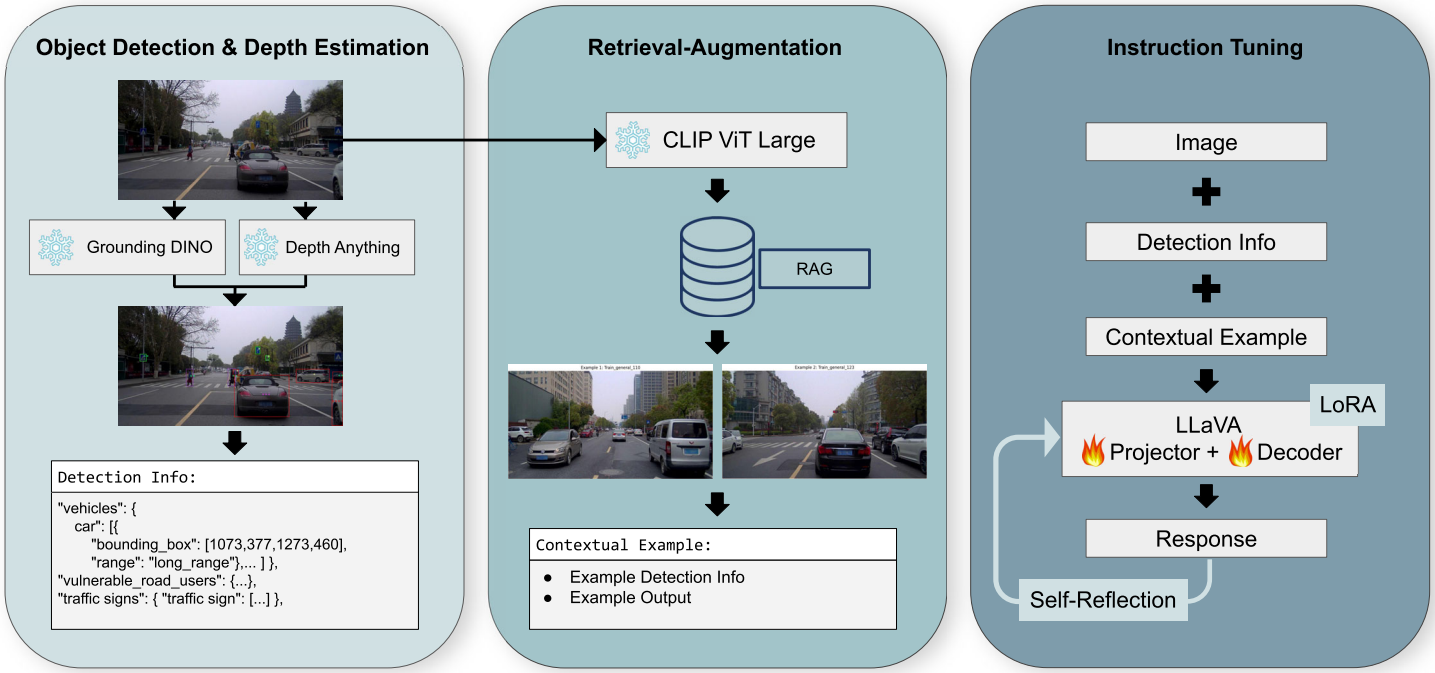
Deep Learning for Computer Vision Final Project Challenge 1
Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

Team 3 csier1213s

R12944033 網媒碩二 吳昇陽, P12922004 資工碩二 洪嘉駿, R13922050 資工碩一 楊詠絜, R13922051 資工碩一 黃唯秩

OVERVIEW



METHOD

Object Detection & Depth Estimation

- We first utilize object detection and depth estimation models to extract spatial information from images.
- We select GroundingDINO as the object detector due to its excellent zero-shot capability.
- We utilize Depth-Anything-v2 to extract pixel-wise depth information of the image. For each detected object, we take the median of the depth of each pixel in its bounding box as the depth of that object.

Retrieval Augmented Generation

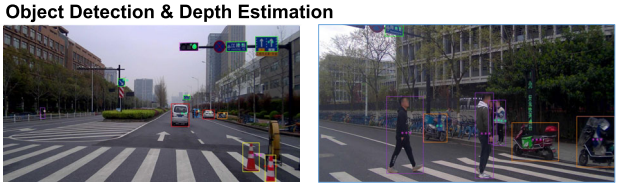
- We use the CLIP ViT Large vision encoder to extract features (CLS token embedding) and store them in a vector database.
- We retrieve the top-k (k=1,2) images based on cosine similarity and include their corresponding GT outputs as examples in the input prompt for the target image.

Self-Reflection

- We make the model refine its responses based on its previous outputs.
- We create the input for self-reflection by inferencing our previous trained model and use the ground truth data as intended output for refinement.

VISUALIZATION

Object Detection & Depth Estimation



∗ ∗ ∗ ∗: immediate ∗ ∗ ∗: short range ∗ ∗: mid range ∗: long range

RAG Results

EXPERIMENTS & ABLATION STUDY

Method	Overall	LLM	General	Regional	Suggestion	Bleu-1	Bleu-2	Bleu-3	Bleu-4
Ablation Study									
Baseline	3.404	4.158	4.87	4.19	3.413	1.091	0.639	0.39	0.242
+Detection information	2.894	3.513	3.71	3.573	3.257	1.167	0.684	0.418	0.26
+RAG	3.736	4.573	4.673	4.87	4.177	1.103	0.639	0.387	0.238
+Detection information +RAG	4.08	4.997	5.423	4.933	4.633	1.166	0.678	0.414	0.257
+Detection information +RAG +Self reflection	**4.135**	5.077	5.807	4.817	4.607	1.059	0.611	0.367	0.224
Self Reflection									
+Reflection once	3.108	3.803	5.557	2.697	3.157	0.912	0.533	0.328	0.207
+Reflection twice	3.667	4.502	5.83	4.62	3.057	0.911	0.532	0.328	0.208
+Reflection three times	3.355	4.109	5.113	4.503	2.71	0.938	0.548	0.339	0.215
Other Attempts									
Images with boxes	3.98	4.87	4.72	4.95	**4.94**	1.206	0.693	0.42	0.26
Providing general results for suggestion	3.903	4.779	4.907	4.927	4.503	1.132	0.655	0.398	0.247
Integration	4.059	4.972	5.033	4.98	4.903	1.167	0.674	0.408	0.252

DISCUSSION & CONCLUSION

Images with bounding boxes

We experimented with inputting images that include bounding boxes colored according to their category and found that this approach significantly improved the driving suggestion results.

Providing general results for driving suggestion

Since general results typically describe scenarios in great detail, we attempted to replace the detection information with general perception results to enable the model to provide driving suggestions.

Integrating different methods

We also explored integrating different methods, such as using original images for general and regional perception, while providing images with bounding boxes for driving suggestions, in order to achieve the best results for each task.

Conclusion

In our method, we used object detection result and depth information to help the model understand the situations better, and utilized retrieval-augmented generation to provide an example to guide the model to generate outputs with desired format. Also, self-reflection gives the model a second chance to provide more detailed answers. By experiment, we proved these strategies help improve the model's performance.