ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation

Supplementary Material

We provide supplementary material to complement the main paper, arranged as follows:

- Appendix A: Details on the Chat-B2D dataset.
- Appendix B: Traning Details.
- Appendix C: More results.

A. Details on the Chat-B2D dataset

To compensate for the absence of a high-quality scene text annotation dataset and promote the application of VLM in the closed-loop simulated driving scenario, we carefully design an automated annotation pipeline to extend the Bench2Drive dataset [24] to support VQA pairs, named Chat-B2D, covering diverse tasks.

A.1. Data Annotation Pipeline

As shown in Fig. A1, the automated annotation pipeline consists of three steps:

Critical object selection. Unlike mainline self-driving perception modules that process all detected objects equally, we emphasize identifying the crucial object that potentially affects the ego vehicle's driving behavior, grounded in human driving strategies. Our selection criteria include: 1) Objects have potential collisions within three seconds. 2) Leading vehicles in current and adjacent lanes. 3) Active traffic signals. 4) The vulnerable road users (VRUs), such as pedestrians/cyclists.

Description generation. We extract video clips comprising the current and five preceding frames. Subsequently, these clips, along with the ego vehicle's status and the ground truth information (*e.g.*, 2D/3D coordinates and velocity, *etc.*) of selected crucial objects, serve as input to Qwen2VL-72B [59] for multi-task generation: 1) the scene description; 2) attributes of key objects and their impact on the ego vehicle; 3) operational meta-commands and action reasoning for autonomous navigation.

History Information. During the generation process, we incorporate a queue mechanism to preserve essential historical information. The stored information comprises two principal components: 1) Environmental dynamics that capture spatiotemporal variations of critical scene elements, and 2) Ego-motion characteristics derived from comparative analysis between current speed/action and their historical counterparts across previous frames.

The generated description and collected historical information are combined with predefined question templates to create VQA pairs. Tab. A4 displays the detailed crafted prompt, and Tab. A5 shows the question templates.

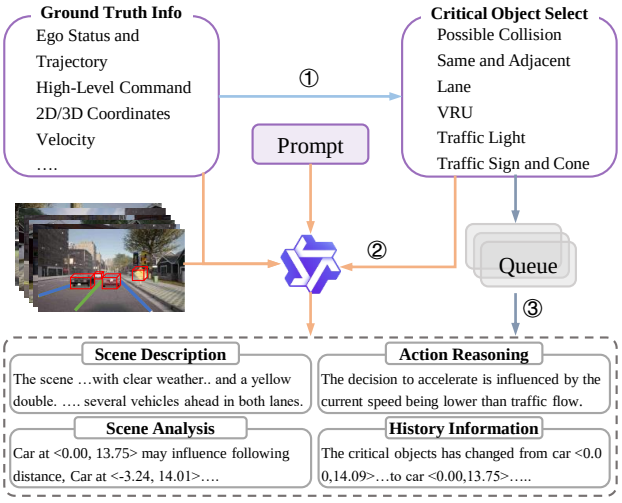

Figure A1. The automated annotation pipeline for the Chat-B2D dataset.

A.2. Chat-B2D Attribute

Through the carefully crafted prompts and the above generation pipeline, we have automatically conducted a large-scale, high-quality VQA dataset for the Bench2Drive [24], creating Chat-B2D. This dataset, including a total of 2.11M VQA pairs for training and 0.12M for validation, supports four primary categories: 1) Scene description, which provides a comprehensive overview of the driving scenarios, including weather, time of day, traffic situations, and road characteristics. 2) Behavior description of critical objects detailing their current state and intentions. 3) Meta-driving decisions and action reasoning of the ego car, such as turning left and lane following. 4) Recall of essential historical information.

B. Training Details

To accelerate the alignment of the vision-reasoning-action space and gradually enhance the reasoning and planning capabilities of our ORION, we adopt a three-stage training strategy. In each stage, the model inherits the weights from the previous stage and continues training. Additionally, we train the model for six epochs per stage with a total batch size of 32. The three-stage training strategy is as follows:

3D Vision-Language Alignment: In this first stage, we primarily train the QT-Former and the VLM while freezing the generative planner. By training on VQA pairs from

Table A1. Comparison of the Open-loop planning in nuScene. †: The ego status and planning trajectory are both processed by LLM in textual modality. ‡: The high-level command is not used during the training and testing phases.

