Supplementary Material for 'Don't Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving'

Ziying Song[1,2,†], Caiyan Jia[1,2,⋆], Lin Liu [1,2], Hongyu Pan[3], Yongchang Zhang[3], Junming Wang[3,7], Xingyu Zhang[3], Shaoqing Xu[4], Lei Yang[5], Yadan Luo[6,⋆]

[1]School of Computer Science and Technology, Beijing Jiaotong University
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence
[3]Horizon Robotics [4]University of Macau [5]THU [6]The University of Queensland [7]HKU

{songziying, cyjia}@bjtu.edu.cn, y.luo@uq.edu.au.

A. Appendix

This supplementary material provides additional descriptions of the proposed MomAD framework, including the following supplementary material:

- **Appendix A.1:** Summary of contributions.
- **Appendix A.2:** The details of Turning-nuScenes dataset.
- **Appendix A.3:** Implementation details.
- **Appendix A.4:** More planning results.
- **Appendix A.5:** Detailed Result Analysis on Robustness.
- **Appendix A.6:** More visualizations of planning results.

A.1. Contributions

Our contributions are summarized below.

1) **MomAD Framework.** We propose MomAD, an end-to-end autonomous driving framework that employs momentum planning. Momentum planning leverages trajectory and perception momentum to enhance current planning through historical guidance, overcoming temporal inconsistency. It addresses key challenges in planning stability and robustness for end-to-end autonomous driving systems.

2) **TTM and MPI.** We propose the Topological Trajectory Matching (TTM) module, which utilizes the Hausdorff Distance to align candidate trajectories with past paths, ensuring temporal coherence and reducing abrupt trajectory changes. Furthermore, we propose the Momentum Planning Interactor (MPI) module. By cross-referencing current and past trajectory data, this module expands the system's perceptual awareness over time, enhancing long-horizon prediction and reducing collision risks.

3) **New*** Turning-nuScenes Validation Dataset. We create the Turning-nuScenes val dataset, derived from the nuScenes full validation dataset. This new dataset focuses on turning scenarios, providing a specialized benchmark for evaluating the performance of autonomous driving systems in complex driving situations.

4) **New*** Trajectory Prediction Consistency (TPC) Metric. We introduce the TPC metric to quantitatively assess the consistency of trajectory predictions in existing end-to-end autonomous driving methods, addressing a critical gap in the evaluation of trajectory planning.

A.2. The Detail of Turning-nuScenes dataset

When turning, vehicles need to quickly and accurately adjust their direction, making turning scenarios particularly challenging for the model's ability to maintain stable planning. However, there is currently no dataset specifically designed for evaluating models in turning scenarios. Based on the nuScenes val dataset, we selectively extracted data involving the ego vehicle in turning situations from the validation set to create the Turning-nuScenes dataset.

1) **Preparation Work.** We extract the data information from the *val* dataset based on the annotations of NuScenes dataset. Specifically, we establish a correspondence between *sample_token* (the unique identifier of each sample) and *scene_token* (the unique identifier of each scene) grounded in the provided data annotation information as illustrated in formula 1. We also extracted the future trajectory T_{fut} of the ego vehicle for each sample in the validation dataset over the next three seconds.

$$dict_{sa}^{sc}[sample\_token] = scene\_token \tag{1}$$

2) **Sample Select.** Considered that the ego vehicle's driving direction aligns with the y-axis of the world coordinate system, significant changes in the x-coordinate will occur during turns. Thus, we assess potential future turns of the ego vehicle based on changes in its x-coordinate, recording the unique identifier of each sample (*sample_token*). The specific criteria for judgment are as outlined in the formula

† Intern of Horizon Robotics, ⋆ Corresponding author.

