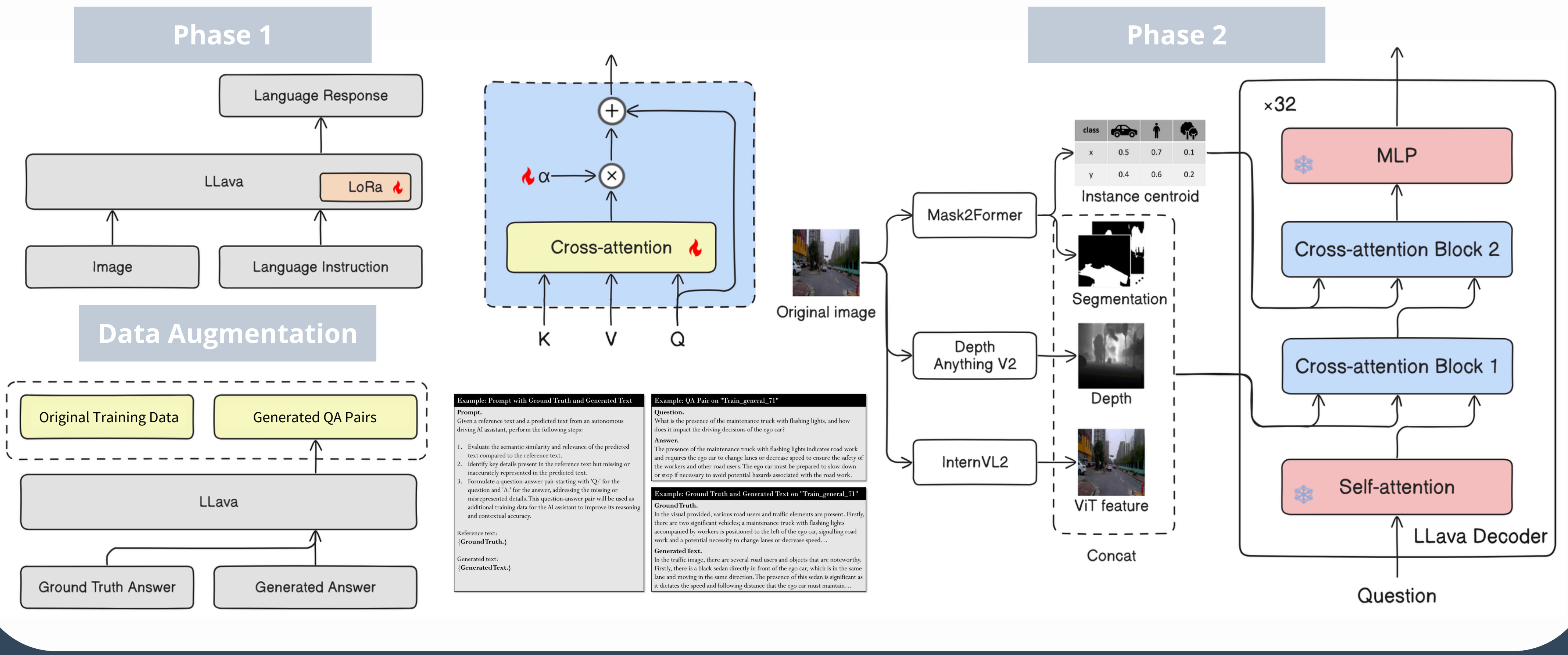


Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving

黃彥傑 (電機碩一 R13921038) 李維釗 (電信碩二 R12942089) 郭瑋羽 (電機碩一 R13921054) 吳家萱 (電機碩一 R13921068)

NTU DLCV 2024 Fall uber603 🚗

Methodology



Introduction

To tackle **Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving**, a challenging problem defined in Track 1 of the ECCV 2024 Workshop on Corner Case Scene Understanding, we propose a two-phase approach:

- In **Phase 1**, we **fine-tune LLaVA with LoRA** to enhance its ability in processing driving scene inputs and aligning the model's output style with the ground truth annotations.
- In **Phase 2**, we extract comprehensive scene information from images using multiple encoders to obtain **segmentation, instance, depth, and vision transformer (ViT) features**. These features are integrated into LLaVA through new cross-attention layers added during second-stage fine-tuning, enabling the model to reason over diverse visual cues.