Method	VLM-Based	Ego Status		L2 (m) ↓				Collision (%) ↓			
		BEV	Planner	1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3	-	-	-	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [19]	-	-	-	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
UniAD	-	✓	✓	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37
VAD-Base [26]	-	-	-	0.69	1.22	1.83	1.25	0.06	0.68	2.52	1.09
VAD-Base	-	✓	-	0.41	0.70	1.06	0.72	0.04	0.43	1.15	0.54
VAD-Base	-	✓	✓	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33
Ego-MLP [66]	-	-	✓	0.15	0.32	0.59	0.35	0.00	0.27	0.85	0.37
BEV-Planner [32]	-	-	-	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59
BEV-Planner++	-	✓	✓	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34
DriveVLM† [56]	✓	-	-	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27
DriveVLM-Dual [56]	✓	✓	-	0.15	0.29	0.48	0.31	0.05	0.08	0.17	0.10
OmniDrive‡ [61]	✓	-	-	1.15	1.96	2.84	1.98	0.80	3.12	7.46	3.79
OmniDrive	✓	-	-	0.40	0.80	1.32	0.84	0.04	0.46	2.32	0.94
OmniDrive++	✓	✓	✓	0.14	0.29	0.55	0.33	0.00	0.13	0.78	0.30
Senna [27]	✓	-	-	0.37	0.54	0.86	0.59	0.09	0.12	0.33	0.18
Senna	✓	✓	✓	**0.11**	**0.21**	**0.35**	**0.22**	0.04	**0.08**	**0.13**	**0.08**
EMMA† [20]	✓	-	-	0.14	0.29	0.54	0.32	-	-	-	-
ORION (**Ours**)	✓	✓	-	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37

Table A2. Ablation study of different LLMs. DS and SR denote Driving Score and Success Rate separately.

Different LLMs	Closed-loop		Runtime
	DS ↑	SR(%) ↑	
Qwen2-0.5B	66.10	36.57	**654.9ms**
LLaMA3-1.1B	72.23	48.33	684.8ms
Vicuna-7B	**77.74**	**54.62**	1106.2ms

Table A3. Ablation study of training strategy. V/L/A indicates vision/language/action space. DS and SR denote Driving Score and Success Rate separately. C/B/R refers to CIDEr/BLEU/ROUGE-L.

ID	V→L	L→A	V→L→A	Closed-loop	
				DS ↑	SR(%) ↑
1		✓		57.96	26.32
2	✓	✓		65.10	38.83
3	✓	✓	✓	**74.65**	**49.31**

Chat-B2D, we focus on aligning the vision space with the reasoning space.

Language-Action Alignment: In this stage, we unfreeze the generative planner and train the entire model except for the LLM, which is trained by LoRA [17], to predict planning trajectories without auxiliary VQA pairs. This stage primarily focuses on transmitting world knowledge from the reasoning space to the action space.

End-to-End Fine-tuning: We follow the training settings from the previous stage, with the only difference being the incorporation of joint training on VQA and planning tasks. This step further facilitates the alignment of the vision-reasoning-action space.

C. More Results

C.1. Experiments on nuScenes dataset

nuScenes Dataset. nuScenes [7] is a popular autonomous driving benchmark typically used for detection and open-loop planning evaluation. The dataset contains 1000 scenes from Singapore and Boston, with 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. Each scene spans 20 seconds and is annotated at 2 Hz. nuScenes utilizes the L2 error and collision rate as planning metrics.

Results on nuScenes. We compare the ORION with previous SOTA end-to-end autonomous driving methods on the nuScenes dataset. Here, for a fair comparison with other VLM-Based methods, we modify ORION by replacing QT-Former with the Q-Former from OmniDrive [61], and without the explicit ego status in the generative planner. As shown in Tab. A1, our ORION achieves comparable performance to classic SoTA methods [19, 26, 32] without VLM. However, compared with other VLM-Based methods, our ORION is suboptimal. We argue that this is due to the latent space of VAE being more suitable for multimodal trajectory distributions of Bench2Drive [24]. In contrast, the nuScene dataset follows a uni-modal Gaussian distribution (with straight trajectories accounting for about 70%).

Additionally, as highlighted in BEV-Planner [32] and Ego-MLP [66], even a simple MLP decoder with ego sta-

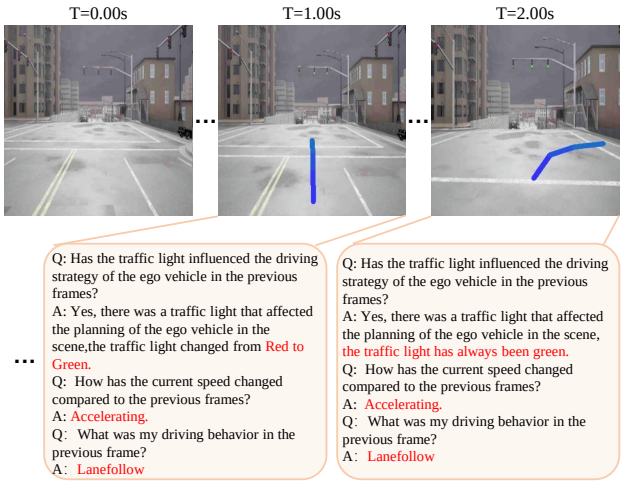

Figure A2. Qualitative results of historical information memory and retrieval on Bench2Drive open-loop validation set.

tus can achieve strong open-loop planning performance on nuScenes. Thus, in the main paper, we primarily focus on evaluating ORION's closed-loop performance on the Bench2Drive dataset.