2,

$$\begin{cases} S_T & |T_{\text{fut}}[0] - T_{\text{fut}}[5]| \geq \varepsilon \\ S_S & |T_{\text{fut}}[0] - T_{\text{fut}}[5]| < \varepsilon \end{cases} \tag{2}$$

where S_T and S_S represent the states of the ego vehicle during turning and going straight, respectively. And ε represents the judgment threshold, with a default setting of 25. 3) **Generate Dataset.** After sample select, we obtained a series of sample_tokens associated with turning scenarios, denoted as $sample\_token_{select}$. Based on the mapping relationship $dict_{sa}^{sc}$ from *scene_token* to *sample_token*, we derive a series of driving scenarios involving the ego vehicle's turning maneuvers. The Turning-NuScene dataset comprises 17 scenes with 680 samples and includes diverse urban turning scenarios, such as intersections, T-junctions, roundabouts, traffic islands, and alleyway turns. The visualization of some data from Turning-nuScenes dataset is shown in Fig. 1.

A.3. Implementation Details

The training process of MomAD is divided into two stages following SparseDrive [7]. In stage 1, we train the sparse perception module, including 3D object detection, multi-object tracking, and online mapping, from scratch to learn sparse scene representations. In stage 2, we train the sparse perception, motion, and planning modules without freezing the weights of the sparse perception module. For MomAD, we use ResNet50 [2] as backbone network and the input image size is 256 $\times$ 704. For detection, the perception range is a circle with a radius of 55m. For online mapping, the perception range is 60m $\times$ 30m longitudinally and laterally. For motion and planning, the number of stored frames H in the instance memory queue is set to 3, and the number of modes K_m in motion is set to 6, accounting for six trajectory proposals. All experiments are conducted on 8 NVIDIA RTX 4090 24GB GPUs.

Stage-1 Overall Objectives. In alignment with SparseDrive [7] and VAD [5], MomAD does not enforce tracking constraints during the identity assignment process. As a result, we do not include a tracking loss in our framework. The loss function for the supervised process during the first phase is defined as follows,

$$\mathbf{L}_1 = \mathbf{L_D} + \mathbf{L_M}. \tag{3}$$

Stage-2 Overall Objectives. MomAD is trained utilizing the losses from all tasks, which include 3D object detection, multi-object tracking, online mapping, motion prediction, and planning. This training is conducted over a duration of 10 epochs, employing a total batch size of 48 and a learning rate of $3 \times e^{-4}$. The loss function for the supervised process during this stage is defined as follows,

$$\mathbf{L}_2 = \mathbf{L_D} + \mathbf{L_M} + \mathbf{L_{MP}}. \tag{4}$$

Detection Loss. The detection loss is formulated as a linear combination of the Focal Loss [6] for classification and the L1 Loss for box regression.

$$\mathbf{L_D} = \lambda_c \mathbf{L_{D}}_c + \lambda_r \mathbf{L_{D}}_r, \tag{5}$$

which λ_c and λ_r are set to 2 and 0.25, respectively.

Online Mapping Loss. In accordance with VAD [5] and SparseDrive [7], we define the online mapping loss as the following equation,

$$\mathbf{L_M} = \lambda_c \mathbf{L_{M}}_c + \lambda_r \mathbf{L_{M}}_r, \tag{6}$$

which λ_c and λ_r are set to 1 and 10, respectively.

Motion and Planning Loss. We compute the average displacement error (ADE) between the multi-modal outputs and the ground truth trajectory. The trajectory with the lowest ADE is designated as the positive sample, while the remaining trajectories are treated as negative samples. In addition, for the planning component, the ego state is also predicted. We employ Focal Loss for classification and L1 Loss for regression,

$$\mathbf{L_{MP}} = \lambda_c^m \mathbf{L_{MO}}_c + \lambda_r^m \mathbf{L_{MO}}_r + \lambda_c^p \mathbf{L_{p}}_c + \lambda_r^p \mathbf{L_{p}}_r + \lambda_s^p \mathbf{L_s} \tag{7}$$

which λ_c^m and λ_r^m are set to 0.2 and 0.2, λ_c^p, λ_r^p and λ_s^p are set to 0.5, 1.0 and 1.0, respectively.

