# 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

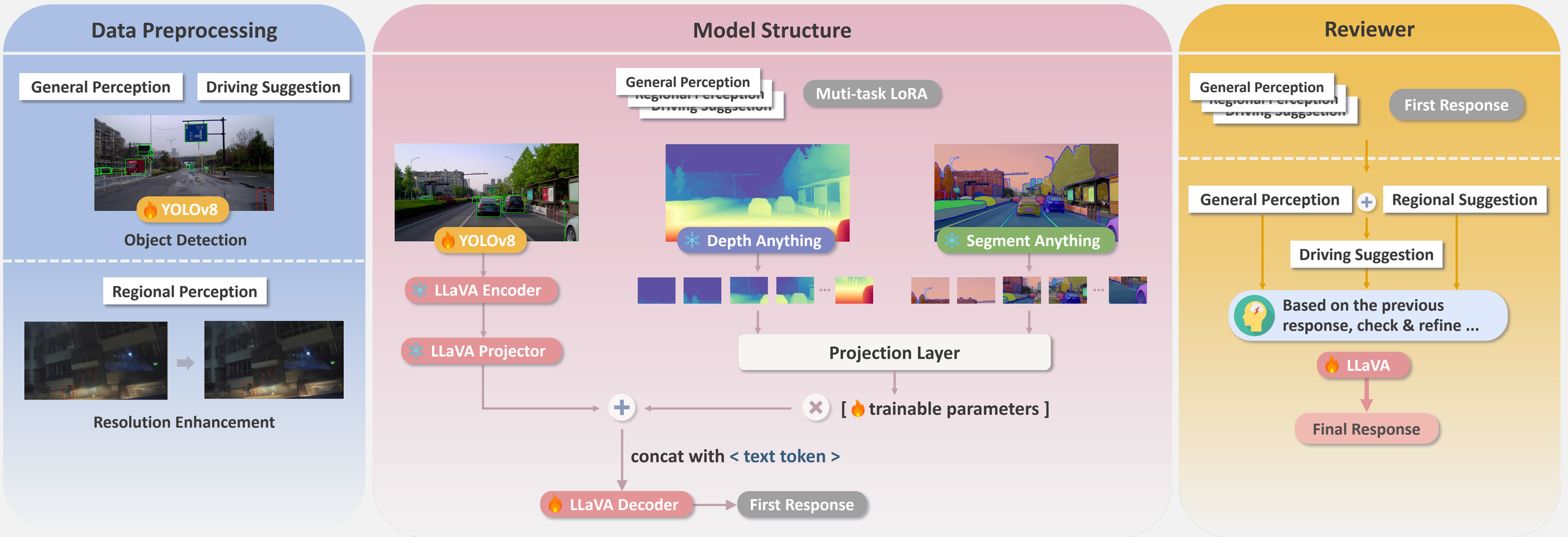
## Abstract

Real-world traffic scenarios pose significant challenges for Visual Language Models (VLMs) like LLaVa, particularly in detecting small objects and handling complex traffic interactions. To address these limitations, we enhance LLaVa-1.5-7b through an optimized three-stage preprocessing pipeline. First, we integrate YOLO-based detection with resolution enhancement techniques to improve image quality and object recognition. Second, we enhance the model's perception through a multi-task learning framework that integrates depth and segmentation information using LoRA-based adaptation. Finally, we implement a specialized validation module to refine and verify model responses. Our enhanced system demonstrates superior performance in global scene understanding, object detection accuracy, and contextual reasoning for autonomous driving scenarios.

## 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.

## Method




## Prompting Mechanism Design


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

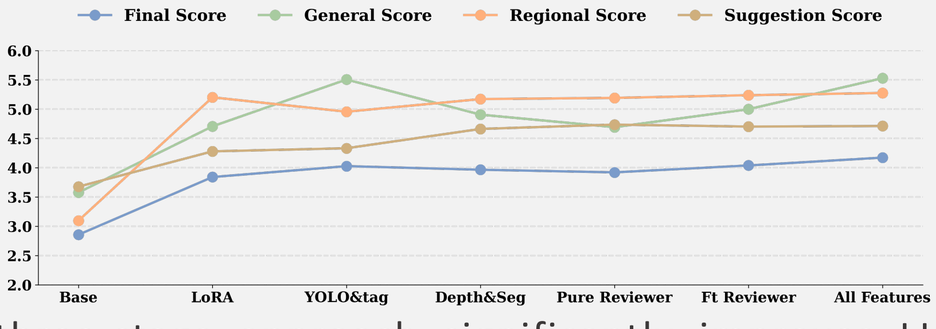
This is an image of traffic captured from the perspective of the ego car. Here are the key elements in the image { tag[sample_id] } ...

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

| Index | Setting | Belu 1 | Belu 2 | Belu 3 | Belu 4 | General | Regional | Suggestion | LLM judge | Final Score |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LLaVA (Base) | 1.735 | 0.883 | 0.481 | 0.278 | 3.577 | 3.097 | 3.677 | 3.450 | 2.856 |
| 2 | Finetune 1 (LoRA) | 0.820 | 0.472 | 0.283 | 0.172 | 4.707 | 5.203 | 4.280 | 4.730 | 3.841 |
| 3 | 2 + YOLOv8 + tag | 1.262 | 0.699 | 0.405 | 0.242 | 5.507 | 4.957 | 4.333 | 4.932 | 4.027 |
| 4 | 2 + Depth & Seg | 0.476 | 0.280 | 0.170 | 0.106 | 4.907 | 5.173 | 4.663 | 4.914 | 3.966 |
| 5 | 2 w/ Reviewer (Pure 1) | 0.299 | 0.176 | 0.108 | 0.068 | 4.693 | 5.193 | **4.737** | 4.874 | 3.921 |
| 6 | 3 w/ Reviewer (Finetuned 1) | 0.565 | 0.333 | 0.201 | 0.124 | 5.000 | 5.239 | 4.703 | 4.999 | 4.039 |
| 7 | 2 w/ all features | 0.485 | 0.287 | 0.174 | 0.107 | **5.530** | **5.277** | 4.713 | **5.173** | **4.173** |




Our three-stage approach significantly improves LLaVa's performance, with the combined method achieving a 46% increase in Gemini score (from 2.856 to 4.173) through enhanced Perception and reviewing process.

## Conclusion

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.
- **Resolution Enhancement**: Improved image quality to support accurate visual processing.
- **Multi-Task Perception Enhancement**:
  - *Depth Anything* for depth estimation and spatial understanding.
  - *Segment Anything* 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.