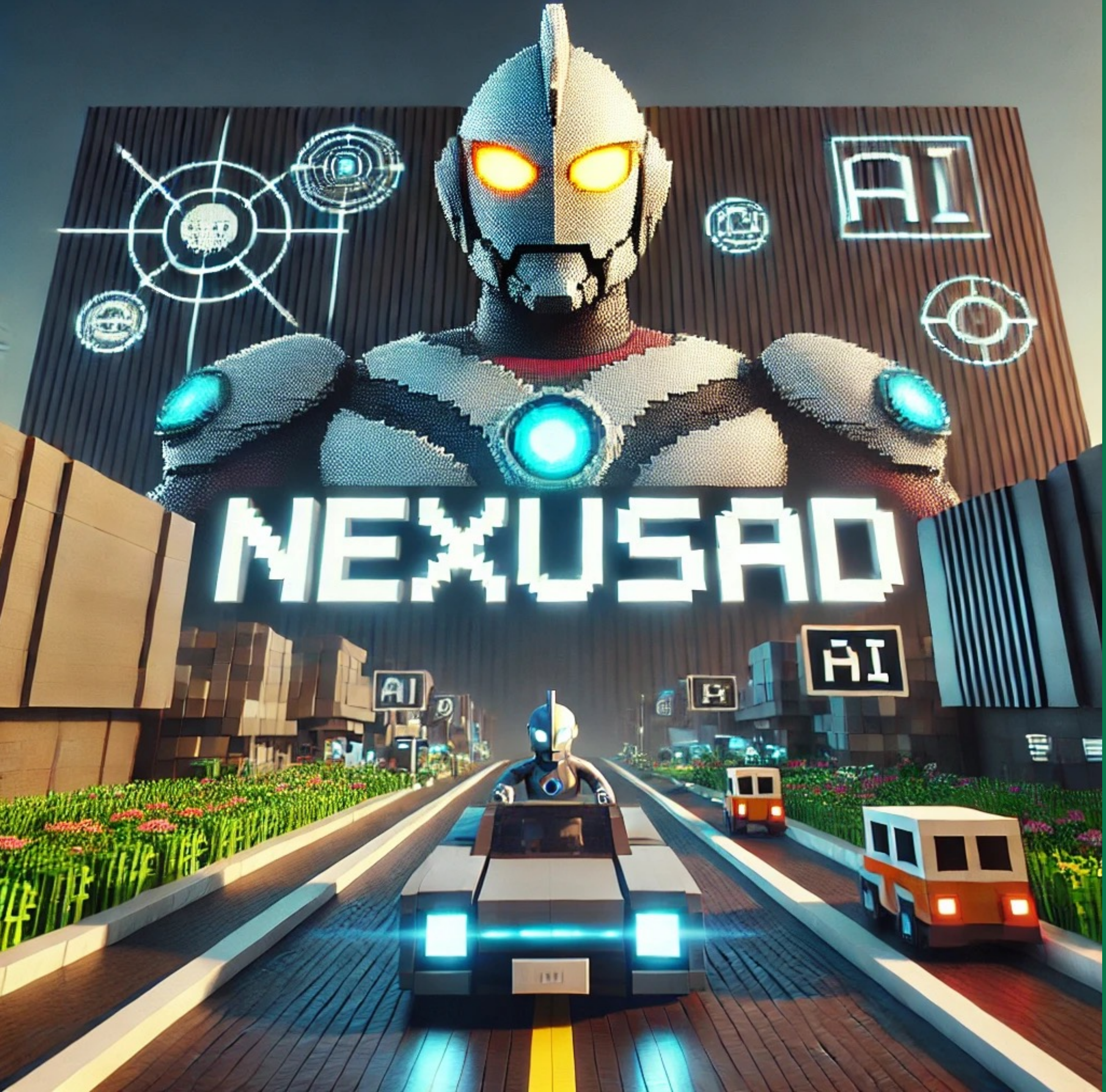
Mengjingcheng Mo, Jingxin Wang, Like Wang, Haosheng Chen, Changjun Gu, Jiaxu Leng *, Xinbo Gao*