A.4. More Planning Results

We have extended the results of Tables 2 and 3 in the main by including UniAD [4] and VAD [5] to provide additional experimental data. As shown in Tables 1 and 2, our conclusion is consistent with those presented in the main text: end-to-end autonomous driving methods represented by UniAD [4], VAD [5], and SparseDrive [7] suffer challenges in turning scenarios. Our TPC metric demonstrates issues of robustness in temporal consistency, as these methods enable seamless integration of perception and planning but often rely on one-shot trajectory prediction, which may lead to unstable control and vulnerability to occlusions in single-frame perception. Overall, our proposed MomAD addresses key challenges in planning stability and robustness for end-to-end autonomous driving systems.

A.5. Detailed Result Analysis on Robustness

As shown in Table 3, we furthur evaluated MomAD on **nuScenes-C** [1], which benchmarks robustness against diverse corruptions including extreme weathers. Our MomAD consistently outperforms SparseDrive across all tasks, by **22.9%** (*detection*), **27.1%** (*tracking*), **25.1%** (*mapping*), **24.2%** (*motion*), and **40.0%** (planning) on average. These results highlight the robustness of MomAD against various noise perturbations.

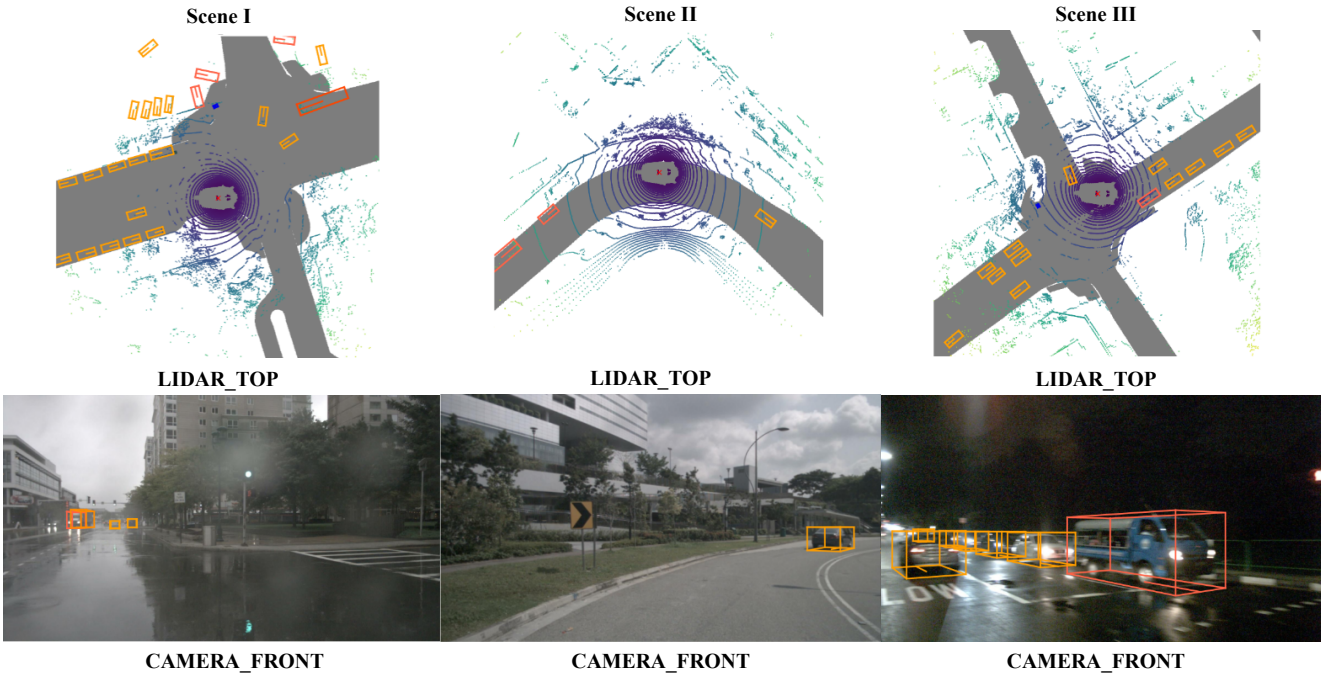

Figure 1. Visualization of turning scenarios in the **Turning-nuScenes dataset**. "LIDAR_TOP" represents the visualization of the corresponding scene from BEV. While "CAMERA_FRONT" refers to the images captured by the front camera of the ego vehicle in the respective scene.

