CVPR Nashville JUNE 11-15, 2025

MPDrive: Improving Spatial Understanding with Marker-Based Prompt Learning for Autonomous Driving

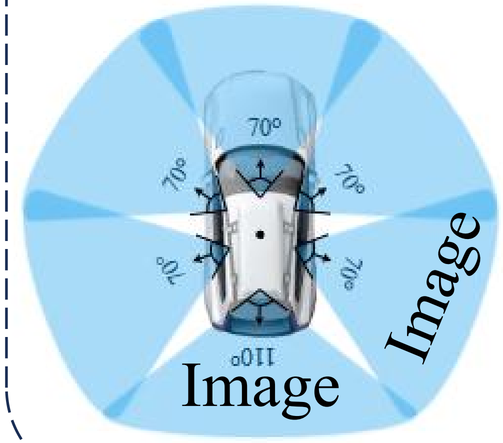
Zhiyuan Zhang*, Xiaofan Li*, Zhihao Xu*, Wenjie Peng, Zijian Zhou, Miaojing Shi, Shuangping Huang

Overview

Autonomous Driving Visual Question Answering

- Based on the six-view images of the vehicle, conduct visual question answering in autonomous driving scenarios, with related questions including perception, prediction, and planning.

Input



Question: What are the **important** objects in the current scene?

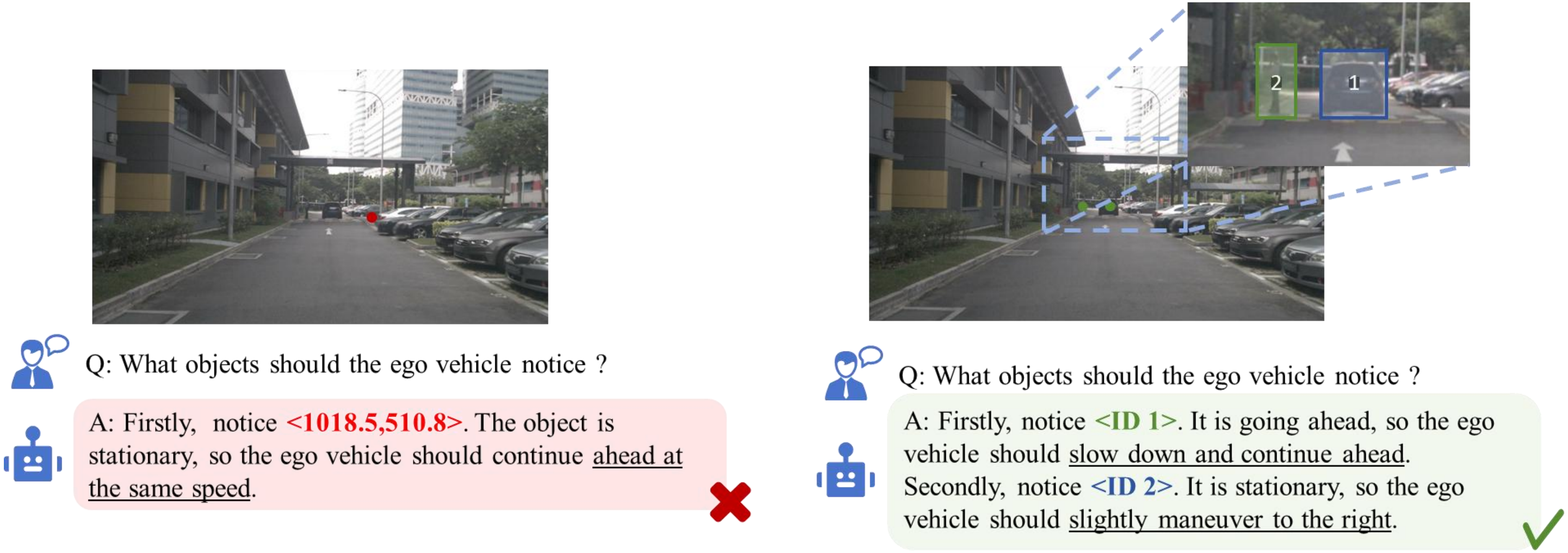
Output

Answer: There is a brown SUV at the back of the ego vehicle.

