HERMES: A Unified Self-Driving World Model for Simultaneous 3D Scene Understanding and Generation

Supplementary Material

S1. Additional Experiments

S1.1. Training Details

The BEV-based tokenizer utilizes the OpenCLIP ConNext-L backbone [3, 6, 8], while other modules in the tokenizer and Render are trained from scratch. The LLM is derived from InternVL2-2B [1, 2]. The resolution of the input image is 1600×900, while the BEV-based world tokenizer adopts the same hyperparameters as BEVFormer v2-base [10], with the size of the encoded scene set to $w = h = 200$ and a BEV channel dimension of 256. The z and c' in the BEV-to-point clouds Render are set to 32. For future generation, we forecast scene evolution over 3 seconds, i.e., $\Delta t = 3$. The frame-wise weights in Eq. 4 of the main paper are empirically defined by $\lambda_i = 1 + 0.5 \times i, i \in \{0, \cdots, 3\}$, corresponding to the point clouds from 0 to 3s. The training of HERMES is structured into three stages and detailed below. Additional details are provided in Tab. S1.

Stage-1: Tokenizer Traning. In initial stage, we train the world tokenizer $\mathcal{E}$ and Render $\mathcal{R}$ to convert current images (I_t) into point clouds (P_t), following $P_t = \mathcal{R}\left(\mathcal{E}\left(I_t\right)\right)$. We utilize 12Hz data from the nuScenes training set for the tokenizer and Render learning.

Stage-2: BEV-Text Alignment and Refinement. This stage encompasses BEV-Text alignment and refinement tuning phases. The alignment phase aims to establish vision-language alignment between the input and output BEV of the LLM, training only the in-projections for flattened BEV embeddings and out-projections for the encoded BEV. To alleviate data deprivation, we propose a simple data augmentation involving masking one of the multi-view images, splicing the caption from the visible view, and using the unprocessed multi-view scene descriptions. This approach increases the multi-view image-text pairs to ∼200K, a sevenfold increase from the nuScenes keyframes. In the refinement phase, all parameters are unfrozen, and the LLM is fine-tuned using LoRA [4]. The alignment phase employs NuInteract [13] dense caption data, while the refinement phase adapts labeling styles using scene description data from OmniDrive-nuScenes [9].

Stage-3: Understanding and Generation Unification. Building on the understanding gained in the first two stages, we introduce future generation modules to generate point clouds at different moments. We train using nuScenes keyframes, descriptions, and general conversation annotations from OmniDrive-nuScenes.

Table S1. Training details of HERMES. -/- in Stage 2 indicates BEV-text alignment/refinement.

Config	Stage 1	Stage 2	Stage 3
Optimizer	AdamW	AdamW	AdamW
Learning Rate	2e-4	2e-4/4e-4	4e-4
Training Epochs	6	3/6	36
Learning Rate Scheduler	Cosine	Cosine	Cosine
Batch Size Per GPU	1	4	4
GPU Device		32×NVIDIA H20	

Table S2. Ablation on scaling potential of the LLM.

# LLM Params	Generation				Understanding		
	0s ↓	1s ↓	2s ↓	3s ↓	MTETOR ↑	ROUGE ↑	CIDEr ↑
0.8B	0.668	1.015	1.379	1.809	0.372	0.318	0.703
1.8B	0.645	**0.984**	1.333	1.718	0.377	0.321	0.720
3.8B	**0.643**	0.991	**1.321**	**1.701**	**0.381**	**0.325**	**0.730**

Table S3. Ablation on the number n of world queries.

n	Generation				Understanding		
	0s ↓	1s ↓	2s ↓	3s ↓	MTETOR ↑	ROUGE ↑	CIDEr ↑
1	0.658	0.996	1.328	1.725	0.376	0.320	0.712
2	0.656	0.995	**1.324**	1.720	0.377	0.321	0.714
4	**0.645**	**0.984**	1.333	**1.718**	0.377	**0.321**	**0.720**
8	0.667	1.028	1.361	1.744	0.376	0.321	0.713
16	0.658	0.999	1.354	1.748	**0.378**	0.321	0.716

S1.2. Additional Ablation Study

Unless otherwise specified, we perform ablation studies trained on a quarter of the nuScenes training scenes. Default settings are marked in green.

