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

Supplementary Material

In this supplementary material, we provide further details on MPDrive and present more ablation studies. We first define the accuracy and match metrics used to evaluate performance on the DriveLM dataset. We then conduct an ablation study investigating the impact of different detection experts and image token lengths. Furthermore, the qualitative ablation examples illustrate the impact of each component on the generated responses, while the qualitative examples demonstrate the performance of MPDrive in comparison to InternVL-2.

1. More Evaluation Details

Accuracy Metric For the DriveLM dataset, both multi-choice questions and yes/no questions are used to calculate the accuracy score. The multi-choice questions include perception and behavior prediction. For perception questions, the question is "What is the moving status of the object?". We will provide 7 candidate options, randomly selecting 3 options from the incorrect answers and incorporating the correct answer to construct the multiple-choice question. Similarly, for behavior prediction questions, the question is "Predict the behavior of the ego vehicle.", with a total of 21 candidate options. The yes/no questions include perception, prediction, and planning, and the ground truth annotations only contain "yes" or "no".

Given m predicted responses $\hat{S} = (\hat{r}_1, \hat{r}_2, ..., \hat{r}_m)$ and the ground truth answers $R = (r_1, r_2, ..., r_m)$, the accuracy score can be calculated as follows:

$$Acc = \sum_{i=1}^{m} \frac{\hat{r}_i == r_i}{m}, \tag{4}$$

where $\hat{r}_i == r_i$ is a boolean expression: it equals 1 if the predicted response matches the ground truth, and 0 otherwise.

2. Ablation Study on Different Image Token Lengths

Match Metric For the DriveLM dataset, we extracted l_{gt} center coordinates $P = [p_1, p_2, ..., p_{l_{gt}}]$ from the ground truth responses and l_p center coordinates $\hat{P} = [\hat{p}_1, \hat{p}_2, ..., \hat{p}_{l_p}]$ from the predicted responses. We then calculated the proportion of coordinates in the predicted responses that have an Euclidean distance of less than 16 from the ground truth coordinates, thus obtaining the matching ratio, formulated as:

$$Match = \frac{min(\left|\left|P - \hat{P}\right|\right|_2) < 16}{l_{gt}}, \tag{5}$$

where $min(\left|\left|P - \hat{P}\right|\right|_2) < 16$ represents the number of pairs of points between the P and $\hat{P}$ for which the minimum Euclidean distance is less than 16 among all possible matches.

3. Ablation Study on Different Detection Experts

To investigate the impact of detection expert performance on spatial localization accuracy, we conducted a comparative analysis using two distinct detection models: StreamPetr and DETR3D, which achieve mAP scores of 48.20 and 50.10, respectively, on the NuScenes Val Set, as shown in the Table 5. Experimental results indicate a positive correlation between detector performance and spatial localization accuracy. Higher-performing detectors generally exhibit improved spatial localization.

To examine the effect of different image token lengths, we experiment with compressing scene-level tokens from 256 to 64 per image, thereby reducing the total scene tokens from 1,536 to 384 for six view images. As shown in table 6, this token compression strategy led to a degradation in model performance on the DriveLM dataset. Specifically, the decline in accuracy metrics suggests that reducing the number of image tokens compromised the model's ability to effectively capture and process visual information.

4. Qualitative Ablation Examples

Figure 4 demonstrates the impact of different components of MPDrive on the responses, we display the predicted coordinates from one of the most relevant images, and after introducing the Visual Marker, the predicted coordinates contain one correct answer. Following the incorporation of MCNet, the model output multiple coordinates in the front-view image, all of which were located on objects; however, the answer included irrelevant objects such as barriers and trucks. With the addition of the instance-level visual prompt, the model was able to accurately locate each coordinate. This sample indicates that the Visual Marker and MCNet contribute to the precise representation of the spatial coordinates of objects, ensuring consistency in language expression. Meanwhile, the instance-level prompt enhances

Method	mAP	Spatial↑ Perception	Language↑				
		Match	Accuracy	BLEU-4	ROUGE_L	CIDEr	METEOR
MPDrvie (DETR3D)	**50.10**	**13.76**	83.30	52.40	**76.99**	**3.58**	37.38
MPDrvie (StreamPetr)	48.20	13.43	**85.18**	**52.71**	76.98	3.56	**38.31**

Table 5. Ablation study of different detection experts.

