Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving (ECCV 2024 Challenge)

楊竣凱 陳楚融 陳英睿 梁璿安

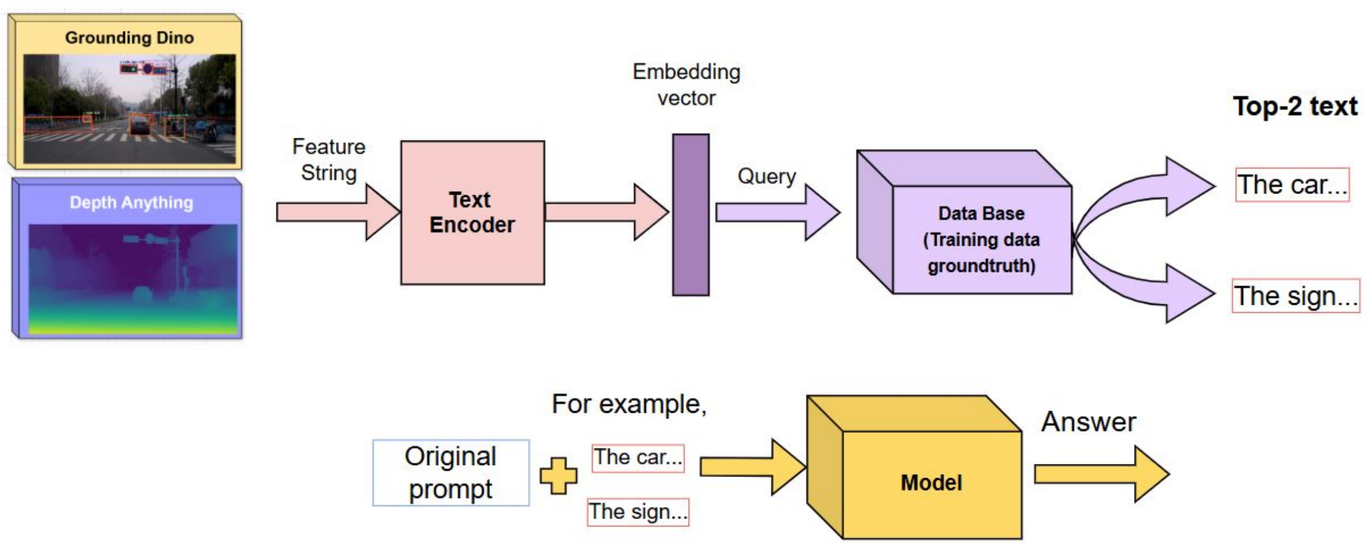
Group 12: To be Frank with you

INTRODUCTIOIN

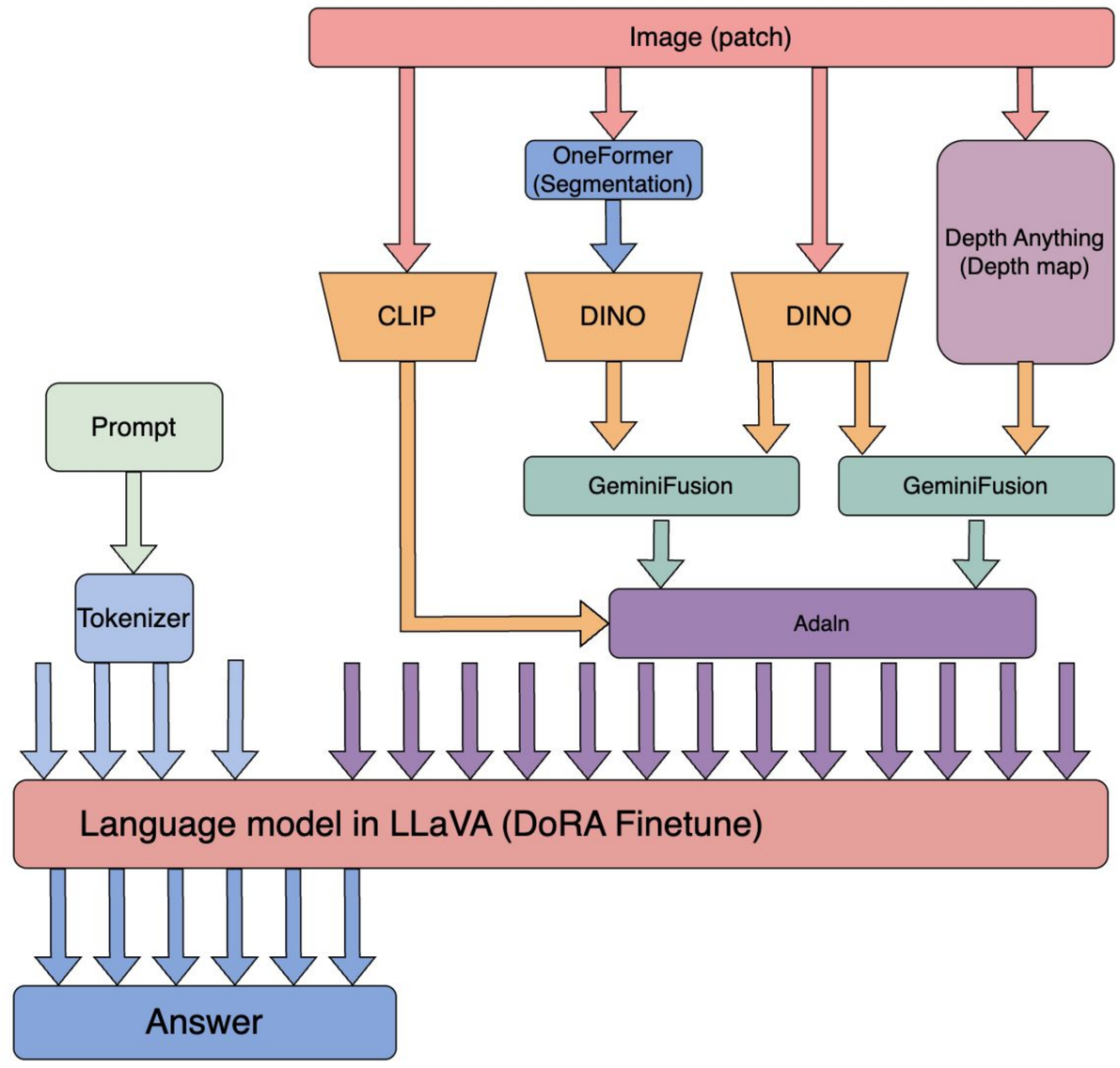
In the field of autonomous driving, state-of-the-art techniques suffer from corner cases. Therefore, RAG is implemented to enhance the scene describing ability of our work. The RAG searching relies only on object detection, in order to ignore weather and background information. Furthermore, we try the method which utilizes both segmentation and depth information to help language model better capture objects in real traffic scenario.