Analysis on the scaling potential of the LLM. We first explore the scaling potential of our HERMES, as shown in Tab. S2. Scaling up LLMs yields consistent gains in 3D scene understanding and point cloud generation, and we utilize the 1.8B LLM form InternVL2-2B [1, 2] as a trade-off. This indicates that the broader world knowledge acquired during pre-training enhances these tasks, suggesting potential benefits from further scaling.

Analysis on the number of world queries. The world queries facilitate knowledge transfer between the LLM and the Render for future scenarios. We then evaluate the impact of the number of queries n for each group, as shown in Tab. S3. We find that world queries do not adversely affect text understanding quality. However, increasing the

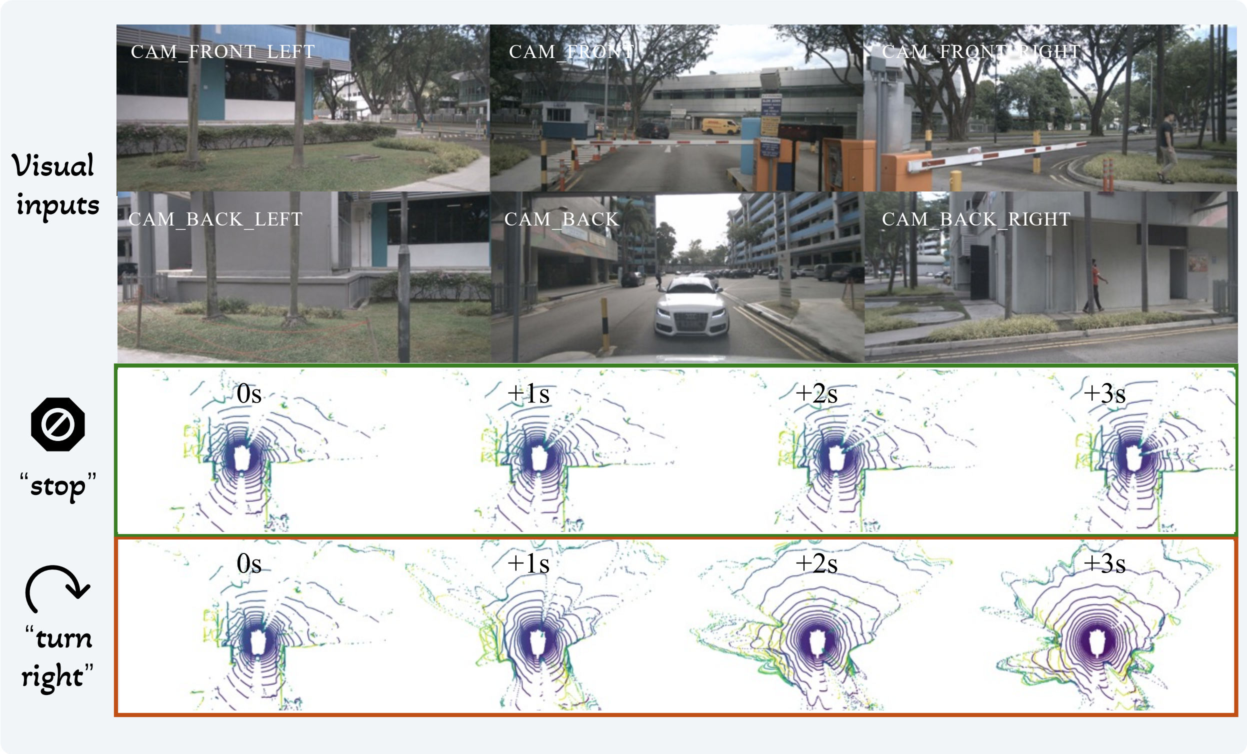


Figure S1. Qualitative results of Hermes conditioned on different future ego-motion conditions. From top to bottom, each sub-figure displays the multi-view input of the current scene, scene evolution predicted with a "stop" future ego-motion, and scene evolution predicted with a "turn right" ego-motion.

Table S4. Comparison of generation ability.

	Generation				Understanding		
	0s ↓	1s ↓	2s ↓	3s ↓	MTETOR ↑	ROUGE ↑	CIDEr ↑
Copy&Paste	-	1.27	2.12	2.66	-	-	-
ViDAR [12]	-	1.12	1.38	1.73	-	-	-
Hermes	**0.59**	**0.78**	**0.95**	**1.17**	**0.384**	**0.327**	**0.741**

Table S5. VQA results on NuScenes-QA.