Chongqing University of Posts and Telecommunications

Speaker: Mengjingcheng Mo

Workshop W-CODA

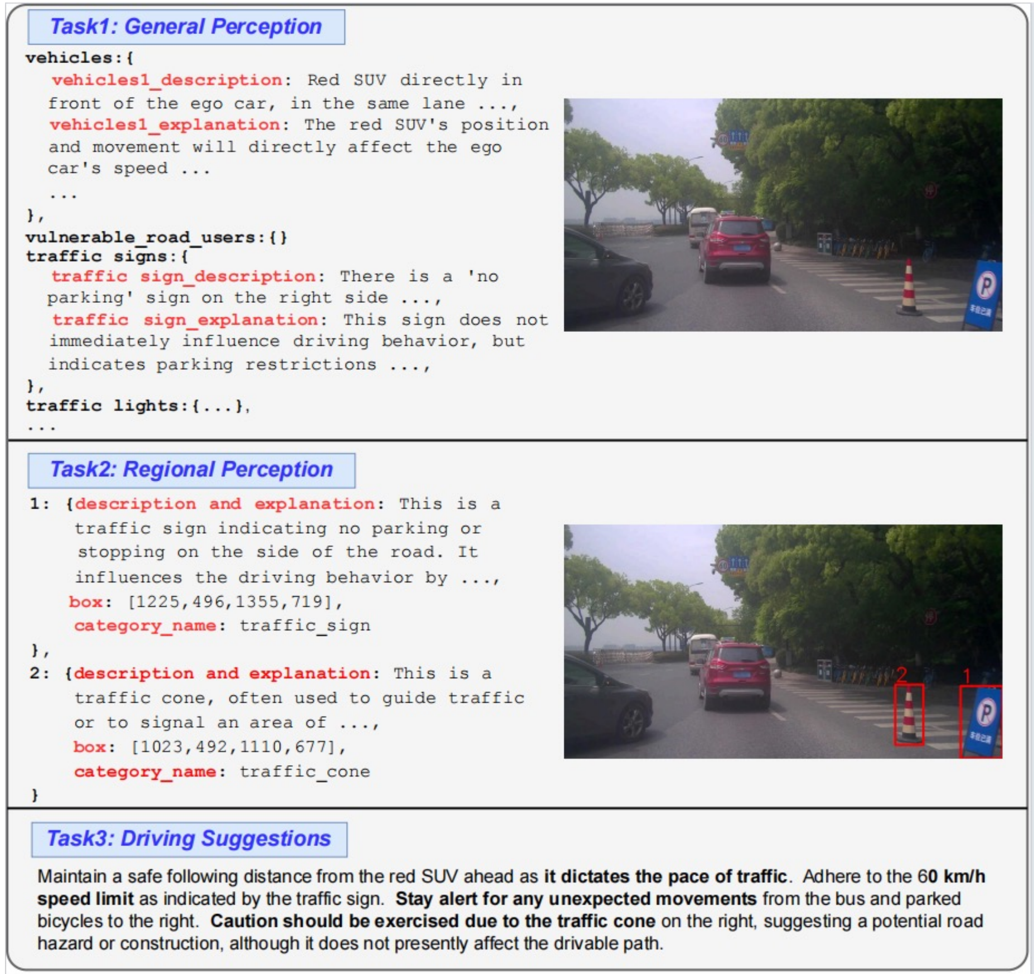
Sep. 30th, 2024 (Mon)