C.2. More Ablation Studies on Bench2Drive

Ablation of different LLMs. We conduct the ablation experiments to investigate the influence of the diverse LLMS. As shown in Tab. A2, although reducing the parameters of LLM decreases the performance, it increases the runtimes. We argue that using more advanced but fewer parameter LLMs could maintain the driving performance while improving runtimes.

Ablation of training pipeline. To facilitate the vision-language-action space alignment of our model, we implement a progressive space alignment training strategy. We validate the effectiveness of the training pipeline, and the results are presented in Tab. A3. Here, the QT-Former of our model does not incorporate collision loss or long-term memory bank with history queries. Specifically, through our second-stage training (ID-2), ORION achieves a significant improvement by +7.14 DS and +12.51% SR compared to direct training planning without the first stage (ID-1). After completing the third-stage training (ID-3), our model further improved the performance and achieved optimal (74.65 DS and 49.32 SR), demonstrating the effectiveness of our training strategy.

C.3. More Qualitative Results

Historical information memory and retrieval. Benefiting from the introduced long-term memory bank and history queries in QT-Former, our ORION could store and retrieve historical information, as illustrated in Fig. A2. Our model could perceive critical elements (*e.g.*, traffic light) changes in previous and current times.

Scene understanding and action reasoning. Fig. A3 shows scene understanding and action reasoning results of ORION. It could be observed that ORION could not only accurately perceive detailed scene information but also identify key objects influencing the ego vehicle's behavior and infer appropriate motion decisions. Even in extreme situations (e.g., a pedestrian suddenly crosses the road in Fig. A3(b)), our model maintains robust performance, highlighting its superior reasoning and decision-making ability.

Users: Could you describe the overall environment and objects captured in the images provided?
ORION: The scene depicts a nighttime driving scenario with wet road conditions, indicated by the reflective surface of the road. The ego vehicle is positioned behind a white sedan, which is stationary. Ahead, there is a vehicle with its brake lights illuminated, suggesting it is either slowing down or stationary. The surrounding area is dark, with minimal visibility beyond the reach of the headlights.

Users: Where are the critical objects in the scene and what impact do they have on the ego vehicle?
ORION: Car at <1.23, 7.86> directly in front of ego vehicle, stationary, may influence braking distance.

Users: Please describe your driving behavior and explain the reasons.
ORION: You should slow down rapidly and follow the lane. The pedestrian crossing the street requires immediate attention to ensure safety. The car in the same lane is moving at a similar speed, so maintaining the current lane is appropriate. The car in the opposite lane is moving faster, so changing lanes is not advisable at this moment.

Users: What is the current behavior of the vehicle?
ORION: Lane follow.

(a)

Users: Can you summarize what the panoramic images show?
ORION: The scene depicts an urban street during daytime with clear weather conditions. The road appears dry and well-maintained. A pedestrian is crossing the street from right to left, approximately 5.65 meters ahead and 3.15 meters to the left of the ego vehicle's center position. The surrounding area includes buildings on both sides of the street, with some storefronts visible on the right side.

Users: Can you pinpoint the critical objects in the scene and describe their influence on the ego vehicle?
ORION: Pedestrian at <0.15, 6.90> is 6.90 meters ahead and 0.15 meters to the right of the ego vehicle, moving slowly and posing a potential hazard.. Car at <-4.01, 25.00> is 25.00 meters ahead and 4.01 meters to the left of the ego vehicle, traveling in the same direction at a speed of 8.05 m/s.

Users: Please describe your driving behavior and explain the reasons.
ORION: You should slow down rapidly and follow the lane. The pedestrian crossing the street requires immediate attention to ensure safety. The car in the same lane is moving at a similar speed, so maintaining the current lane is appropriate. The car in the opposite lane is moving faster so changing lanes is not advisable at this moment. ,

Users: What is the current behavior of the vehicle?
ORION: Lane follow.

(b)

Figure A3. Qualitative results for scene understanding and action reasoning on Bench2Drive open-loop validation. From top to bottom, each sub-figure displays the multi-view input and traffic conditions in Bird's Eye View (BEV) of the current scene, the scene understanding, and the reasoning result. The red rectangles indicate the critical objects influencing the action of the ego vehicle, while the red text highlights our method's correct scene comprehension.

Table A4. Prompts fed into Qwen2VL to generate corresponding response.