Method	Reference	Modality	Acc. (%) ↑
LLaVA [5]	NeurIPS 23	Camera	47.4
LiDAR-LLM [11]	arXiv 23	LiDAR	48.6
BEVDet+BUTD [7]	AAAI 24	Camera	57.0
BEVDet+MCAN [7]	AAAI 24	Camera	57.9
CenterPoint+BUTD [7]	AAAI 24	LiDAR	58.1
CenterPoint+MCAN [7]	AAAI 24	LiDAR	59.5
OmniDrive [9]	CVPR 25	Camera	59.2
Hermes	-	Camera	**61.9**

number of world queries leads to a decline in performance, likely due to redundant information and optimization challenges. Therefore, we choose to include four world queries per group for future generations.

Analysis on generation ability. We finally compare our future point cloud generation ability trained on the full training set against a Copy&Paste baseline, where Copy&Paste simply duplicates the current ground-truth point cloud for future observations. As shown in Tab. S4, this baseline fails to account for point cloud changes due to movement and occlusion, demonstrating that Hermes truly learns to understand 3D scenes and predict their future evolution.

S1.3. Understanding on NuScenes-QA

The NuScenes-QA [7] is another multi-modal VQA benchmark for driving scenarios, featuring primarily single-word answers focused on perception. We fine-tune Hermes on the NuScenes-QA training set to align with its style and length, and the results are shown in Tab. S5. Hermes achieves superior performance, outperforming LLaVA [5] by 14.5% and the point cloud method CenterPoint+MCAN [7] by 2.4%. This showcases Hermes's strong 3D scene understanding capabilities via its unified BEV representation, especially considering it requires no 3D object detection supervision.

S2. Discussion

The integration of Bird's-Eye View (BEV) representations as input for Large Language Models (LLMs) presents distinct advantages in our Hermes. Unlike conventional multi-view processing approaches that process individual camera streams independently, the BEV-based tokenization establishes a unified spatial coordinate system that inherently preserves geometric relationships across views while maintaining object interaction patterns. This spatial consolidation addresses the inherent limitations of vision-language models in interpreting multi-perspective scenarios, where disconnected 2D projections fail to capture the holistic 3D environment context. By strategically compressing high-resolution multi-view inputs (1600×900 per view, for example) into a compact BEV latent space through our downsampling block, we achieve efficient token utilization (2,500 tokens vs. $\sim$47,000 tokens for raw view processing) without exceeding standard LLM context windows. Crucially, the spatial-aware BEV features enable synergistic knowledge transfer between scene understanding and generation tasks through our world query mechanism, i.e., the positional correspondence between text descriptions and geometric features permits causal attention patterns that enrich future predictions with linguistic context. Our experiments on nuScenes demonstrate that this spatial-textual alignment contributes substantially to the 32.4% reduction in generation error and 8.0% CIDEr improvement, validating BEV's dual role as both information compressor and cross-modal interface.

S3. More Qualitative Results

This section presents further qualitative results of HERMES on controllability and the unification ability of understanding and generation.

Potential for Controlled Scene Generation. As shown in Fig. S1, we observe the capability of HERMES to generate future point cloud evolution conditioned on specific ego-motion information, such as “stop” or “turn right”. This showcases the potential of HERMES as world simulator and its ability to understand complex world scenarios deeply.

Unification of understanding and generation. More qualitative results on future generations and scene understanding are illustrated in Fig. S2. Our HERMES effectively captures future scene evolution (with the ground truth ego-motion information for better comparison), such as the corner of the building keeps moving backward, noted in the red circle in Fig. S2b. While HERMES achieve an encouraging integration of understanding and generation, it faces challenges in complex scenes (e.g., significant left turns and occlusions as in Fig. S2a) and low-quality visible light conditions (e.g., nighttime driving as in Fig. S2c). Despite the complexity of the scenarios, HERMES still makes reasonable predictions about the emerging parts of future scenes.