Outline

- Dataset & Tasks
- Challenge
- Method
- Analysis
- Results

/01 Dataset & Tasks

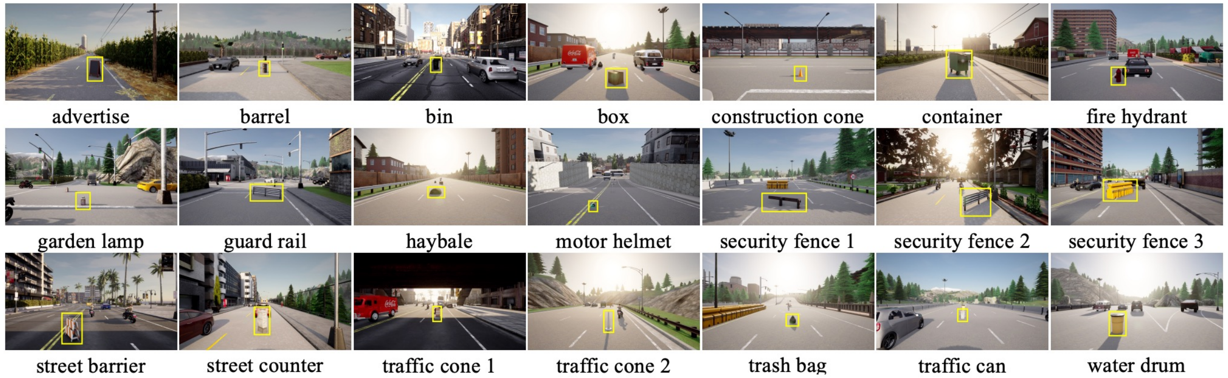


This competition is dedicated to enhancing multimodal perception and comprehension capabilities of MLLMs for autonomous driving, focusing on the **General Perception**, **Regional Perception**, and **Driving Suggestions**.

[A] Li Yanze *et al.*, Automated evaluation of large vision-language models on self-driving corner cases, in *arXiv preprint: 2404.10595*, 2024.