Method	Spatial↑ Perception	Language↑				
	Match	Accuracy	BLEU-4	ROUGE_L	CIDEr	METEOR
MPDrvie (64)	**13.76**	79.37	52.35	76.95	3.54	38.10
MPDrive (256)	13.43	**85.18**	**52.71**	**76.98**	**3.56**	**38.31**

Table 6. Ablation study of different image tokens.

the spatial features of the objects, further improving the spatial perception capabilities of MPDrive.

5. More Qualitative Examples

In this section, we present more qualitative examples of MPDrive responses. Figure 5 illustrates a comparison between the response results of MPDrive and InternVL-2. In the first sample of Figure 5, for the question of identifying whether the mentioned pedestrian is an important object, InternVL-2 incorrectly answers that the pedestrian crossing the street is not an object that should be considered, however, the pedestrian on the left side is indeed significant because the ego vehicle is making a left turn, and MPDrive provides an accurate assessment in this scenario. Similarly, in the second sample, for the question of understanding the relationship between the mentioned vehicle and the traffic light, InternVL-2 incorrectly assumes that the car is unrelated to the traffic light. However, the traffic light signals influence the vehicle's position. MPDrive, with its excellent spatial perception abilities, can accurately recognize the relationship between the car and the traffic light. In the last two samples, for the questions of identifying the dangerous behaviors between the ego vehicle and other vehicles, InternVL-2 struggles to recognize the relative spatial relationships between the ego vehicle and the relevant vehicles due to a lack of strong spatial perception capabilities, thereby limiting its ability to identify potential dangerous behaviors accurately. In contrast, MPDrive successfully perceives the spatial positions of the relevant vehicles, because of its superior spatial perception abilities, allowing it to make accurate planning decisions.

In summary, MPDrive demonstrates an advantage in scenarios requiring precise spatial perception. Its ability to accurately interpret spatial relationships and identify critical objects allows it to make more informed and safer planning decisions. This enhanced spatial understanding is crucial for the effective navigation and safety of autonomous systems, highlighting the potential of MPDrive for superior performance in complex driving environments.

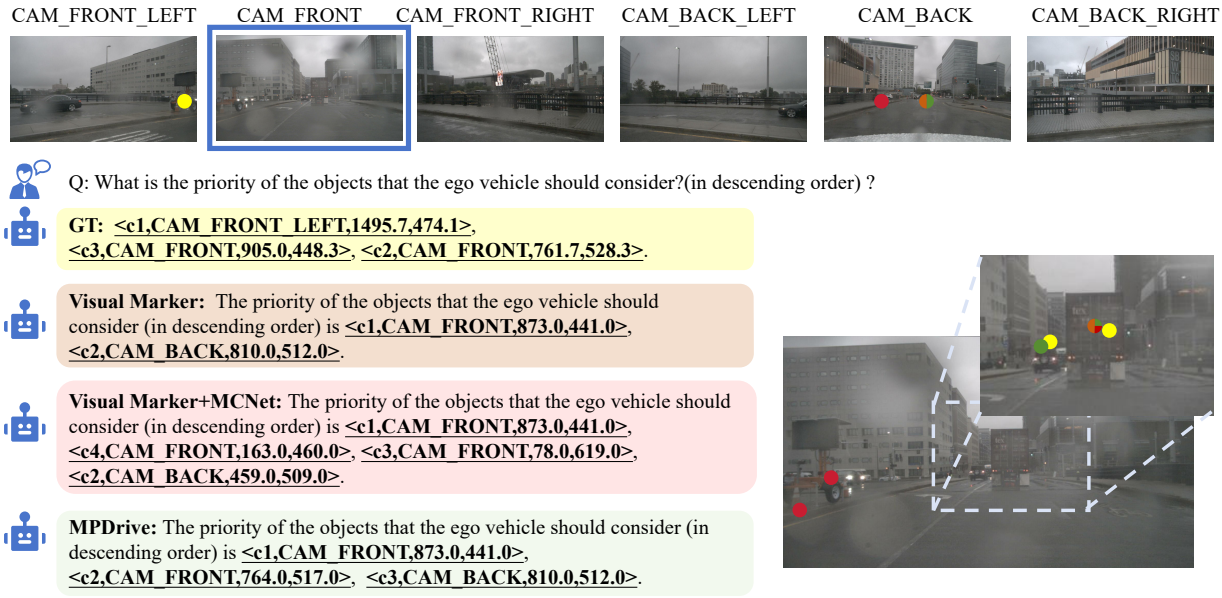

Figure 4. Comparison of different components of MPDrive on the responses. The yellow (■) area and dots represent the response and coordinates of ground truth (GT), the brown (■) area and dots indicate the response and coordinates after adding the Visual Marker, the red (■) area and dots denote the response and coordinates after adding the Visual Marker and the MCNet, and the green (■) area and dots indicate the response and coordinates of MPDrive.