References

[1] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024. 1

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024. 1

[3] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023. 1

[4] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *Proc. of Intl. Conf. on Learning Representations*, 2021. 1

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. of Advances in Neural Information Processing Systems*, pages 34892–34916, 2023. 2

[6] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2022. 1

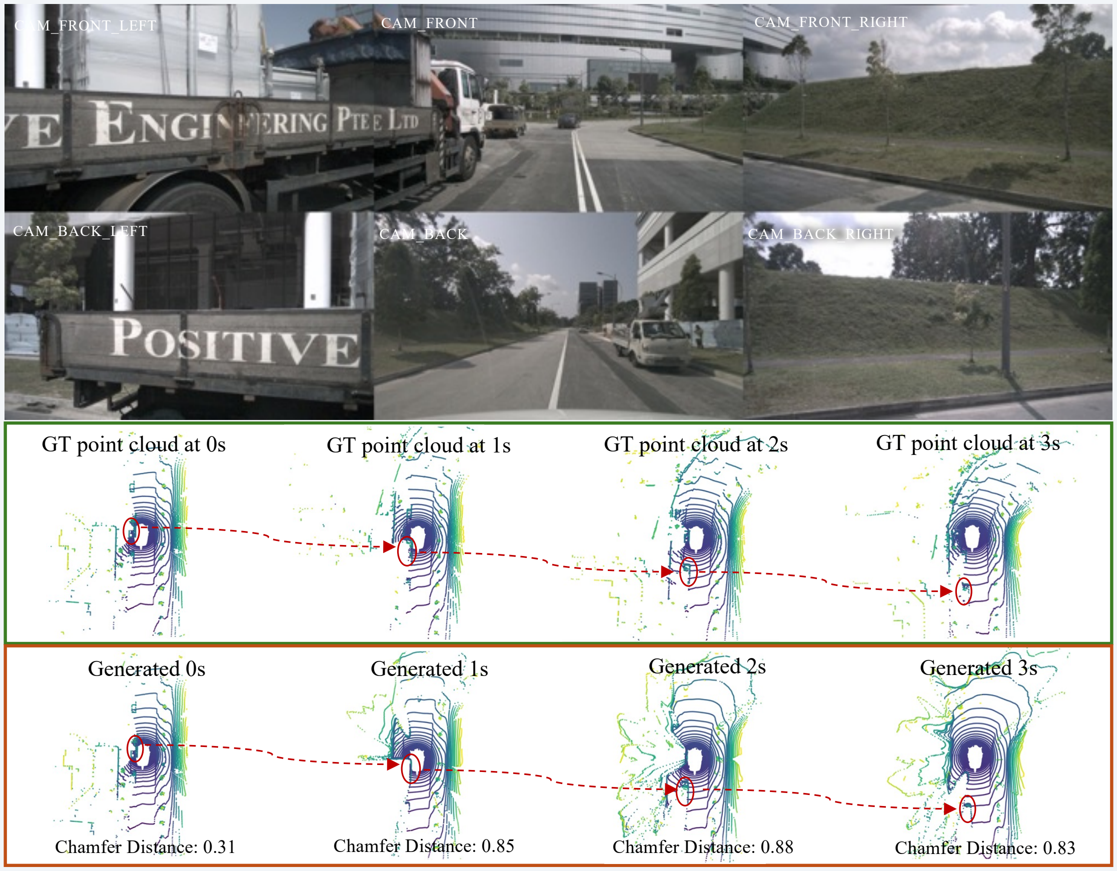
[7] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proc. of the AAAI Conf. on Artificial Intelligence*, pages 4542–4550, 2024. 2

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Intl. Conf. on Machine Learning*, 2021. 1

[9] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025. 1, 2

[10] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023. 1

[11] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of



Q: Can you provide a summary of the current driving scenario based on the input images?
A: The vehicle is positioned on a clear, marked road during what appears to be daytime with good visibility and dry weather conditions. To the left, there's a **stationary construction vehicle**, indicating potential work nearby or a temporary halt in traffic. Straight ahead, the road is free of obstructions with a few vehicles at a distance, suggesting a smooth flow of traffic. The road is bordered by **greenery on both sides**, with a **grassy slope to the right** and trees lining the road, contributing to a serene driving environment. In the rear view, there are **no vehicles closely following**, providing a sense of ease for potential maneuvers. The overall scene suggests a non-congested urban or suburban setting with minimal traffic activity.

Q: Is there a traffic signal or stop sign visible?
A: **No**, there is no traffic signal or stop sign visible in the current view.

Q: What type of vehicles are visible in the current driving scene?
A: **There is a construction vehicle on the left side** and a white SUV on the right side of the road.

Q: Are there any pedestrians or cyclists on the road?
A: **No, there are no pedestrians** or cyclists visible on the road in the current view.

Q: What is the condition of the road surface?
A: The road surface appears to be in good condition, with **clear lane markings and no visible potholes or debris**.

