



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

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



琶洲实验室
PAZHOU LAB

Overview

1 Background

2 Method

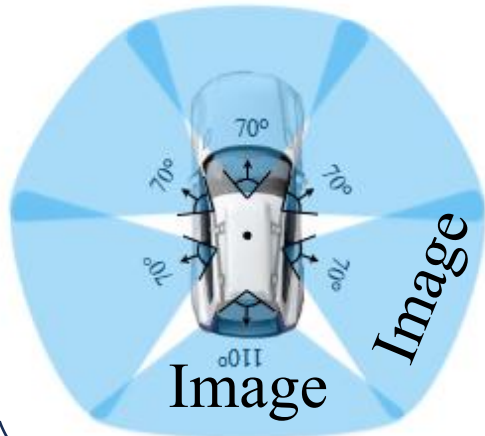
3 Experimental Results

4 Conclusion

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.

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.



Q: What objects should the ego vehicle notice ?



A: Firstly, notice **<1018.5,510.8>**. The object is stationary, so the ego vehicle should continue ahead at the same speed.



Q: What objects should the ego vehicle notice ?



A: Firstly, notice **<ID 1>**. It is going ahead, so the ego vehicle should slow down and continue ahead. Secondly, notice **<ID 2>**. It is stationary, so the ego vehicle should slightly maneuver to the right.



