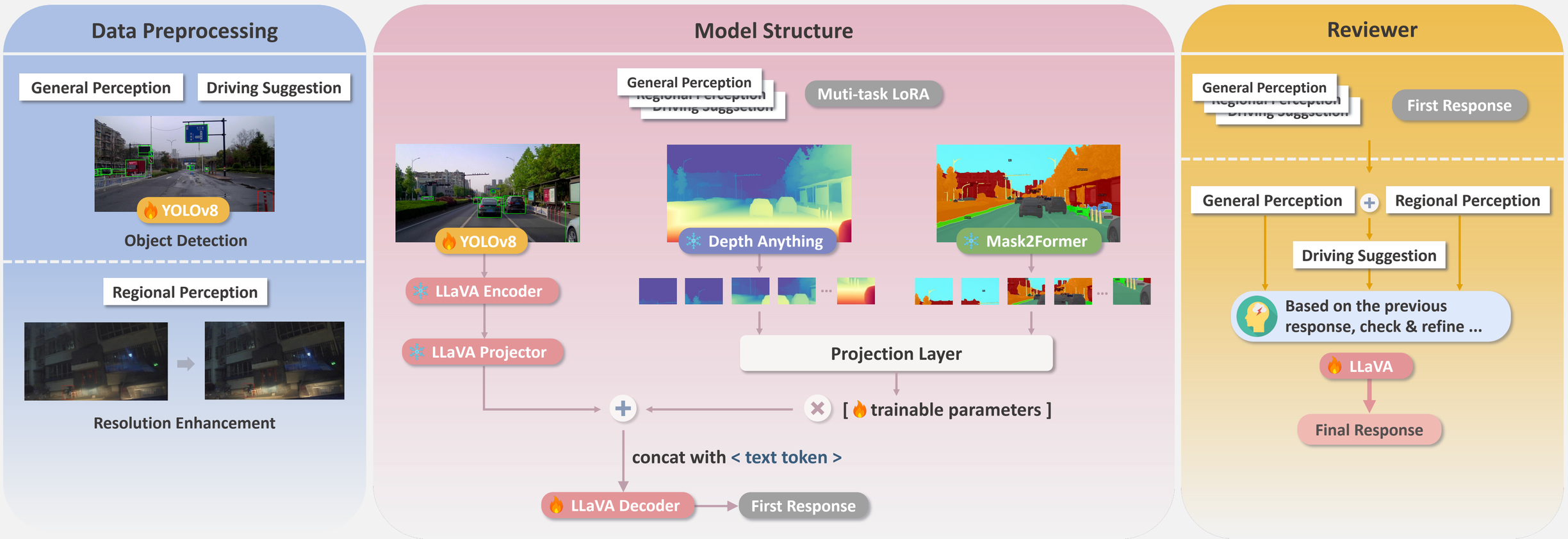
# Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

電機碩二 R12921057 林佳儀, 電機碩二 R12921125 楊沛蓉, 電機碩一 R13921093 李彥璋, 網媒碩二 R12944030 簡志宇

Panda603

## Method




## Prompting Mechanism Design


▼ Integrate identified tags in image , split into sub prompt and implement multi-run prompting in order

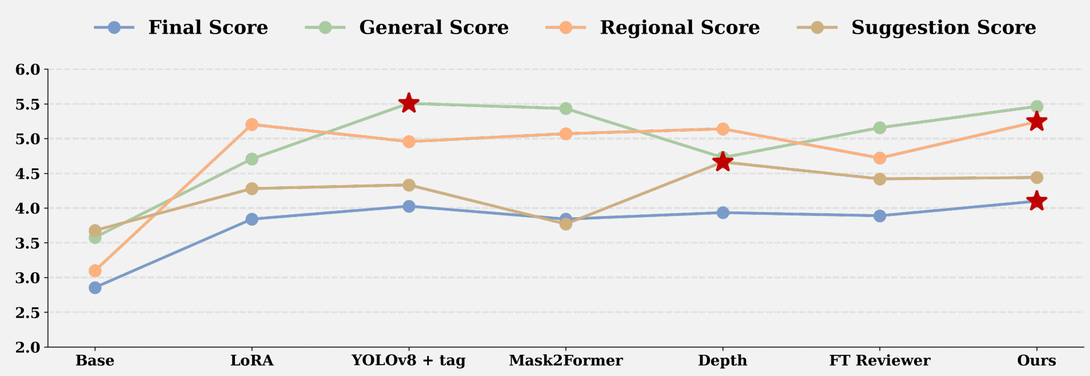
This is an image of traffic captured from the perspective of the ego car. Here are the key elements in the image { tag[sample_id] } ...

traffic cone, vehicle ,traffic sign, suv

1. Describe the objects in this traffic scene, including their relationship ...
2. Explain how these objects within the 7 categories contribute to the traffic scene, especially { tag[sample_id] } ...
3. Discuss their impact on traffic patterns, and give suggestion ...

## Experiment

| Id | Setting | General | Regional | Suggestion | LLM judge | Final Score |
|---|---|---|---|---|---|---|
| 1 | LLaVA (Base) | 3.577 | 3.097 | 3.677 | 3.450 | 2.856 |
| 2 | Finetune 1 (Multi-Task LoRA) | 4.707 | 5.203 | 4.280 | 4.730 | 3.841 |
| 3 | 2 + YOLOv8 + tag | **5.507** | 4.957 | 4.333 | 4.932 | 4.027 |
| 4 | 2 + Mask2Former | 5.433 | 5.070 | 3.770 | 4.758 | 3.841 |
| 5 | 2 + Depth | 4.727 | 5.140 | **4.663** | 4.483 | 3.934 |
| 6 | 3 w/ Reviewer (Finetuned 1) | 5.157 | 4.720 | 4.420 | 4.766 | 3.889 |
| 7 | Ours (2 w/ all features) | 5.463 | **5.243** | 4.440 | **5.049** | **4.099** |




Our two-stage approach significantly improves LLaVa's performance, with the combined method achieving a 44% increase in Gemini score (from 2.856 to 4.099) through enhanced perception and reviewing process.

## Idea Key Points

Our proposed framework integrates several advanced components, These innovations collectively contribute to our successful results in our metric. Key highlights include:

- **YOLO-based Image Detection**:
  - Finetuned YOLO to enhance small object detection and recognition, specially on General Task.
- **Resolution Enhancement**:
  - Improved image quality to support accurate visual processing, specially on Regional Task.
- **Multi-Task Perception Enhancement**:
  - *Depth Anything* for depth estimation and spatial understanding.
  - *Mask2Former* for precise object and region segmentation.
  - *Recognize Anything* for robust object and feature identification.
- **Lora-based Adaption**: Efficient fine-tuning of LLaVa-1.5-7b for seamless integration of multi-modal information.

## Ablation Study

- **LoRA Rank Effects (8-64)**: We observed that higher ranks showed decreased performance under constant training time. Extended training periods led to overfitting issues, significantly reducing the model's ability to generalize to new scenarios.
- **Handling Missed Scene Factors**: The model initially missed critical scene elements like traffic cones. Integration of *Recognize Anything* for automated tagging of important scene factors during inference significantly improved contextual understanding.
- **Reviewer Module Findings**: The reviewer let the model revisit images to refine its initial response. Interestingly, we found that using a model with lower testing score to train the reviewer resulted in better refinement ability compared to a higher-scoring one. We hypothesize that this is due to lower-scoring models developing stronger revising capabilities by learning to transform low-quality responses into high-quality ones.