Q: Are there any indications of construction work or road maintenance?
A: The presence of the **construction vehicle and the fenced-off area to the left** suggest **ongoing construction or maintenance work**.

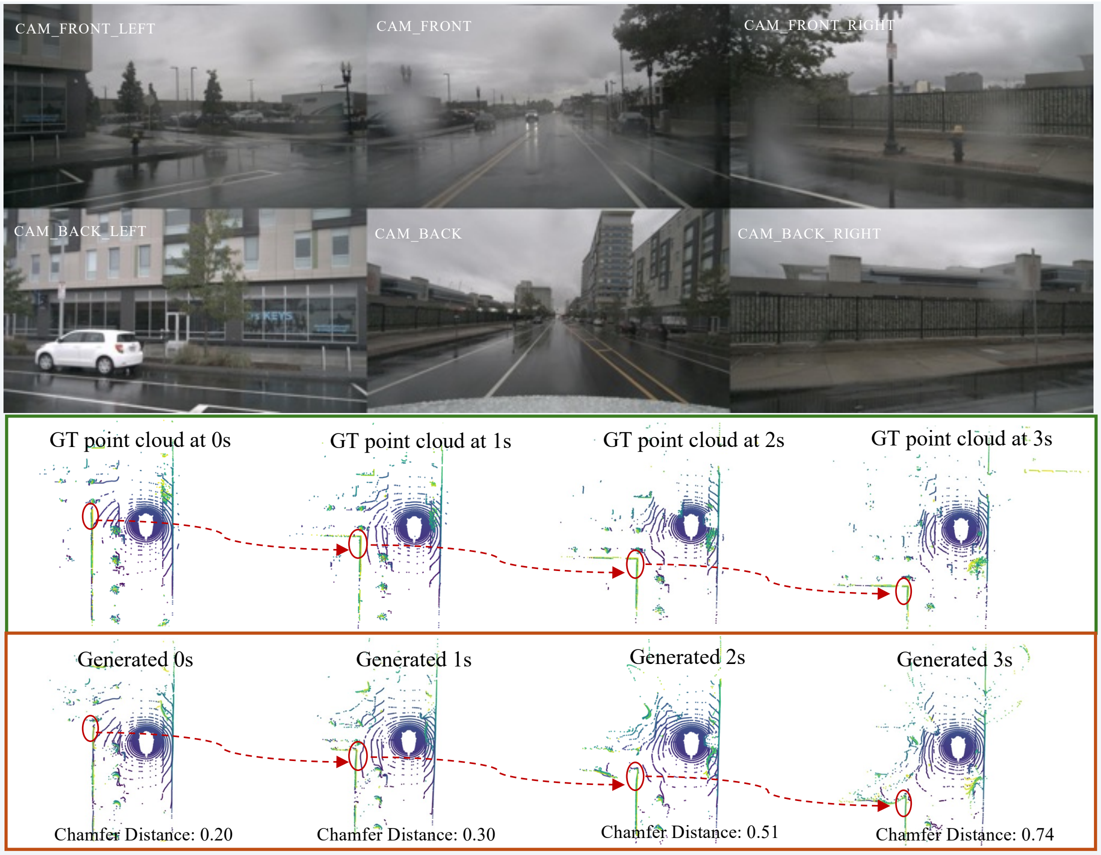
Q: How would you describe the driving conditions based on the weather?
A: The driving conditions seem to be favorable, with **clear weather and daylight** providing good visibility.

Q: Based on the surroundings, what kind of area is the vehicle driving through?
A: The vehicle is driving through an urban area with **modern buildings** and well-maintained greenery, suggesting a well-developed and possibly a business or residential district.

Q: Considering the clear road ahead and the traffic conditions, what would be the best driving practice in this scenario?
A: The best driving practice would be to continue at a moderate speed, maintain a safe following distance from the vehicle ahead, and stay vigilant for any potential movements from **the construction vehicle on the left**.

Q: What action should be taken in the current driving scenario?
A: In this situation, the vehicle should continue at a moderate speed and maintain the current lane.\n- Proceed straight, following the road markings and keeping a safe distance from the vehicles ahead.\n- Be aware of potential movements from the construction vehicle on the left, in case it starts to move or workers step onto the road.\n- Regularly check the rearview mirror to stay informed about the traffic situation behind, even though there's no immediate pressure from following vehicles.\n- Since the road ahead is clear, **focus on any unexpected obstacles that may appear from the sides of the road or from the construction area**.

