Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

Challenge 1 - Team 11 Saturn

電信所碩一 呂則諺 (R13942070)、電信所碩二 沈鶩毅 (R12942160)、電子所碩一 黃昱翔 (R13943016)、電信所碩二 吳育澤 (R12942080)

Introduction

This challenge addresses the robustness of autonomous driving systems in handling complex and unexpected "corner cases." It focuses on leveraging multimodal data, such as images and text prompts, to enhance perception and understanding.

The three question types are:

- General perception,
- Regional perception,
- Driving suggestions.

Our approach involves using task-specific model weights to better distinguish the objectives of the three tasks, ensuring each task is optimized independently. We enhance prompts to improve input quality and integrate depth information into images to provide richer contextual data.

Methodology

Task-specific Model Weights

In order for our model to better distinguish objectives of the three tasks, we prepared three independent sets of model weights. For each set, we only finetune the model on one specific type (general/regional/suggestion) of data. In our inference pipeline, we first recognize the data question type then process the data with the corresponding model weights.

Enhanced Prompts

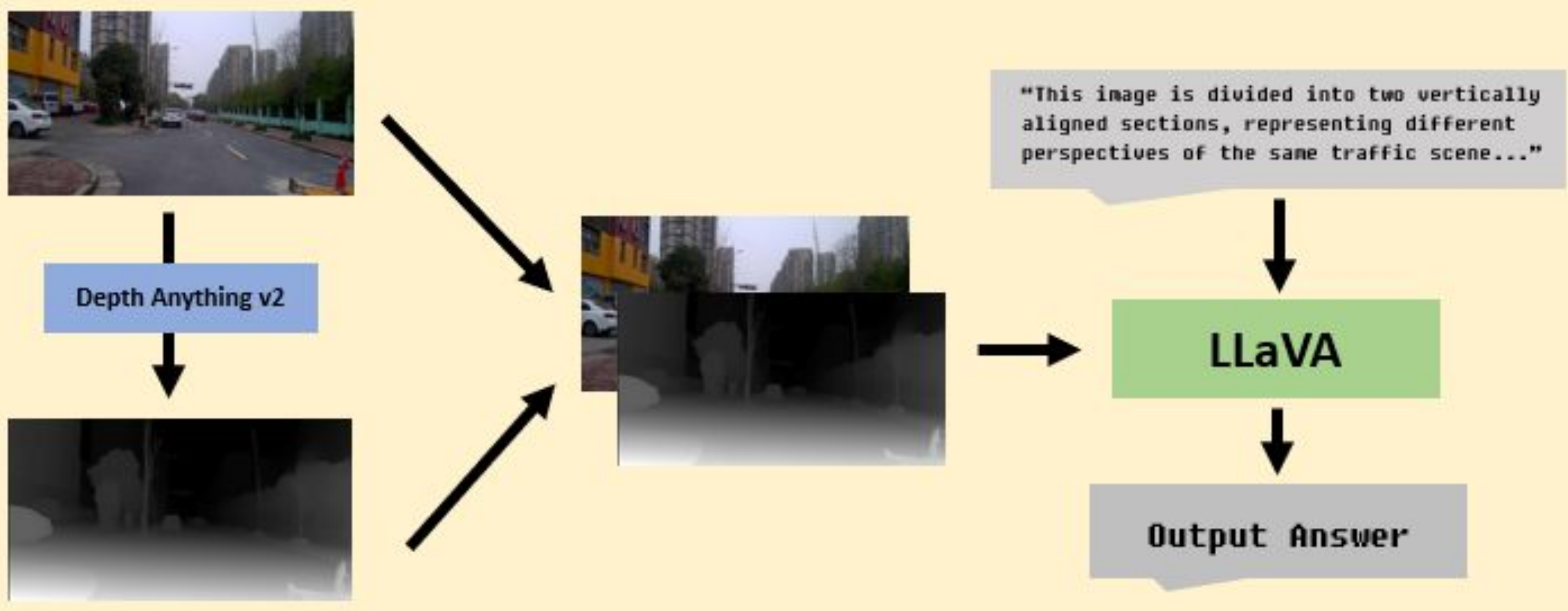
We refer to the methods in [1] to provide more information in LLaVA's input prompt, such as target detection results from DINO [2] and range information from Depth Anything V2 [3]. One detection result is as follows, where object classes, bounding boxes in the image and their relative distances are organized in a Python dict().

```
“traffic signs”: [
    {
        “bounding_box”: [
            482,
            413,
            542,
            437,
        ],
        “range”: “long range”
    }
]
```