/02 Challenges in Corner Cases

Varied Scenes

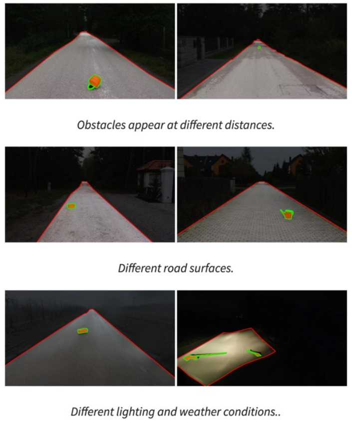


Diverse Objects

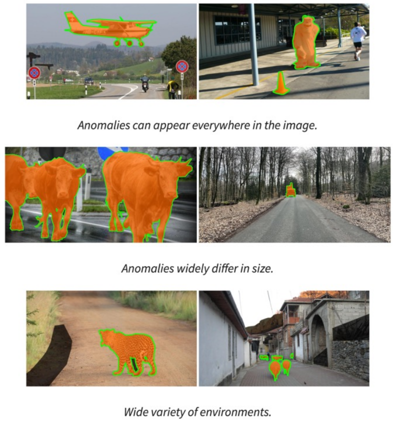
Novel class

Novel instance

Varied Scenes



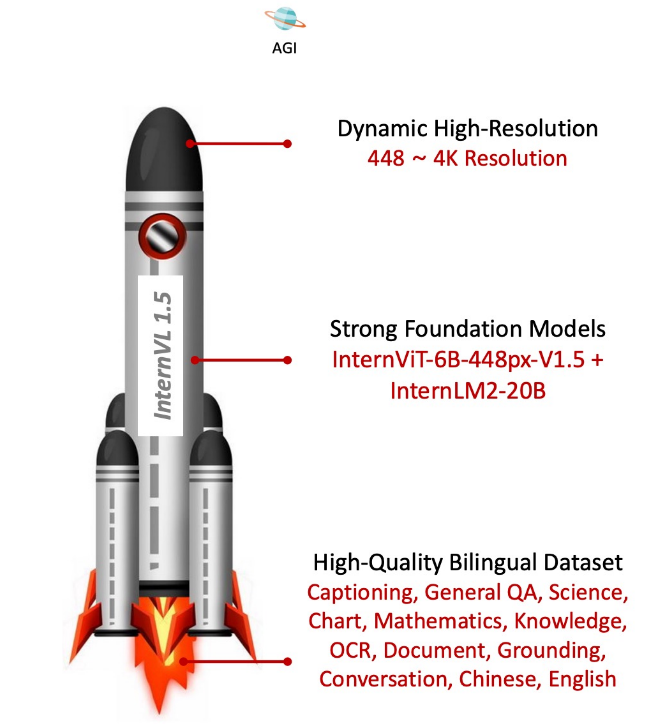
Small Objects

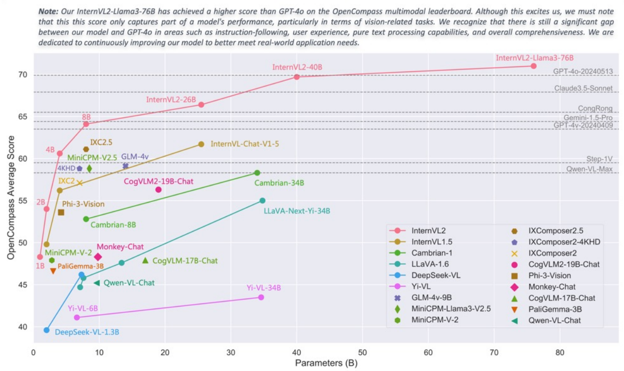


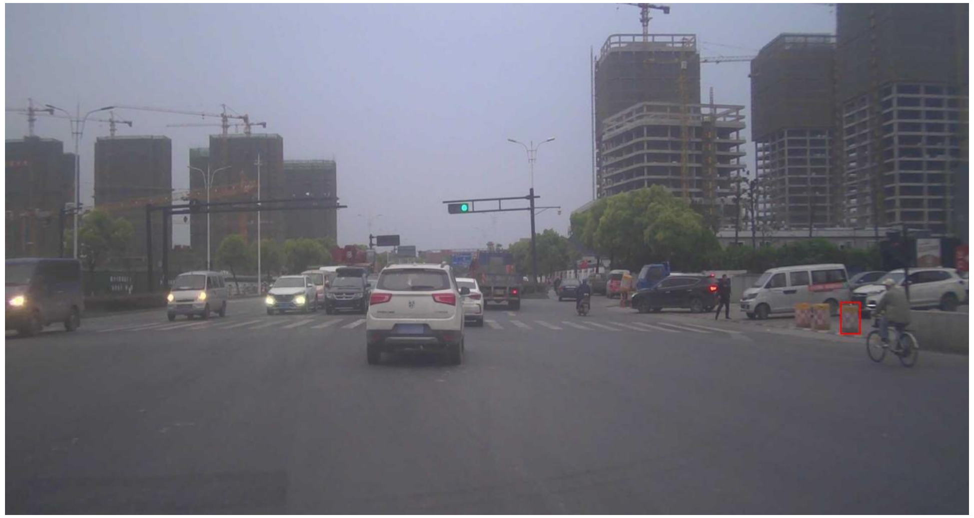
Contextual Anomaly

[A] Li Kaican *et al.*, Coda: A real-world road corner case dataset for object detection in autonomous driving, in *ECCV*, 2022.

/02 Regional Perception via Vision-Language Models







Query: Please describe the object inside the red rectangle in the image and explain why it affect ego car driving.

Answer: A barrier with orange and white stripes, indicating a temporary obstruction or hazard. The ego car must remain vigilant for any changes in the driving lane or potential rerouting due to this obstruction. It could indicate construction, a road closure, or an accident ahead, and drivers should be prepared to slow down or follow detour signs.

The object inside the red rectangle in the image is a cyclist. This individual appears to be riding on or near an intersection, which could impact how vehicles navigate through that area.

Description of Object:

1. **Location:** The cyclist seems positioned close to where cars might turn right.
2. **Direction and Movement:** It's unclear whether they are actively moving into traffic lanes or waiting at the side but their presence poses potential risks due to unpredictable movements typical among cyclists navigating city streets.

How it Affects Ego Car Driving:

1. Potential Risk for Collision

- If the cyclist decides to move suddenly across the path of your car while turning right, there's a risk of collision unless you remain alert and prepared with defensive driving maneuvers such as slowing down and maintaining visibility checks around blind spots common during turns.