(a)



Q:* *Can you provide a summary of the current driving scenario based on the input images?
A: The vehicle is positioned on a **multi-lane road** with clear lane markings, driving straight ahead against a backdrop of urban buildings under an overcast sky, suggesting it might be a cloudy day or early evening. The **road appears wet**, indicating recent rain, which could affect driving conditions. To the left, **a parking lot with several cars** is visible, while on the right, **a pedestrian sidewalk** is bordered by **a fence**, beyond which lies a large building complex. The road ahead is relatively clear, with a few vehicles in the distance and no immediate obstructions. In the rear view, the road is also clear with no vehicles in close proximity, and the surroundings are consistent with the front view, **featuring urban infrastructure and wet road conditions**.

Q: What kind of buildings are visible on the left side of the road?
A: On the left side of the road, there are **multi-story urban buildings** that appear to be commercial or residential.

Q: Is there any on-street parking visible?
A: Yes, there is **a parking lot visible to the left of the vehicle**, with several cars parked.

Q: Are there any visible traffic signs or signals?
A: **No traffic signs or signals** are visible from the current perspective.

Q: Can you see any pedestrians in the vicinity?
A: **No pedestrians** are visible in the immediate vicinity of the vehicle.

Q: How would you describe the weather conditions?
A: The weather is overcast, and **the road is wet, suggesting recent rainfall**.

Q: Is there any visible movement behind the vehicle?
A: No, there is **no visible movement behind** the vehicle; the road appears clear.

Q: What type of road markings can you see on the road ahead?
A: There are **white lane** markings on the road ahead, indicating the separation of lanes and the direction of traffic flow.

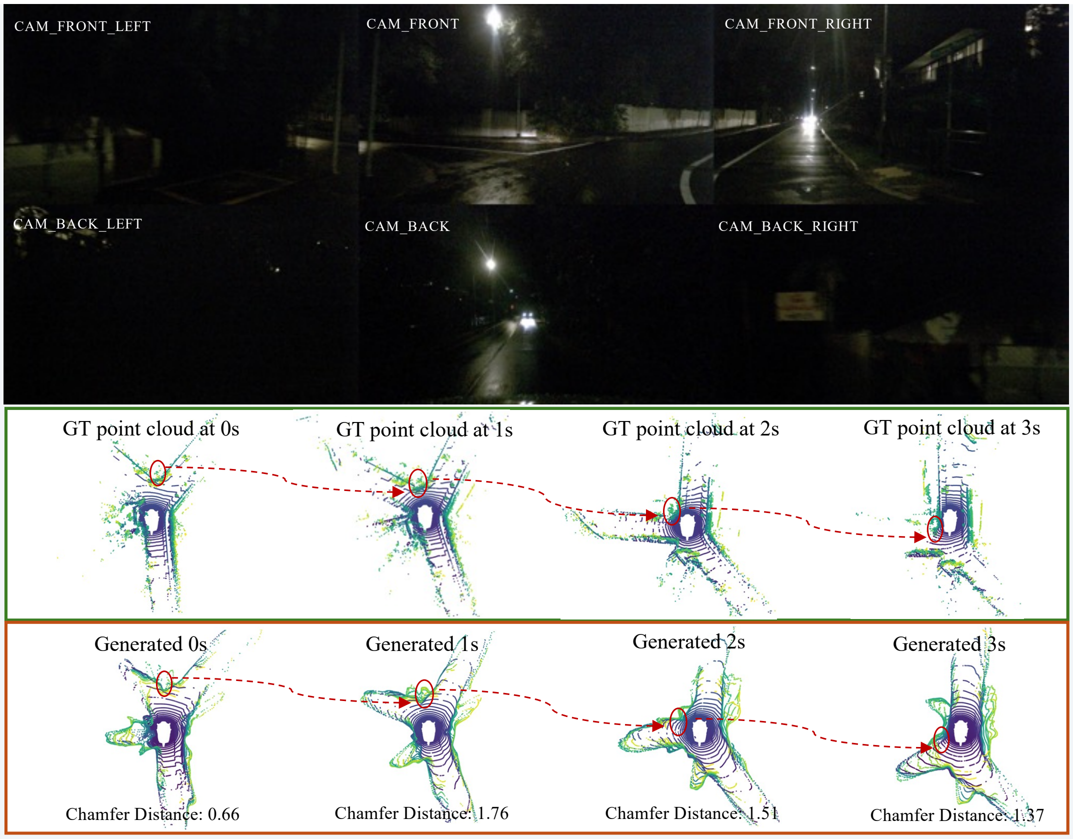
Q: Considering the wet road conditions, what driving behavior should be adopted?
A: The driver should maintain a moderate speed, increase following distance, and stay vigilant for pedestrians and other potential hazards due to the reduced traction on wet roads.