Prompt 1: Scene Description

Suppose you are driving, generate a description of the driving scene which includes the key factors for driving planning, including the traffic conditions, weather, time of day and road conditions, traffic signs, and traffic lights that affect the driving of the ego vehicle if it exists, indicating smooth surfaces or the presence of obstacles; The description should be concise, and accurate to facilitate informed decision-making. Please make sure the traffic light colors you provide are accurate; otherwise, give 'unknown.'

Prompt 2: Critical Objects Analysis

I will provide you with several critical objects that are most important to my short-term command in the last image of the video. I provide you with 2d coordinates, which are two points of the top-left and bottom-right coordinates, and the 3d position and velocity information of these critical objects: {objects_desc}. Please describe their action and explain why they are most important, including their speed, position, heading, and influence on ego vehicle. Please associate these objects with the objects in the image and please remember the ego vehicle is located at the **center of the bottom edge of the picture**.

Prompt 3: Expert Meta-Decision

Besides, I will provide you speed, historical trajectory and future driving behaviors of ego vehicle, which can be divided into SPEED decisions and COMMAND decisions, SPEED includes keep, accelerate, decelerate, while COMMAND includes left, right, straight, lane follow, change lane left, change lane right. Your current speed is {ego_vel} m/s, historical trajectory is {ego_his_trajs}. The next SPEED decision is {speed_decision}, the next COMMAND decision is {path_decision}. Please analyze the reasons for the future driving behaviors or the reason why ego vehicle can {path_decision} based on the driving environment, including the behavior of other traffic participants, especially the critical objects, road conditions, and traffic light status.

Example:

You should refer to the following example and format the results like {"description": "xxx","critical_objects": "xxx", "action": "{speed_decision} and {path_decision}"}}:

{{ "description": "The scene captures a moment of urban life framed by a red traffic light in mid-transition. To the right, a pedestrian crossing, ..., waiting for the signal to change. Directly ahead, ... On the left, the sidewalk bustles with people of all ages, ... Behind this foreground of orderly traffic and pedestrian movement, ..."

"critical_objects: "["Car at <-0.24, 7.56> directly in front of ego vehicle, ...", "Car at <-2.64, 10.00> ..., moving at a slower speed, may influence left change."]"

"action": "Slow down and right lane change. - The decision to change lanes is influenced by the need to overtake Car at <-0.24, 7.56> in front of the ego vehicle. - There are no traffic lights for the vehicle,... - Pedestrians are visible on the sidewalk to the right, ..." }}

If it has no critical_objects, you should refer to the following example and format the results like {{"description": "xxx", "critical_objects": [], "action": "xxx"}}.

Table A5. A list of question templates for diverse VQA tasks.

Type 1: Scene Description

1. What can you tell about the current driving conditions from the images?
2. What can be observed in the panoramic images provided?
3. Can you provide a summary of the current driving scenario based on the input images?
4. What can you observe from the provided images regarding the driving conditions?
5. Please describe the current driving conditions based on the images provided.
6. Can you describe the current weather conditions and the general environment depicted in the images?
7. Please describe the current driving conditions based on the input images.
8. Could you summarize the current driving conditions based on the input images?
9. Please provide an overview of the current driving conditions based on the images.
10. Can you summarize what the panoramic images show?
11. Can you describe the overall conditions and environment based on the images?
12. Could you describe the overall environment and objects captured in the images provided?

Type 2: Critical Objects Analysis

1. Where are the critical objects in the scene and what impact do they have on the ego vehicle?
2. Identify the significant objects in the scene and their specific impacts on the ego vehicle.
3. Can you pinpoint the critical objects in the scene and describe their influence on the ego vehicle?
4. Which objects in the scene are critical, and what effects do they have on the ego vehicle's movement?
5. Please describe the critical objects in the scene, their positions, and the influence they have on the ego vehicle.

Type 3: Interpretable Action of Ego Vehicle

1. Please describe your driving behavior and explain the reasons.
2. What is the current behavior of the vehicle?

Type 4: Historical Information

1. What are the differences between the current scene and the past scene in terms of critical objects?
2. How do the critical objects in the current scene differ from those in the past scene?
3. What changes have occurred in the critical objects between the current and past scenes?
4. What are the differences in critical objects between the present scene and the previous scene?
5. What distinctions exist between the critical objects of the current scene and those of the past scene?
6. In the past few frames, has a traffic light affected the driving strategy of the ego vehicle?
7. Within the recent frames, has the driving strategy of the ego vehicle been influenced by a traffic light?
8. In the last few frames, has the driving strategy of the ego vehicle been impacted by a traffic light?
9. Has the driving strategy of the ego vehicle been affected by a traffic light in the past few frames?
10. Has the traffic light influenced the driving strategy of the ego vehicle in the previous frames?
11. How has the current speed changed compared to the previous frames?
12. What was my driving behavior in the previous frame?