2. Safety Concerns During Turning Maneuver

- When making a turn (left/right), drivers must always check surroundings carefully before executing any maneuver because bikes can often appear unexpectedly from behind parked cars or crosswalk areas adjacent roads.

3. Impact on Speed Management

- Presence of bicycles necessitates more cautious speed control particularly when approaching intersections since sudden braking by cyclists isn't uncommon if startled or reacting quickly towards other road hazards like pedestrians jumping onto paths unannouncedly.

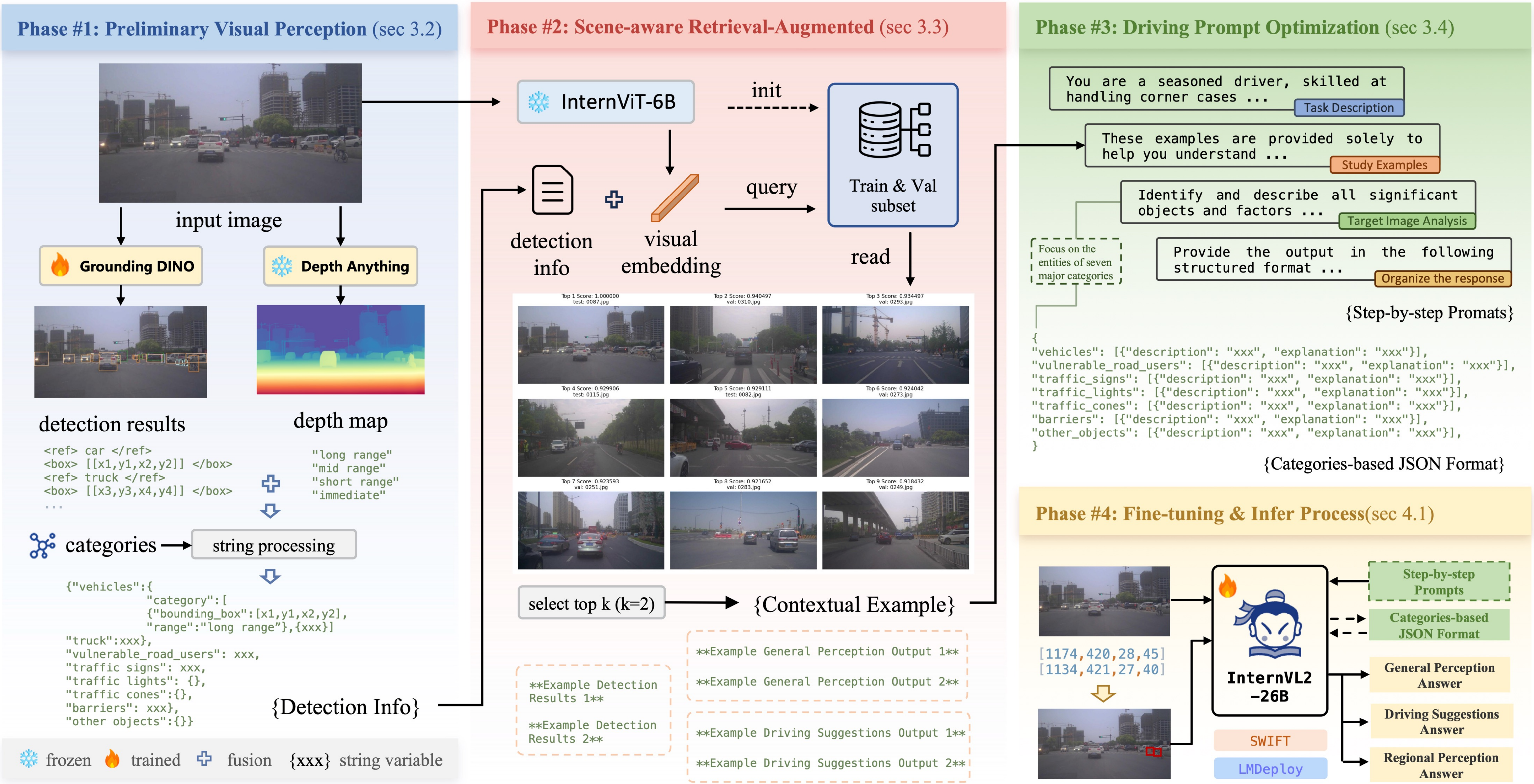
4. Visibility Hazard

```
- Cyclists pose challenges especially under poor weather conditions given ma
```

