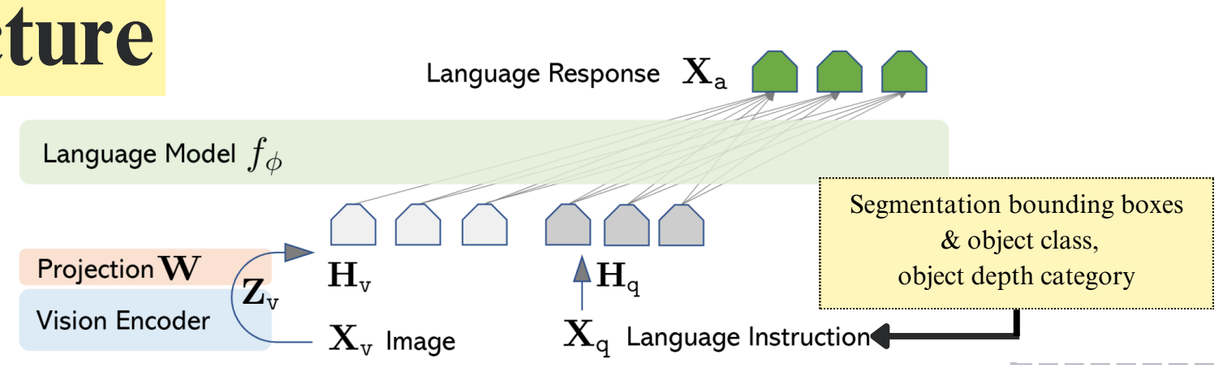
Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

Team Name: FRANK
師大資工系 大二 41247038s 陳哲堯
台大電機系 大四 b10901151 林祐群
台大資工所 碩二 r12922187 雷舜清
台大資工所 碩二 r12922166 張璟榮

Abstract

In this ECCV challenge, we aim to generate detailed descriptions of scenes or specific objects (e.g., motorcycles, fences, cars) from a car's perspective, as well as provide actionable driving recommendations based on the observed viewpoint. Through extensive experimentation, we discovered that the most effective approach involves integrating segmentation, depth map generation, and advanced prompt engineering. These processed outputs are then utilized as inputs for LLaVA, enabling highly contextual and accurate scene understanding.