Furthermore, we also performed **Data Augmentation**. We used the results produced by the fine-tuned model as the **generated text** and fed them, along with the **ground-truth text**, into the LLaVA model. By identifying inaccuracies and gaps, the model generates multiple new QA pairs, which can then be used in Phase 1 or Phase 2 to further **enhance training**.

Experiment Result

Table 1. Experiment results across different LoRA ranks

LoRA Rank	Bleu 1	Bleu 2	Bleu 3	Bleu 4	General	Reginal	Suggestion	LLM judge	Final score
32	1.49	0.86	0.51	0.03	3.00	3.09	4.44	3.51	2.91
64	1.28	0.75	0.45	0.28	5.08	5.39	4.89	5.12	4.19

Table 2. Experiment results across different settings

Index	Setting	Bleu 1	Bleu 2	Bleu 3	Bleu 4	General	Reginal	Suggestion	LLM judge	Final score
1	LoRA only (rank=64)	1.28	0.75	0.45	0.28	5.08	5.39	4.89	5.12	4.19
2	cross-attn only (all features)	1.28	0.75	0.46	0.29	5.06	5.18	4.55	4.93	4.04
3	1 + vit	1.30	0.76	0.47	0.30	5.01	5.55	4.95	5.17	4.23
4	1 + segmentation	1.31	0.77	0.47	0.30	5.57	5.51	4.53	5.20	4.26
5	1 + depth	1.31	0.76	0.46	0.29	5.07	5.48	4.82	5.12	4.19
6	1 + instance	1.31	0.77	0.47	0.30	5.55	5.48	4.42	5.15	4.21
7	1 + all features	1.30	0.76	0.46	0.29	5.72	5.56	4.64	5.31	4.34

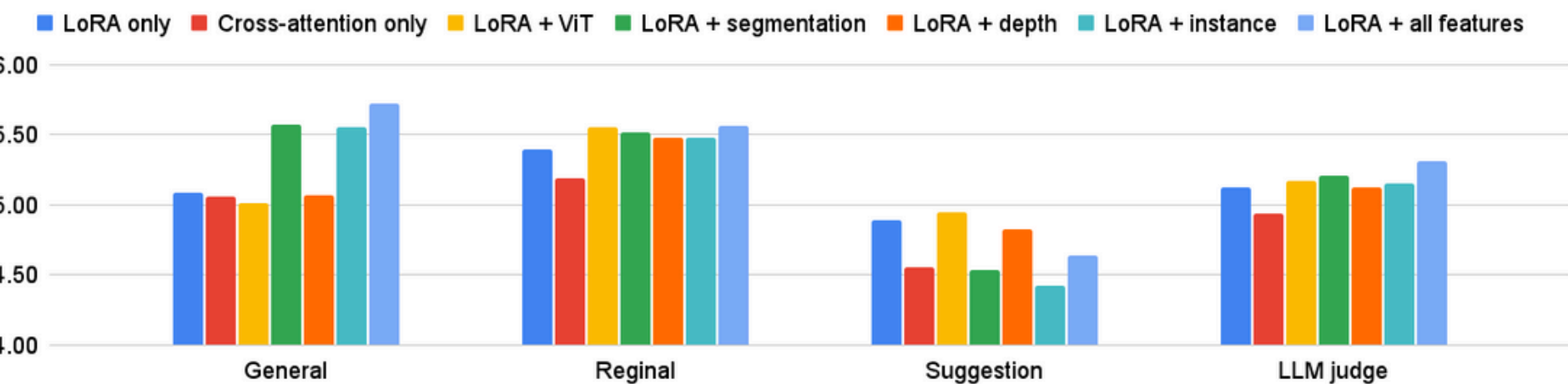


Fig 1. Bar chart comparing different settings

In Table 2, we take the LoRA only model (rank=64) as a **baseline**, as it has the best performance in Table 1, to further examine the importance of Phase 2 and the contributions of each feature individually.