Table 1. Planning results on the Turning-nuScenes validation dataset. UniAD [4] and VAD [5] are SOTA end-to-end deterministic planning methods. SparseDrive [7] is a SOTA end-to-end multi-modal trajectory planning method. Red indicates improvement. We follow the ST-P3 [3] evaluation metric.

Method	L2 (m) ↓				Col. Rate (%) ↓				TPC (m) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD [4]	0.52	0.88	1.64	1.01	0.16	0.51	1.41	0.69	0.47	0.81	1.58	0.95
VAD [5]	0.48	0.80	1.55	0.94	0.07	0.41	1.20	0.56	0.38	0.78	1.51	0.89
SparseDrive [7]	0.35	0.77	1.46	0.86	0.04	0.17	0.98	0.40	0.34	0.70	1.33	0.79
MomAD (Ours)	**0.33**-0.02	**0.70**-0.07	**1.24**-0.22	**0.76**-0.10	0.03-0.01	**0.13**-0.04	**0.79**-0.19	**0.32**-0.08	0.32-0.02	**0.54**-0.16	**1.05**-0.28	**0.63**-0.16

Table 2. Long trajectory planning results on the nuScenes and Turning-nuScenes validation sets. We train models for 10 epochs for 6s-horizon prediction. T-nuScenes indicates the challenging Turning-nuScenes. We follow the ST-P3 [3] evaluation metric.

Split	Method	L2 (m) ↓			Col. Rate (%) ↓			TPC (m) ↓		
		4s	5s	6s	4s	5s	6s	4s	5s	6s
nuScenes	UniAD [4]	1.91	2.57	3.21	0.91	1.66	2.51	1.49	1.81	2.41
	VAD [5]	1.82	2.23	3.01	0.89	1.71	2.41	1.55	1.73	2.17
	SparseDrive [7]	1.75	2.32	2.95	0.87	1.54	2.33	1.33	1.66	1.99
	MomAD	1.67	1.98	2.45	0.83	1.43	2.13	1.19	1.45	1.61
		-0.09	*-0.34*	*-0.50*	*-0.04*	*-0.11*	*-0.20*	*-0.14*	*-0.21*	*-0.38*
T-nuScenes	UniAD [4]	2.45	2.98	3.76	1.21	1.99	3.25	1.81	2.75	3.42
	VAD [5]	2.27	2.87	3.46	1.08	1.86	2.81	1.68	2.56	3.21
	SparseDrive [7]	2.07	2.71	3.36	0.91	1.71	2.57	1.54	2.31	2.90
	MomAD	1.80	2.07	2.51	0.85	1.57	2.31	1.37	1.58	1.93
		-0.27	*-0.64*	*-0.85*	*-0.06*	*-0.14*	*-0.26*	*-0.17*	*-0.73*	*-0.97*

Table 3. Robustness analysis on nuScenes-C [1].