Overview

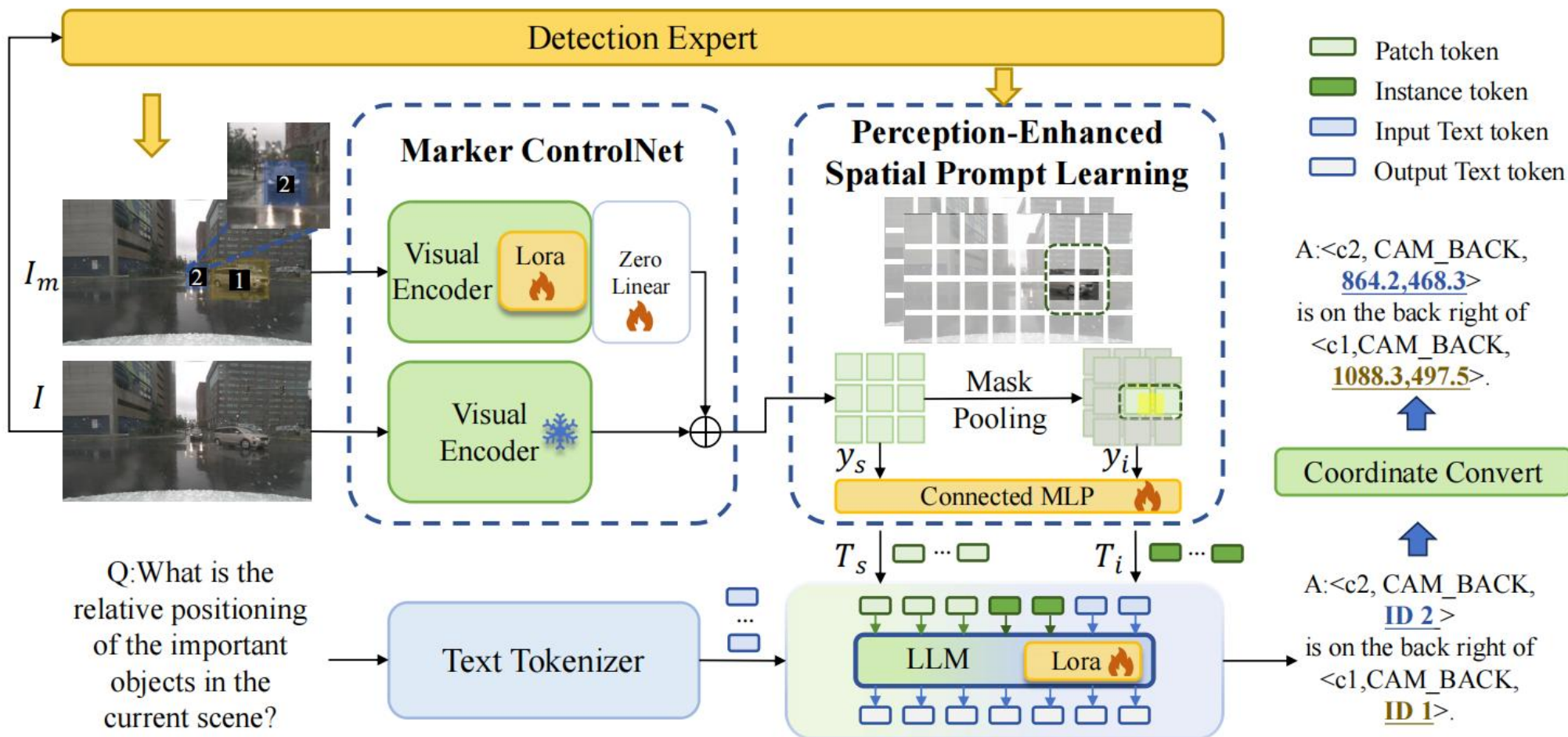
1 Background

2 Method

3 Experimental Results

4 Conclusion

Overall Pipeline



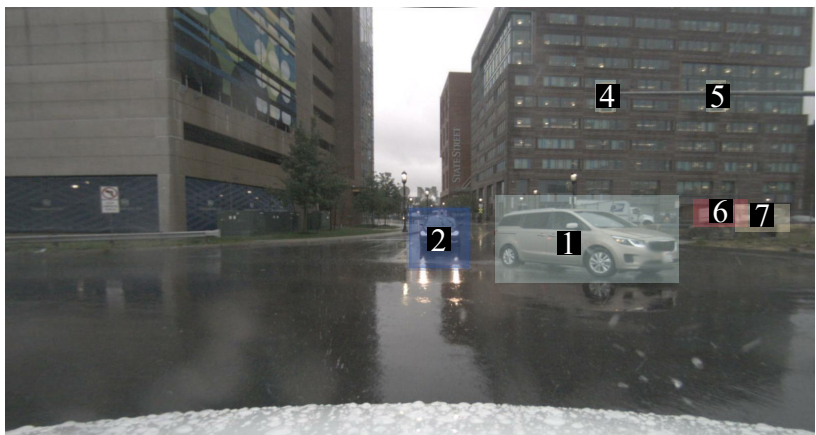
- MPDrive represents spatial coordinates using concise visual markers