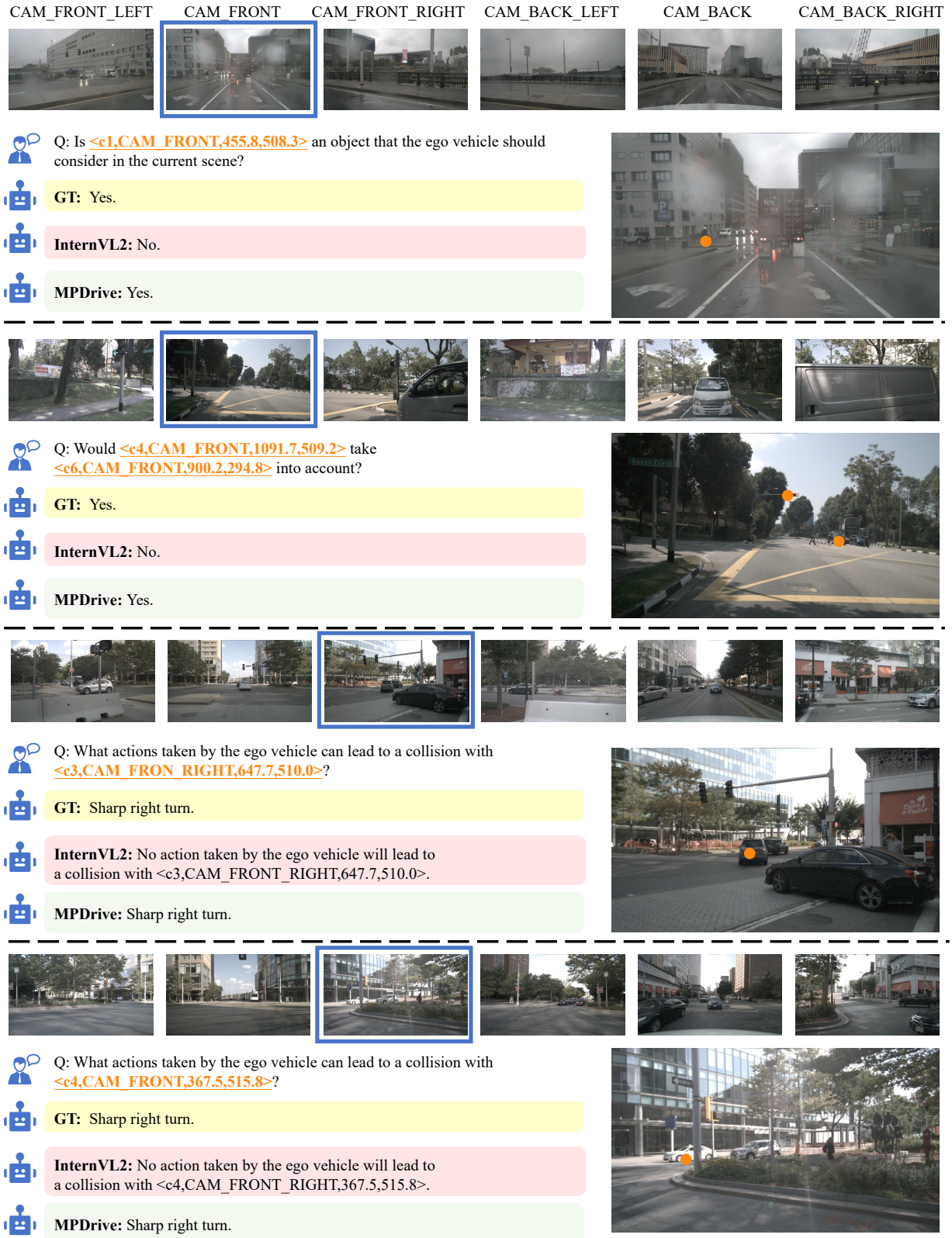

Figure 5. Comparison of the responses between InternVL-2 and our proposed MPDrive.