Sima, H. (2025). DriveLM: Driving with Graph Visual Question Answering. In Computer Vision – ECCV 2024 (pp. 256–274).

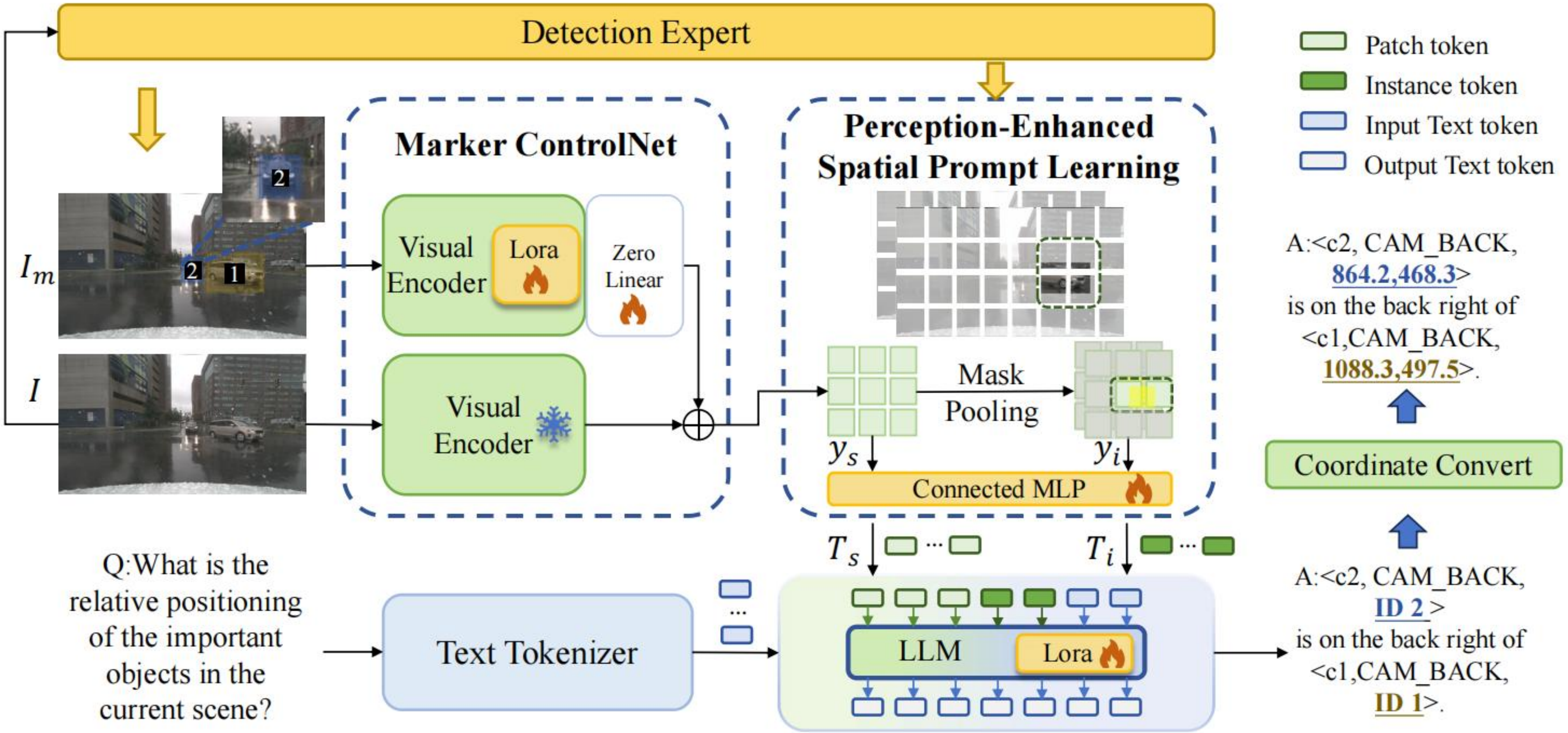
Challenges

- Main challenge: **Spatial Understanding**–semantic gaps between visual coordinate representations and textual descriptions.
- Previous methods enhanced the spatial understanding LLMs through data-driven approaches and the detection experts, but they did not address the semantic gap between spatial information and text coordinates.