Scene	Method	Detection		Tracking	Mapping	Motion		Planning		
		mAP ↑	NDS ↑	AMOTA ↑	mAP ↑	mADE ↓	L2 ↓	Col. ↓	TPC ↓	
Clean	SparseDrive	0.418	0.525	0.386	55.1	0.62	0.61	0.08	0.57	
	MomAD	**0.423**	**0.531**	**0.391**	**55.9**	**0.61**	**0.60**	**0.09**	**0.54**	
Snow	SparseDrive	0.140	0.161	0.133	22.3	0.95	0.85	0.30	0.79	
	MomAD	**0.172**	**0.195**	**0.169**	**27.9**	**0.72**	**0.71**	**0.18**	**0.66**	
Rain	SparseDrive	0.232	0.254	0.198	30.7	0.96	0.87	0.31	0.83	
	MomAD	**0.270**	**0.293**	**0.222**	**34.8**	**0.71**	**0.67**	**0.18**	**0.67**	
Fog	SparseDrive	0.294	0.312	0.260	41.2	0.93	0.84	0.36	0.80	
	MomAD	**0.348**	**0.356**	**0.299**	**43.2**	**0.68**	**0.64**	**0.19**	**0.61**	

A.6. More Qualitative Study of Planning Results

To better illustrate the exceptional planning capabilities of MomAD, we selected planning results from complex traffic scenarios for visualization, such as turning maneuvers and congested scenes. We provide three qualitative results: (1) planning for 3s trajectory prediction, (2) planning for