MODEL STRUCTURE

RAG Flow:

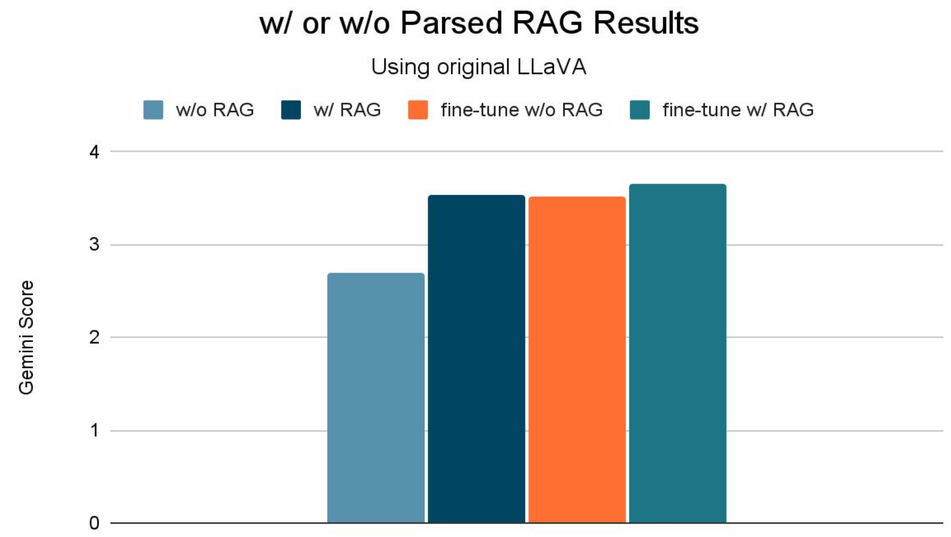


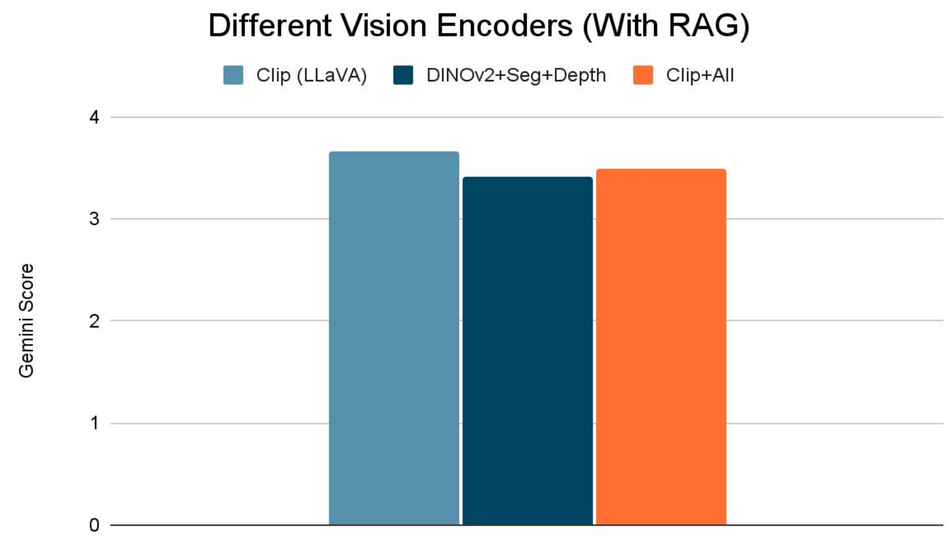
Model Structure:

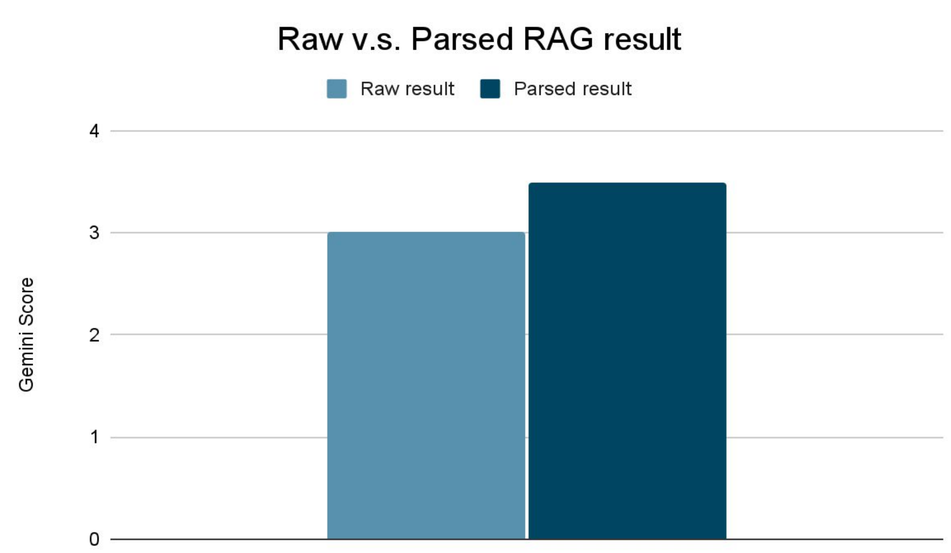


ABLATION STUDY

We conduct each experiment with Gemini Evaluation on validation dataset (500 samples)







CONCLUSION

Our experiment shows that the proposed approach (3encoder RAG) can indeed improve the performance of the model . RAG can capture data with similar objects distribution and use it as an example to guide language model generation. However, the segmentation and depth features fused into our image feature, doesn't better capture the spatial correlation than original LLaVA according to our experiment. In future work, we plan to refine our RAG methodology and modify our model structure to get better result.

RAG VISUALIZATION

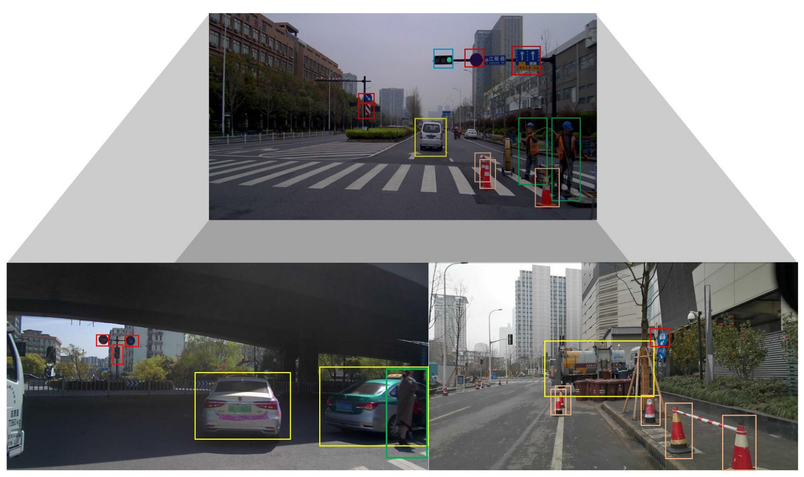
Target Picture:

<a round sign | distance: mid | position: center-right | bbox(in px): [906, 135, 960, 188]>, <a motorcycle ...

Search Result:

<a round sign | distance: mid | position: center-right | bbox(in px): [954, 70, 1017, 133]>, <a truck...

<a truck | distance: mid | position: center | bbox(in px): [417, 276, 690, 439]>, <a round sign...



REFERENCE

[1] Jain, Jitesh, et al. "Oneformer: One transformer to rule universal image segmentation." *CVPR*. 2023.

[2] Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data."CVPR. 2024.

[3] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." *ECCV* 2025.

[4] Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2024).

[5] JIA, Ding, et al. GeminiFusion: Efficient Pixel-wise Multimodal Fusion for Vision Transformer. *arXiv preprint arXiv:2406.01210*, 2024.

[6] GUO, Yunhui, et al. Adaln: a vision transformer for multidomain learning and predisaster building information extraction from images. *Journal of Computing in Civil Engineering*, 2022