Overview

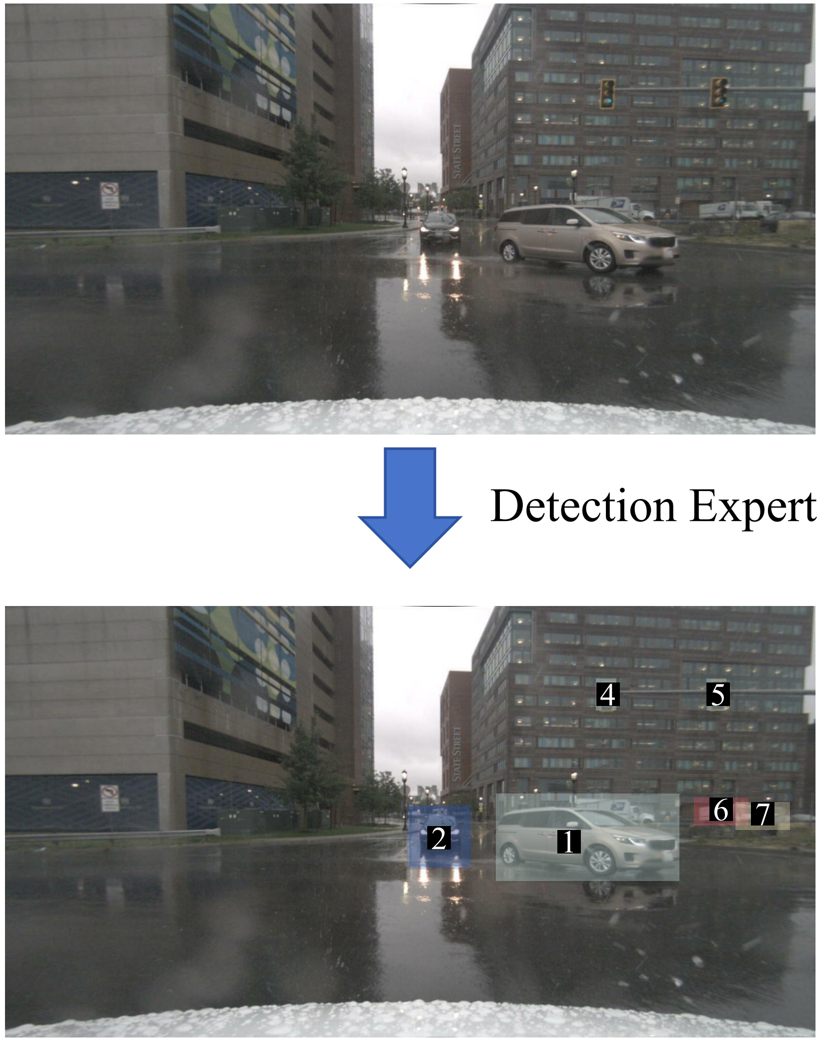
Overall Pipeline



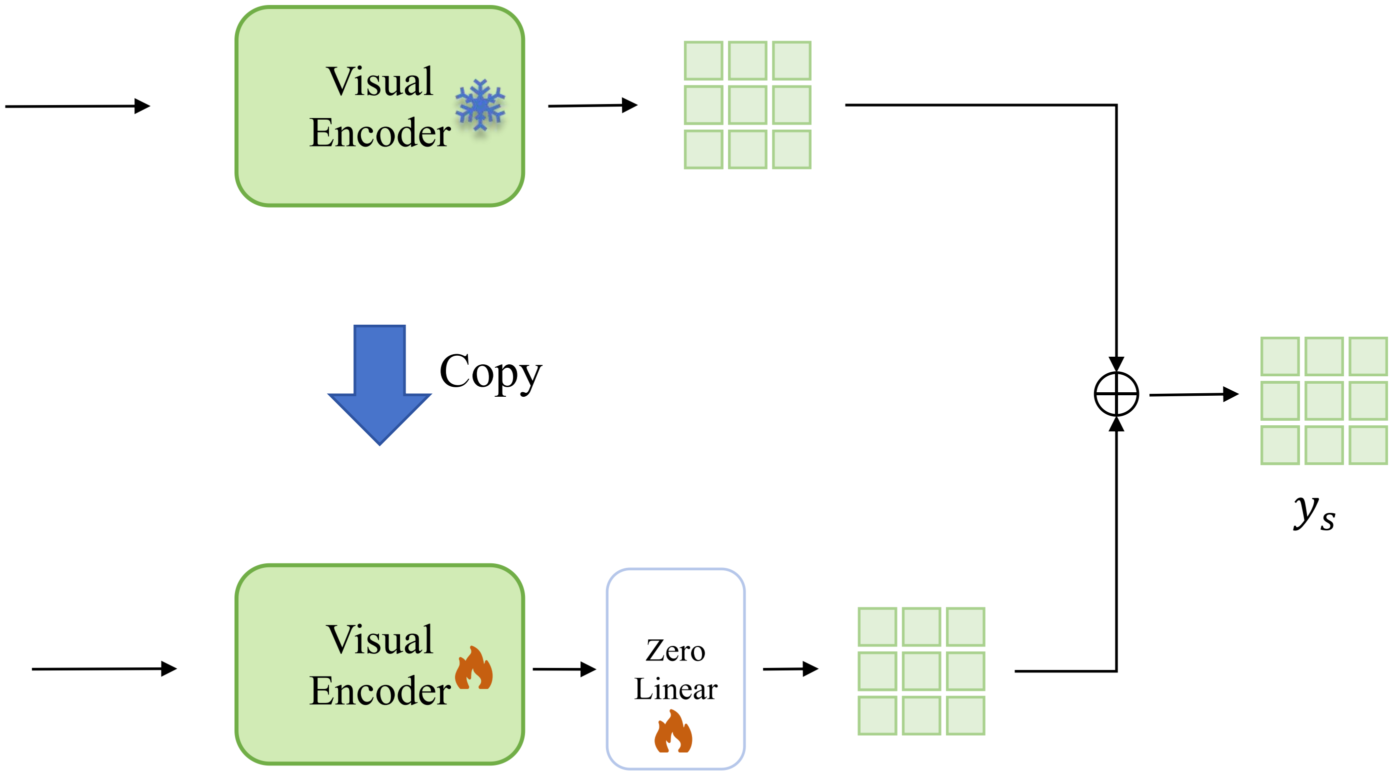
- MPDrive represents spatial coordinates using concise visual markers