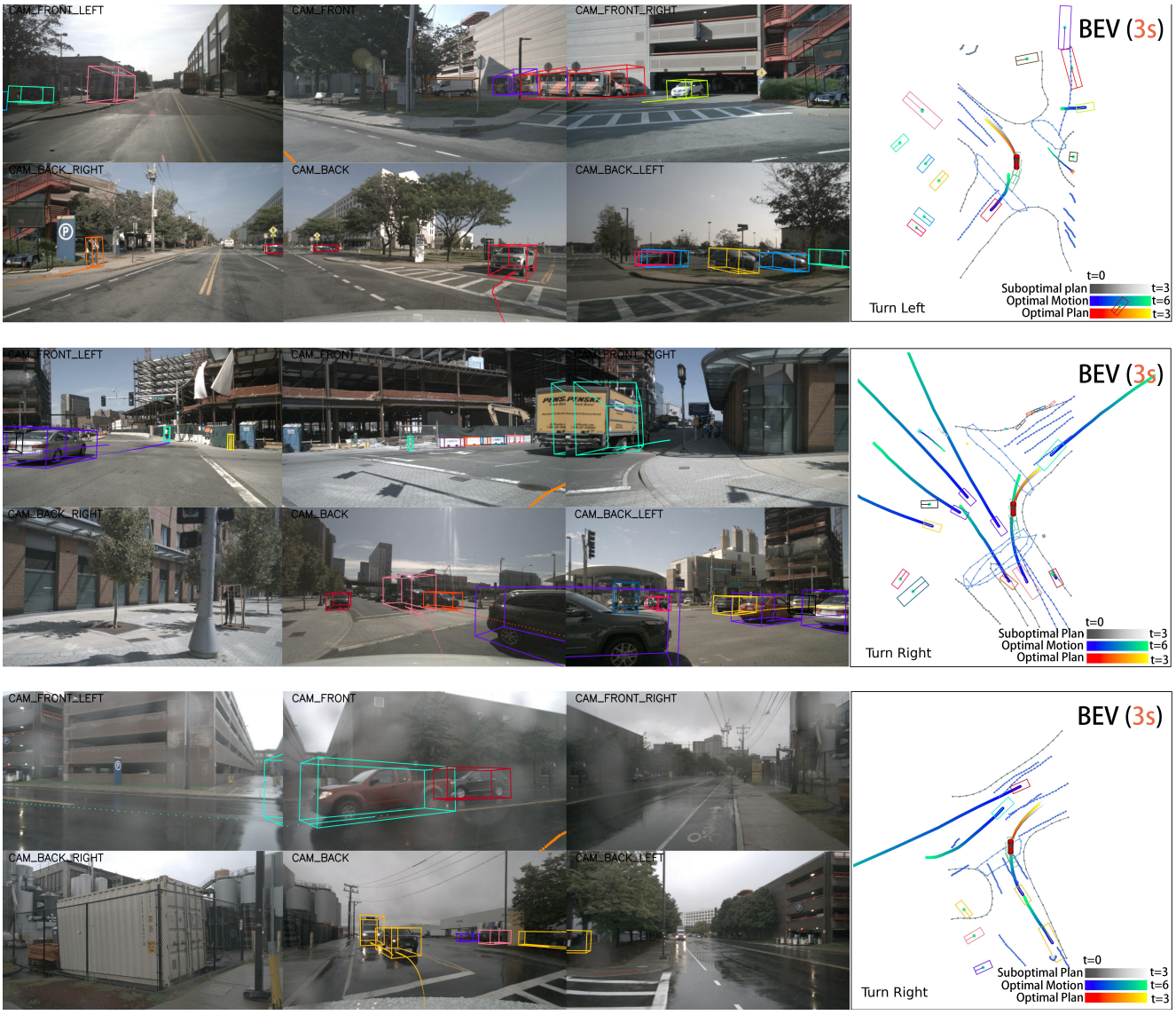

Figure 2. Visualization results (Planning for **3s** Trajectory Prediction). We visualize results for detection, online mapping, motion prediction, and planning. MomAD demonstrates stable and temporally consistent planning across various complex turning scenarios, especially in crowded environments. For motion prediction, we present the model's selected trajectory from multi-modal proposals, with each trajectory spanning a 6-second duration. For planning, the selected (optimal) trajectory is visualized in **red**, alongside two suboptimal (proposal) multi-modal trajectories in **gray**.

6 trajectory prediction, and (3) trajectory prediction across multiple frames.

(1) Planning for 3s Trajectory Prediction. Consistent with most end-to-end autonomous driving methods, we provide conventional 3-second prediction results, including the selected optimal trajectory and multi-modal proposal trajectory, as well as the optimal motion trajectory. As shown in Fig. 2, MomAD performs well across various turning scenarios, successfully executing large-angle turns without any collisions.

(2) Planning for 6s Trajectory Prediction. Unlike most end-to-end autonomous driving methods, we offer long-horizon trajectory predictions with a 6-second horizon. As depicted in Fig. 3, even under more challenging conditions, MomAD maintains superior planning performance. Specifically, the predicted trajectory remains smooth and consistent even over a long-horizon trajectory. This strong performance can be attributed to the proposed MomAD's effective use of historical trajectory data. By incorporating past trajectories, MomAD is able to predict and adapt to dynamic changes in the environment, ensuring smoother navigation and more accurate decision-making during turns.