Q: Are there any indications of a residential area nearby?
A: Yes, the presence of a **large building complex to the left** suggests there may be a residential area nearby.

Q: What does the presence of a fence on the right-rear indicate?
A: The **fence on the right-rear** indicates a boundary for the public or private property adjacent to the road.

Q: What action should be taken in the current driving scenario?
A: Given the current road and weather conditions, the vehicle should:\n- Maintain a moderate speed to account for the wet road surface, which could increase stopping distances.\n- Stay in the current lane, following the clear lane markings to ensure a safe distance from other vehicles.\n- Proceed straight, as there are no immediate turns or intersections visible.\n- Remain vigilant for any pedestrians that may unexpectedly enter the roadway, especially given the reduced visibility from the weather conditions.

(b)



Q: Can you provide a summary of the current driving scenario based on the input images?
A: The environment is enveloped in the **darkness of night**, with only **streetlights and the headlights of vehicles** providing illumination. Directly ahead, the road is **glistening with moisture**, suggesting recent rain, and it appears to be slick, requiring careful navigation. To the right, the **glow of tail lights** indicates the presence of other vehicles, possibly moving or parked along the side of the road. On the left, **the road curves, hinting at a right turn ahead**. The rear views show **no immediate signs of traffic**, but the low visibility necessitates heightened awareness of any potential vehicles or obstacles that may be obscured by the darkness.

Q: Are there any traffic lights or stop signs visible?
A: **No traffic lights or stop signs are visible** in the immediate vicinity, indicating the absence of immediate traffic control devices in this section of the road.

Q: Can you describe any road signs and their possible meanings?
A: There are a few road signs visible, such as a 'No Entry' sign for a side road which suggests that it is not a through street and other regulatory signs that may not be fully visible due to the darkness.

Q: What is the condition of the road surface, and how should it affect driving behavior?
A: The **road surface is wet**, likely from recent rain, which should lead to a more cautious driving approach with reduced speed to prevent skidding.

Q: Is there any other vehicle around you?
A: Yes, there are **headlights visible in the distance** from another vehicle, suggesting there are cars on the road, but there are no vehicles in close proximity to us.

Q: Considering the wet road conditions and the solitary oncoming vehicle, what precautions should the driver take?
A: The driver should maintain a lower speed to account for **longer stopping distances on wet roads**, use headlights to increase visibility, and be vigilant for the oncoming vehicle, ready to react if it enters our lane.

Q: What can be inferred about the area surrounding the road?
A: The presence of trees and sparse buildings, along with the lack of immediate traffic in the rear views, suggests that this is a suburban or semi-rural area with lower traffic density.

Q: Is there any immediate traffic behind the vehicle?
A: There is no immediate traffic visible directly behind the vehicle, as the rear views show a clear path with **no other vehicles in close proximity**.

Q: How should the driver execute a right turn in these conditions?
A: The driver should signal well in advance, **reduce speed due to the wet road**, and check mirrors and blind spots for unseen traffic or cyclists before making the right turn.

Q: What action should be taken in the current driving scenario?
A: Given the low-light conditions and wet road, the vehicle should proceed with caution, preparing for a right turn.\n- Ensure the headlights and taillights are on for visibility.\n- Slow down to account for the longer stopping distance on the wet road surface.\n- Check the right side mirror and blind spot for any cyclists or vehicles that may be approaching from behind.\n- Signal well in advance to inform any possible vehicles behind of the intention to turn right.

(c)

Figure S2. Qualitative results for future generation and scene understanding. From top to bottom, each sub-figure displays the multi-view input of the current scene, the ground truth scene evolution, the generated scene evolution, and the scene understanding result.

large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 2

[12] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024. 2

[13] Zongchuang Zhao, Haoyu Fu, Dingkang Liang, Xin Zhou, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Extending large vision-language model for diverse interactive tasks in autonomous driving. *arXiv:2505.08725*, 2025. 1