Method Details

Visual Marker

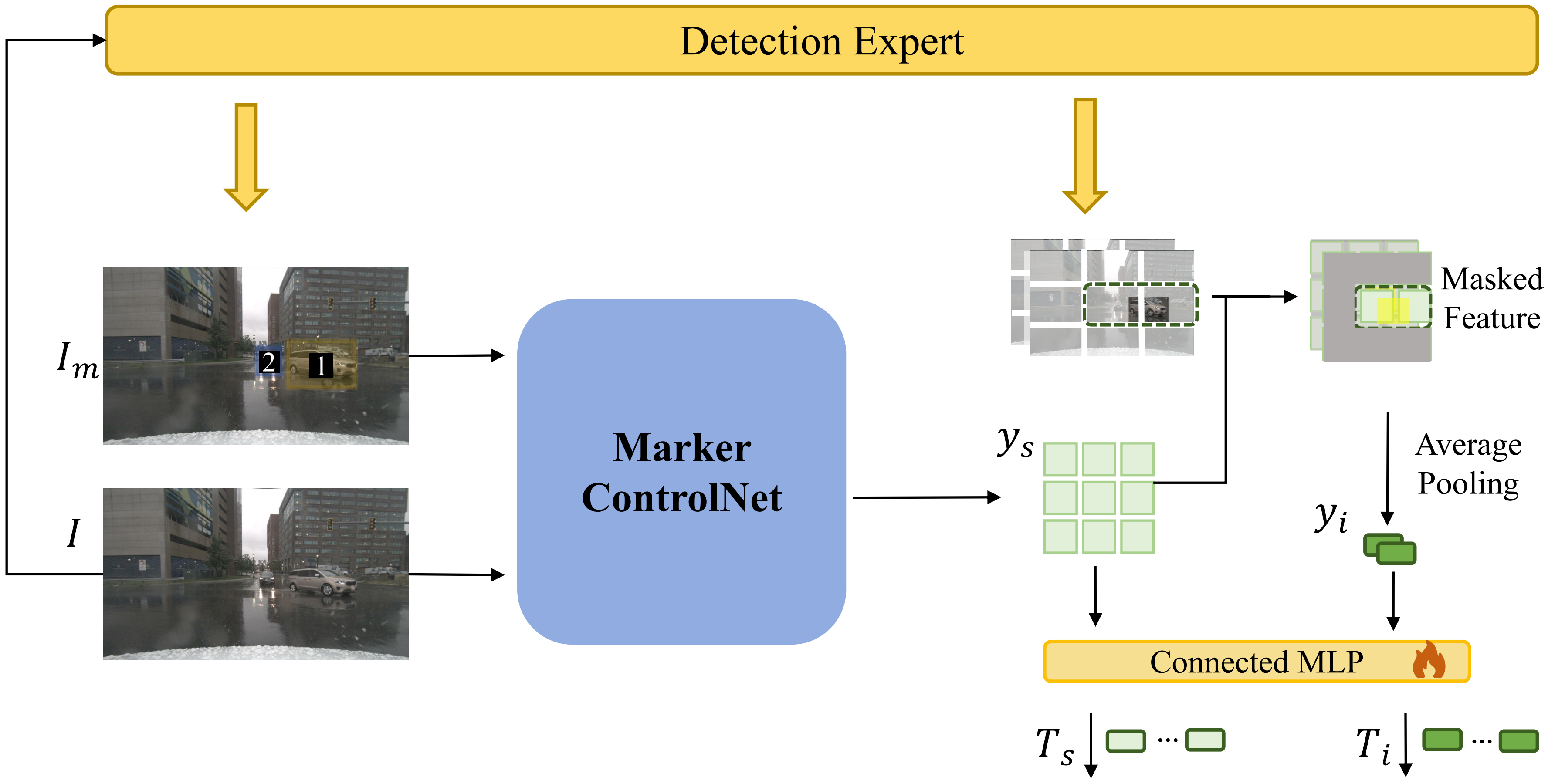


Marker ControlNet



Method Details

- Perception-Enhanced Spatial Prompt Learning



Overview

Experimental Results

- Quantitative evaluation on the DriveLM dataset

Method	Inference Schema	Spatial↑ Perception	Language↑				
		Match	Accuracy	BLEU-4	ROUGE_L	CIDEr	METEOR
DriveLM-Agent [39]	Graph	-	-	**53.09**	66.79	2.79	36.19
EM-VLM4AD [14]	Single	-	-	45.36	71.98	3.20	34.49
MiniDrive [58]	Single	-	-	50.20	73.50	3.32	37.40
LLaMA-Adapter [59]	Single	1.48	66.66	45.96	69.78	3.07	33.66
InternVL-2 [6]	Single	7.59	82.54	51.42	**77.08**	3.53	37.12
Ours: MPDrive	Single	**13.43**	**85.18**	52.71	76.98	**3.56**	**38.31**

- MPDrive outperforms existing single-round reasoning methods in both spatial awareness and language understanding, with spatial positioning metrics exceeding the second place by 76%.

Experimental Results

Qualitative evaluation on the DriveLM dataset

Overview

结论

Contributions

- We introduce MPDrive, a marker-based prompt learning framework that enhances spatial understanding in AD-VQA tasks using visual marker.
- We propose MCNet to fuses visual marker images for scene features.
- We propose Perception-Enhanced Spatial Prompt Learning (PSPL) to integrates scene- and instance-level visual prompts for multi-level spatial understanding.

Future work

- Long-Horizon Temporal Perception

The End

Thanks for watching!