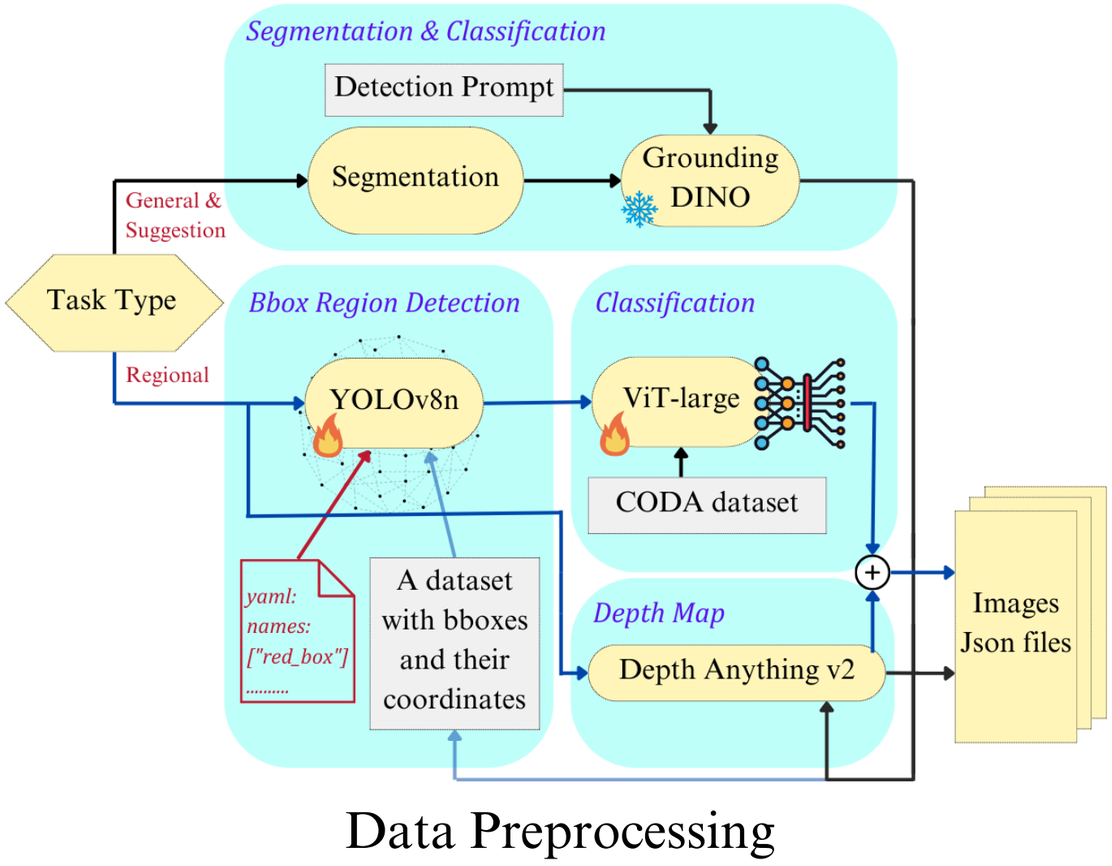
Structure



Prompt Engineering

You are an experienced car driver, enable to point out all details need to be focused while driving. An image from the driver's seat of a ego car is given, corresponded to a question to a question. You need to answer the question with your analysis from image

+ Original Problem



Data Preprocessing

Results

General Task

- a yellow construction vehicle parked on the right side of the road, partially obstructing the lane…
- red and white striped barriers are placed in front of the construction vehicle, marking a work zone or lane closure. These barriers serve as a physical obstruction...

Regional Task

A four-wheeled vehicle is parked on the side of the road. The ego car should maintain a safe distance from the parked vehicle to avoid any potential hazards, such as opening doors or the vehicle pulling out into traffic.

Suggestion Task

Maintain a safe following distance behind the motorcyclist, be prepared to slow down or stop for pedestrians at the crosswalk, and do not attempt to change lanes to the left due to the traffic cones. Monitor the red car on the left for any lane changes or adjustments in speed.

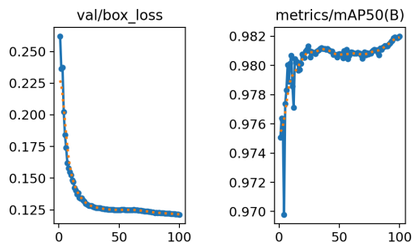
Experiments

strategy	Segmentation	Depth map	RAG	Training Epochs	BLEU_3	General	Regional	Suggestion	LLM_Judge	Total Score
Finetune LORA only	❌	❌	❌	6	0.337	4.753	4.660	4.917	4.777	3.889
Finetune LORA only	❌	❌	✅	6	0.645	3.043	3.907	4.390	3.780	3.153
Prompt engineering	❌	❌	❌	6	0.294	5.763	4.893	4.487	5.048	4.097
Add Seg. Prompt	✅(Prompt)	❌	❌	3	0.356	5.533	5.123	4.403	5.020	4.087
Add Seg. Prompt, Dep	✅(Prompt)	✅	❌	2	0.414	4.447	4.820	4.693	4.653	3.806

strategy	Segmentation	Depth map	RAG	Training Epochs	BLEU_3	General	Regional	Suggestion	LLM_Judge	Total Score
Add Seg. Prompt, Dep, Prompt engineering	✅(Prompt)	✅	❌	3						

YOLOv8n training information:

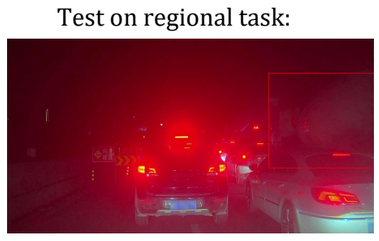
Validation on curated dataset:



Vit accuracy (43 classes):

Training Accuracy: 0.9605
Validation Accuracy: 0.9657
Full CODA 2022 dataset 1: 0.9580
Full CODA 2022 dataset 2: 0.9554