Method Details

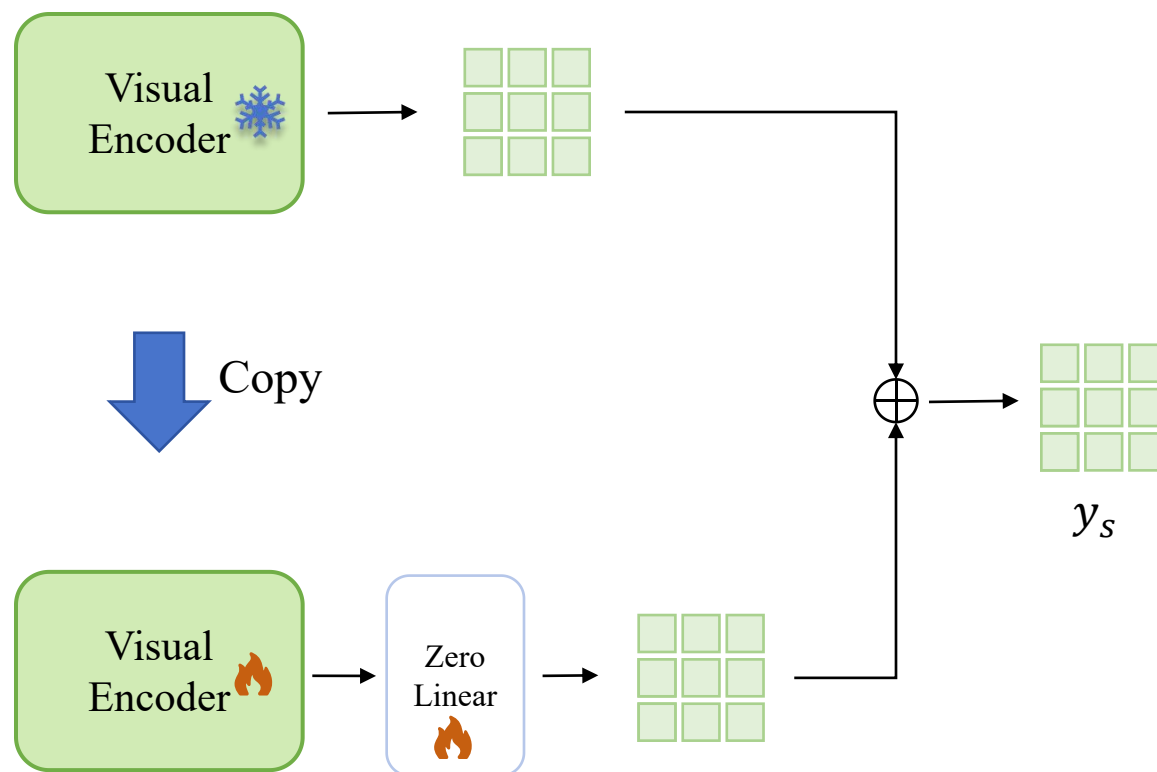
Visual Marker



Detection Expert

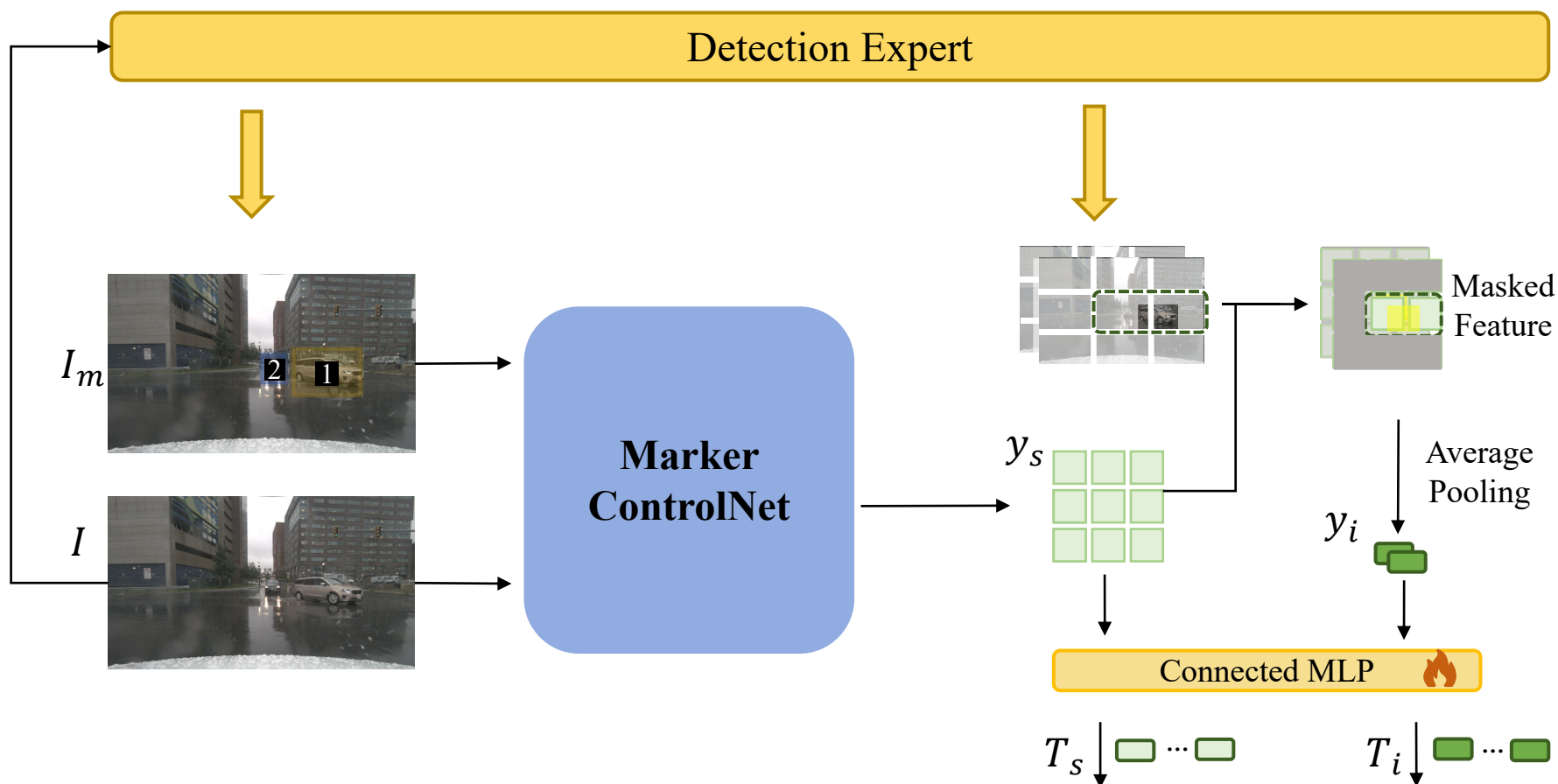


Marker ControlNet



Method Details

■ Perception-Enhanced Spatial Prompt Learning



Overview

1 Background

2 Method

3 Experimental Results

4 Conclusion

Experimental Results

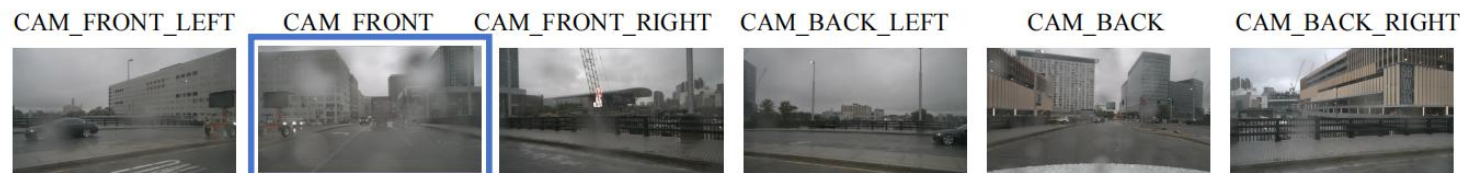
■ Quantitative evaluation on the DriveLM dataset

Method	Inference Schema	Spatial↑ Perception	Language↑				
		Match	Accuracy	BLEU-4	ROUGE _L	CIDE _r	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	<u>37.40</u>
LLaMA-Adapter [59]	Single	1.48	66.66	45.96	69.78	3.07	33.66
InternVL-2 [6]	Single	<u>7.59</u>	<u>82.54</u>	51.42	77.08	<u>3.53</u>	37.12
Ours: MPDrive	Single	13.43	85.18	<u>52.71</u>	<u>76.98</u>	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



Q: What is the priority of the objects that the ego vehicle should consider?(in descending order) ?



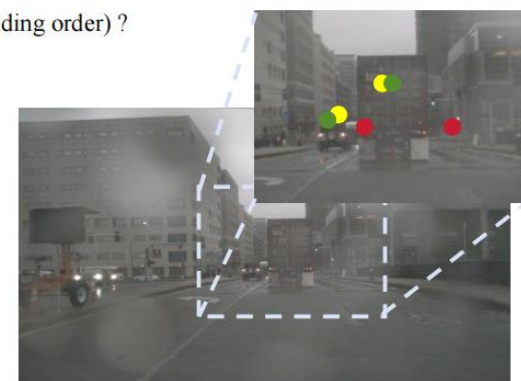
GT: <c1,CAM_FRONT_LEFT,1495.7,474.1>,
<c3,CAM_FRONT,905.0,448.3>, <c2,CAM_FRONT,761.7,528.3>.



InternVL2: The priority of the objects that the ego vehicle should consider (in descending order) is <c4,CAM_FRONT,1020.0,515.8>,
<c1,CAM_FRONT,810.0,530.0>, <c3,CAM_FRONT,13.3,565.8>,
<c2,CAM_BACK,860.8,510.0>.



MPDrive: The priority of the objects that the ego vehicle should consider (in descending order) is <c1,CAM_FRONT_LEFT,1504.0,482.0>,
<c3,CAM_FRONT,873.0,441.0>, <c2,CAM_FRONT,764.0,517.0>.



Q: What actions taken by the ego vehicle can lead to a collision with <c3,CAM_FRONT_RIGHT,1022.5,540.0>?



GT: Moderate right turn.



InternVL2: No such action will lead to a collision.



MPDrive: Moderate right turn.



Overview

1 Background

2 Method

3 Experimental Results

4 Conclusion

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 fuse visual marker images for scene features.
- We propose Perception-Enhanced Spatial Prompt Learning (PSPL) to integrate scene- and instance-level visual prompts for multi-level spatial understanding.

Future work

- Long-Horizon Temporal Perception

The End

Thanks for watching!