In conclusion, this cyclist significantly affects ego vehicle operation primarily through increasing the unpredictability factor influencing safe navigation within urban settings characterized prominently via higher likelihood encounters involving vulnerable road users requiring constant situational awareness & adaptive behavior adjustments whilst ensuring adherence to all safety protocols including regular scan patterns over surrounding environment even post initial observations confirming apparent emptiness visually discerned initially.

[A] **Chen Zhe** *et al.*, Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic task, in *CVPR*, 2024.

/03 **Overview** of the proposed NexusAD framework



/04 Visualization

Our method focuses on scene similarity, not just image similarity, allowing retrieval of the most relevant reference samples for the current driving scenario.

/04 Suggestions provided by ChatGPT

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}]}.
```

Based on the analysis of these topics, I have the following suggestions to further optimize your VLM prompts:

1.Enhance specificity: Ensure that the prompt asks the model to provide precise and detailed operational suggestions, reducing vague or overly broad language. For example, in sections like "Speed Adjustments" and "Lane Positioning and Maneuvers," ask the model not only to indicate when speed adjustments or lane changes are necessary but also to explain how to perform these actions in detail.

2.Emphasize critical traffic elements: Make sure the model consistently focuses on key traffic elements, such as traffic lights, red lights, and important road signs. Consider adding a step in the prompt that directs the model to pay special attention to these critical elements and provide specific response strategies.

3.Avoid overgeneralization: Encourage the model to concentrate on the unique aspects of the current situation, avoiding overly generalized advice. You could include a section in the prompt reminding the model that each scenario is unique and requires a tailored response based on the specific elements detected.

4.Incorporate a feedback mechanism: If possible, implement a mechanism that allows the model to recheck and adjust its recommendations after generating an initial suggestion. This can help reduce low-quality responses caused by omissions or misunderstandings.

5.Minimize the influence of examples: If the examples are causing hallucinations, consider providing only strategic advice from the examples without including specific detection results. This approach can guide the model to focus on structure and depth rather than the specific details of the example.

These suggestions can be integrated into your current prompt version to better align with the characteristics of VLMs, improving the overall scoring of the final predictions.

/05 Results

Leaderboard (Best of All Rounds)

We have removed team submissions violating the competition rules.

Team Name	Date	Final Score	General Perception	Region Perception	Driving Suggestion
llmforad	08-15	72.12	58.70	83.41	74.26
OpenDriver	08-09	69.72	54.41	83.00	71.76
NexusAD	08-15	68.97	57.58	84.31	65.02
123	08-02	68.79	52.98	83.07	70.32
LLMAnything	08-15	68.43	56.24	82.66	66.38
NTHUCVLab	08-15	67.85	55.16	82.88	65.50
Swift Unity Expedition	08-15	67.68	57.38	84.37	61.30
GY	07-26	66.79	46.32	84.50	69.54

Method	FS	GP	RP	DS
GPT-4V	59.02	57.50	56.26	63.30
CODA-VLM	63.62	55.04	77.68	58.14
InternVL-2.0-26B	52.11	43.39	64.91	48.04
NexusAD (Ours)	**68.97**	**57.58**	**84.31**	**65.02**

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

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

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

Thanks for your listening and attention！

GitHub Link

References:

[1] Li, Yanze, et al. "Automated evaluation of large vision-language models on self-driving corner cases." arXiv preprint arXiv:2404.10595 (2024).

[2] Chen, Zhe, et al. "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[3] Li, Kaican, et al. "Coda: A real-world road corner case dataset for object detection in autonomous driving." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

[4] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2024.

[5] Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.