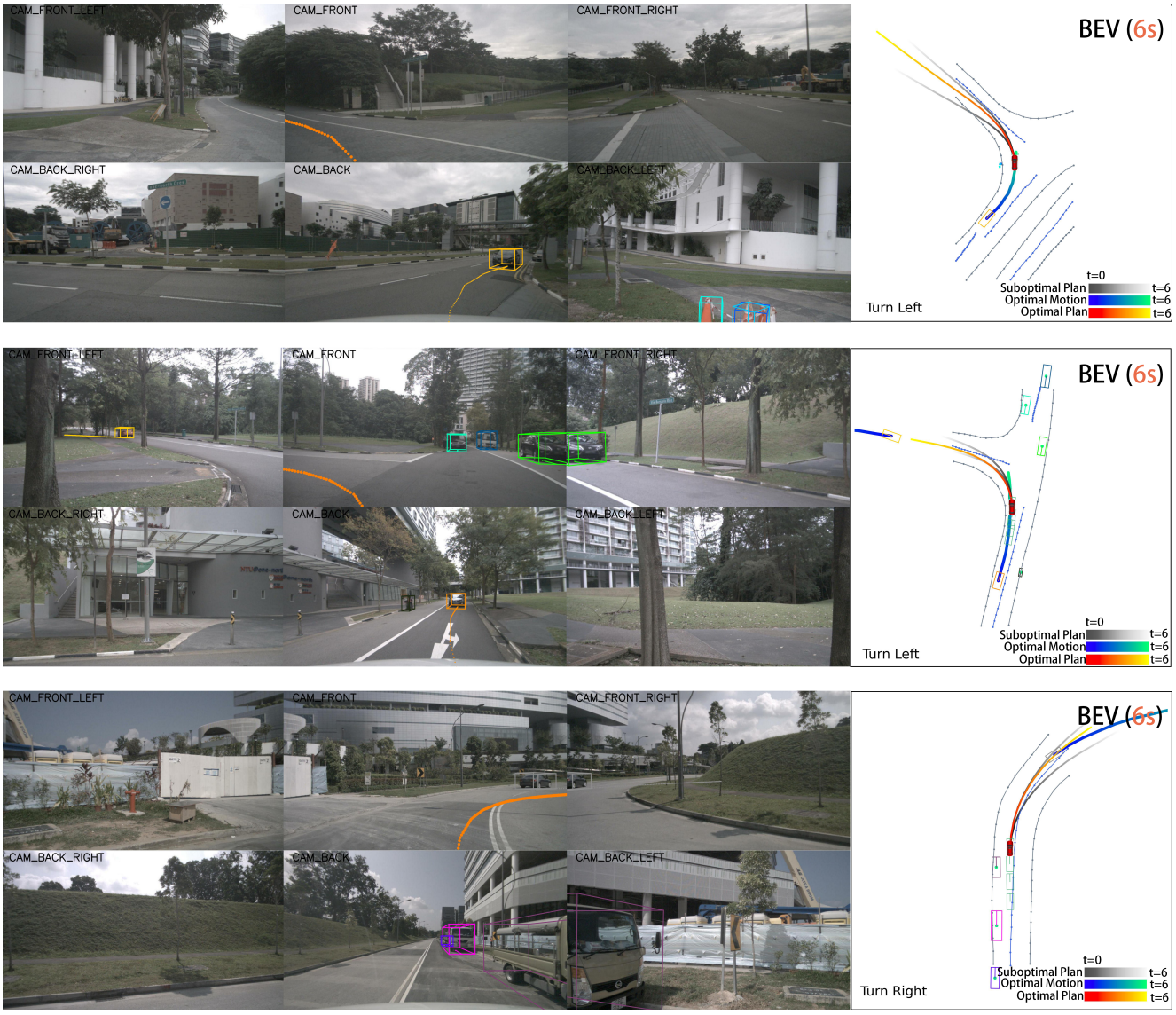

Figure 3. Visualization results (Planning for **6s** Trajectory Prediction). Long-horizon trajectories often face greater temporal consistency issues. We present 6-second trajectory prediction results to demonstrate how MomAD addresses these inconsistencies. Despite the increased challenge of long-horizon trajectories, MomAD continues to exhibit robust and stable performance. For motion prediction, we show the trajectory with the highest score from the model's output, each spanning 6 seconds. For planning, the selected (optimal) trajectory is visualized in **red**, accompanied by two suboptimal (proposal) multi-modal trajectories in **gray**.

(3) Trajectory Prediction across Multiple Frames. As shown in Figure 4, we present two multi-frame qualitative results to highlight the consistency and robustness of the proposed MomAD method. In the turning scenario, MomAD generates a smooth and accurate trajectory, demonstrating its ability to avoid oscillatory behavior during the planning process—a critical factor for ensuring driving safety. In conclusion, the visual results clearly illustrate the superior performance of MomAD in trajectory planning.