- We consider that using **cross-attention only** for fine-tuning results in **suboptimal** performance, as the model tends to allocate excessive capacity to **adapting text styles** rather than **leveraging the additional visual cues** for enhanced reasoning. This limitation emphasizes the effectiveness of LoRA in facilitating alignment between the model's outputs and the ground truth.
- In addition, we observed the following relationships:
 - Segmentation** and **instance** features strengthen **General Perception**.
 - ViT** and **depth** features enhance **Driving Suggestion**.

Finally, based on the results discussed above, we utilized **all features** under the application of LoRA, effectively **combining their strengths** to achieve the optimal model performance.

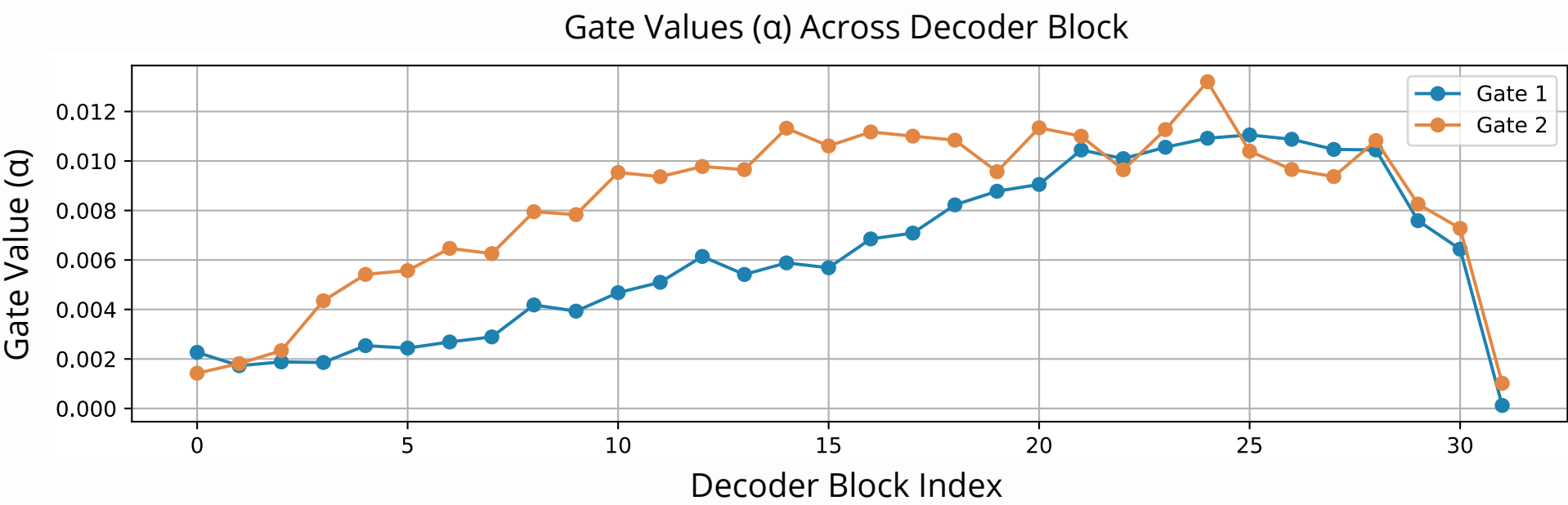


Fig 2. indicate that gate values are **largest in the latter layers** of the decoder, suggesting that **allocating more parameters to these layers** could enhance efficiency by **better utilizing the model's capacity** where it is most impactful.

Table 3. Experimental results before and after applying augmentation

Index	Setting	General	Reginal	Suggestion	LLM judge	Final score
1	LoRA only (rank=64)*	4.35	5.02	4.73	4.70	3.86
2	w/ Augmentation*	4.70	5.26	4.56	4.84	3.97
3	w/ Augmentation	4.81	5.66	4.89	5.12	4.19

* Trained on 4812 samples, with additional 5425 augmented QA-pairs

Reference

- Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2021.
- Bachlechner et al. Rezero is all you need: Fast convergence at large depth. PMLR 2021.
- Bowen Cheng et al. Masked-attention Mask Transformer for Universal Image Segmentation. CVPR 2022.
- Lihe Yang et al. Depth Anything V2. NeurIPS 2024.
- InternVL Family: A Pioneering Open-Source Alternative to GPT-4o. CVPR 2024 Oral.