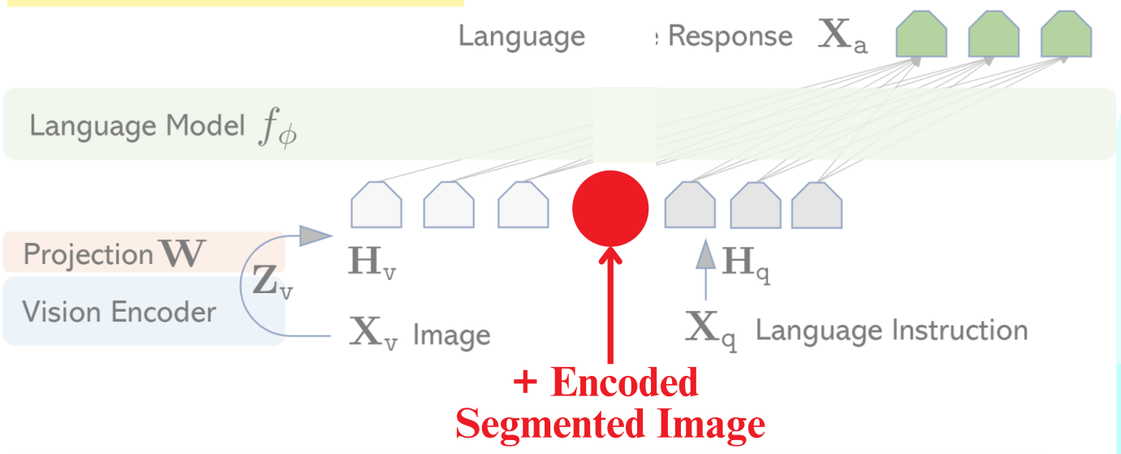
Test on regional task:



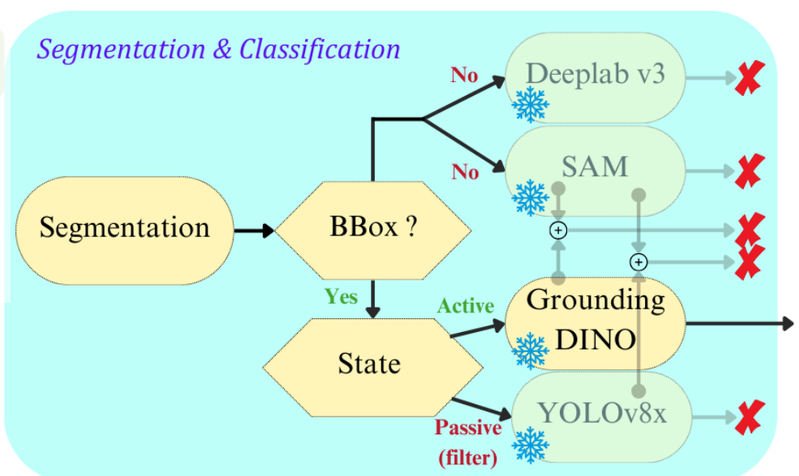
Ground Truth

Prediction

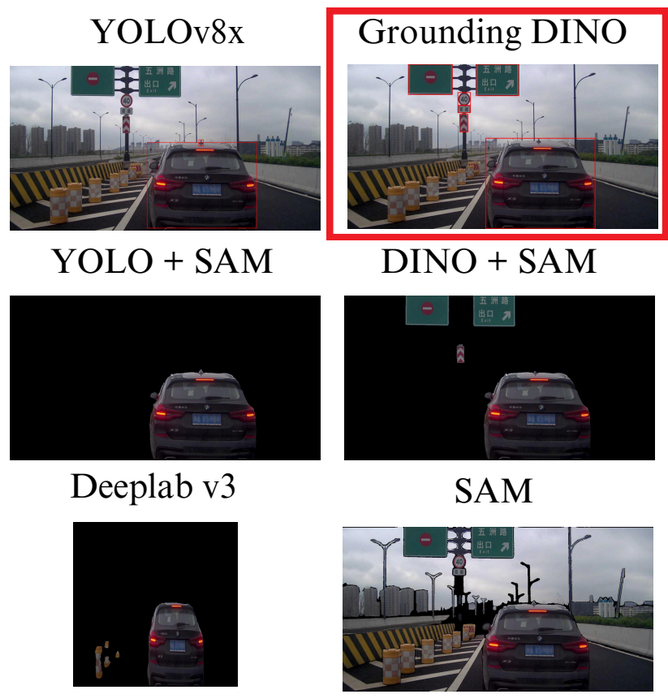
Comparison



strategy	Training Epochs	BLEU_3	General	Regional	Suggestion	LLM Judge	Total Score
Add Seg. Token	3	0.438	4.920	4.873	4.577	4.790	3.92



Model Comparison



References

1. *Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2023). Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. ArXiv. https://arxiv.org/abs/2303.05499*
2. *Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth Anything V2. ArXiv. https://arxiv.org/abs/2406.09414*