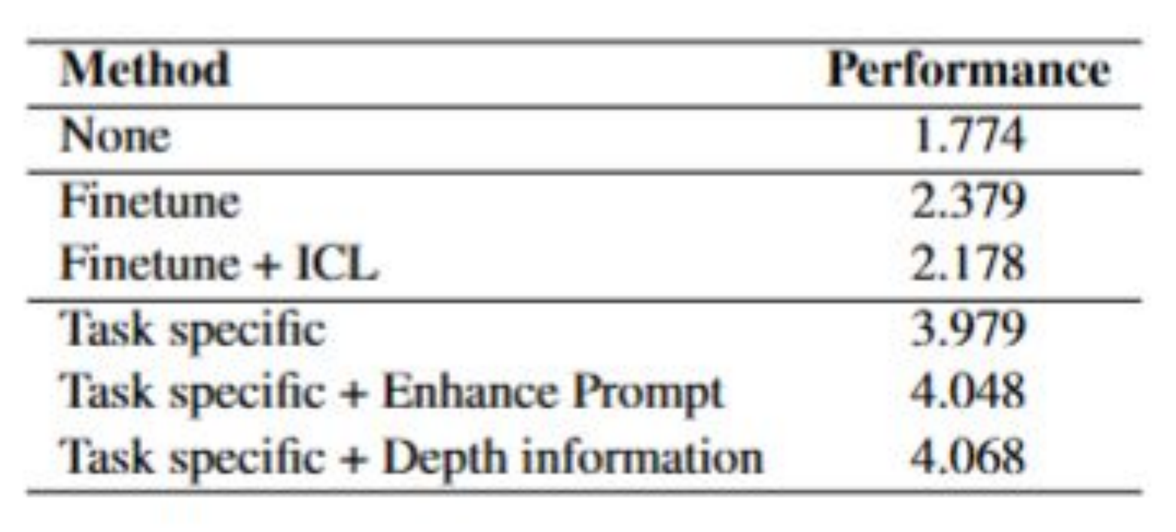
Our prompt is not as detailed as [1] since we observed that long prompts do not necessarily output good predictions and also pose a challenge for our training (GPU memory capacity). Therefore, we started from the original prompt and added the obtained information step by step in a trial-and-error measure.

Embed Depth Information to Images

To help the model better understand the relationships between objects in a scene, we incorporate depth information as a crucial component. By embedding depth data, the model gains a more comprehensive understanding of the spatial structure within traffic scenarios. To achieve this, we leverage the method proposed in [3], utilizing zero-shot inference to extract depth information directly from traffic scene images. This approach enhances the model's ability to perceive and analyze the scene's structural composition effectively.


Depth Anything v2
"This image is divided into two vertically aligned sections, representing different perspectives of the same traffic scene..."
LLaVA
Output Answer

Results & Analysis



Method	Performance
None	1.774
Finetune	2.379
Finetune + ICL	2.178
Task specific	3.979
Task specific + Enhance Prompt	4.048
Task specific + Depth information	4.068

The charts above display the quantitative results. The experimental findings indicate that using task-specific model weights significantly enhances performance. Integrating additional information into the input image data results in a slight performance improvement.

Conclusion

In this project, we found that using task-specific model weights can achieve better results. We hypothesize that this is because such a training framework helps the model focus on specific tasks, allowing it to perform better in those particular tasks. Additionally, we attempted to incorporate more information into the model to help it better understand scene context. However, our experiments reveal that an excess of information leads to only a slight performance improvement. We hypothesize that the bottleneck may lie in the model's capability.

Reference

[1] Mo, M., Wang, J., Wang, L., Chen, H., Gu, C., Leng, J., & Gao, X. NexusAD: Exploring the Nexus for Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving. In ECCV 2024 Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving.

[2] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., ... & Zhang, L. (2025). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision (pp. 38-55). Springer, Cham.

[3] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth Anything V2. arXiv preprint arXiv:2406.09414.