

# 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

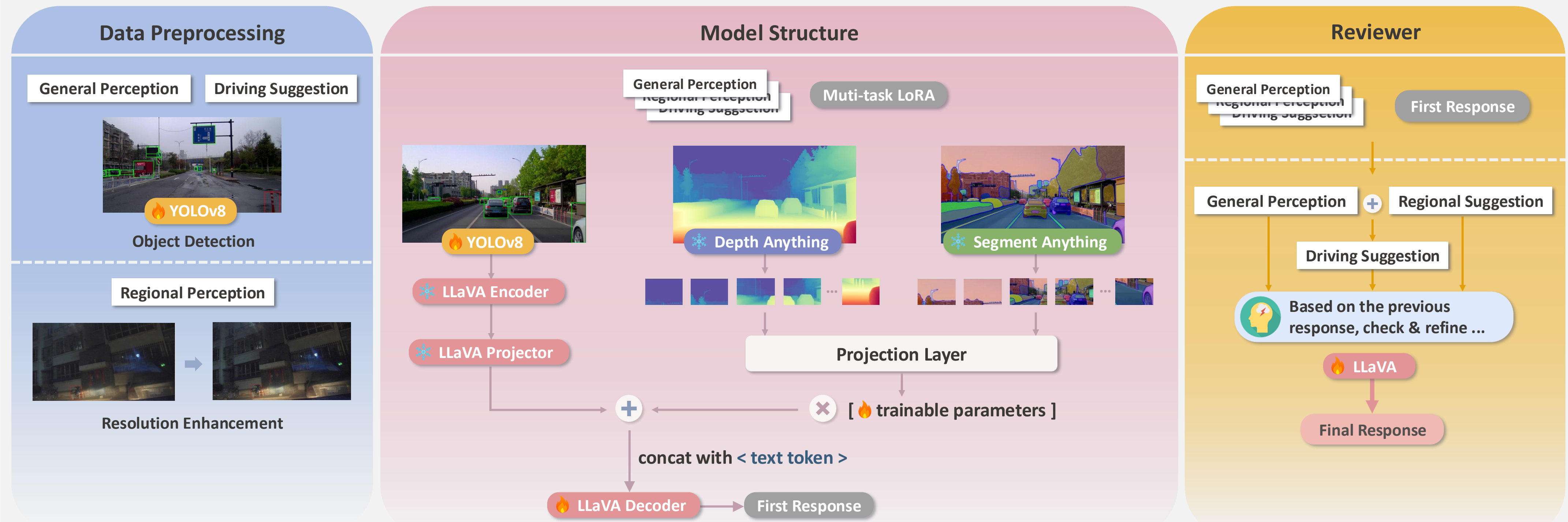
## Abstract

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.

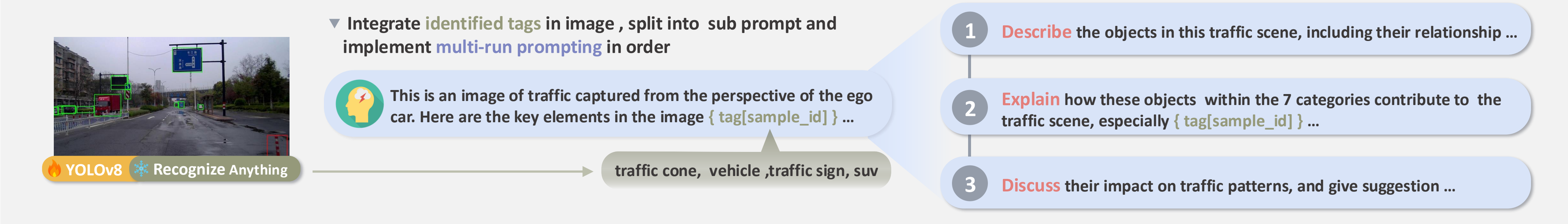
## Ablation Study

- 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



## Experiment

Setting	Gemini Score
LLaVA	2.856
Finetuned LLaVA (Multi-Task LoRA)	3.841
→ w/ YOLOv8	
→ w/ Depth & Segmentaion map	
→ w/ Reviewer (Pure LLaVA)	3.921
→ w/ Reviewer (Finetuned LLaVA)	
Combine All	

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.