References

[1] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1022–1032, 2023. 2, 3

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

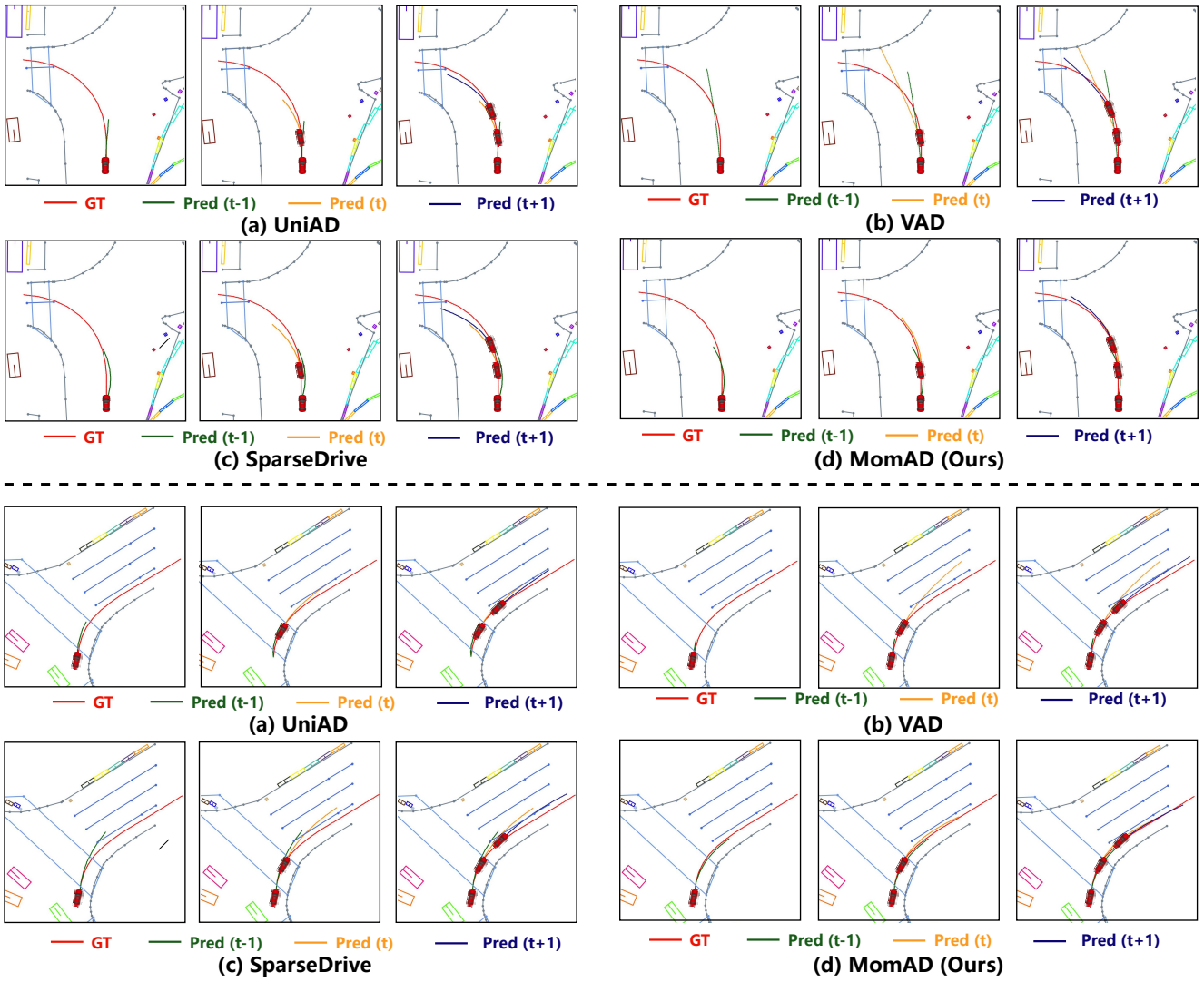

Figure 4. More visualization results of MomAD with SOTA methods across multiple frames.

[3] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 3

[4] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17853–17862, 2023. 2, 3

[5] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 2, 3

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. 2

[7] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 2, 3