Research Article

Don’t Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving

Ziying Song[1,2], Caiyan Jia[1,2], Yadan Luo[3]

1. School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China; 2. Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, China; 3. The University of Queensland, Brisbane, Australia

End-to-end autonomous driving frameworks 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. To address this, we propose the Momentum-Aware Driving (MomAD) framework, which introduces trajectory momentum and perception momentum to stabilize and refine trajectory predictions. MomAD comprises two core components: (1) Topological Trajectory Matching (TTM) employs Hausdorff Distance to select the optimal planning query that aligns with prior paths to ensure coherence; (2) Momentum Planning Interactor (MPI) cross-attends the selected planning query with historical queries to expand static and dynamic perception files. This enriched query, in turn, helps regenerate long-horizon trajectory and reduce collision risks. To mitigate noise arising from dynamic environments and detection errors, we introduce robust instance denoising during training, enabling the planning model to focus on critical signals and improve its robustness. We also propose a novel Trajectory Prediction Consistency (TPC) metric to quantitatively assess planning stability. Experiments on the nuScenes dataset demonstrate that MomAD achieves superior long-term consistency ($\geq 3s$) compared to SOTA methods. Moreover, evaluations on the curated Turning-nuScenes shows that MomAD reduces the collision rate by 26% and improves TPC by 0.97m (33.45%) over a 6s prediction horizon, while closed-loop on Bench2Drive demonstrates an up to 16.3% improvement in success rate. The source code is available at https://github.com/adept-thu/MomAD.

Corresponding authors: Caiyan Jia, cyjia@bjtu.edu.cn; Yadan Luo, y.luo@uq.edu.au

1. Introduction

Autonomous driving[1][2][3] has undergone a transformative shift from modular, manually crafted pipelines to a more integrated, end-to-end paradigm[4][5][6]. Unlike traditional approaches that handle tasks like detection, tracking, mapping, motion prediction, and planning in isolation, the end-to-end framework emphasizes seamless integration. By prioritizing planning, it strategically directs information from upstream perception modules, thereby enhancing robustness and reliability in dynamic driving environments.

Achieving high-quality planning in end-to-end frameworks hinges on accurately predicting the future trajectory prediction for the ego vehicle[4][5][7][8][9][10]. Such future prediction requires a long-horizon understanding of both static and dynamic environmental factors, including map elements and interactions with surrounding agents. For instance, UniAD[4] queries the ego context from detailed bird’s-eye-view (BEV) maps at each timestamp, while VAD[5] uses an ego query to retrieve surrounding context. The retrieved information then informs the planner, which predicts a deterministic trajectory for the vehicle, as illustrated in Figure 1 (a). Nevertheless, optimal trajectory prediction is inherently *stochastic* due to the unpredictability of other road users’ intentions, varying road conditions, and the ambiguity introduced by human driving behaviors. This stochastic nature complicates the regression target, making deterministic predictions suboptimal and even risk-prone, potentially leading to severe collisions. To mitigate these uncertainties, methods such as VADv2[7] and SparseDrive[8] leverage probabilistic modeling to capture the continuous planning action space, producing multi-modal trajectories that consider various possible behaviors of road agents, as shown in Figure 1 (b). While effective, these multi-modal approaches are typically *one-shot* and solely on the current perception frame. This limitation makes them susceptible to occlusion or loss of key visual cues, which can degrade multi-modal trajectory

 doi.org/10.32388/2VUIDZ

quality. Additionally, without temporal consistency, consecutive trajectories may lack coherence, causing unstable vehicle control and introducing undesirable directional shifts and oscillations.

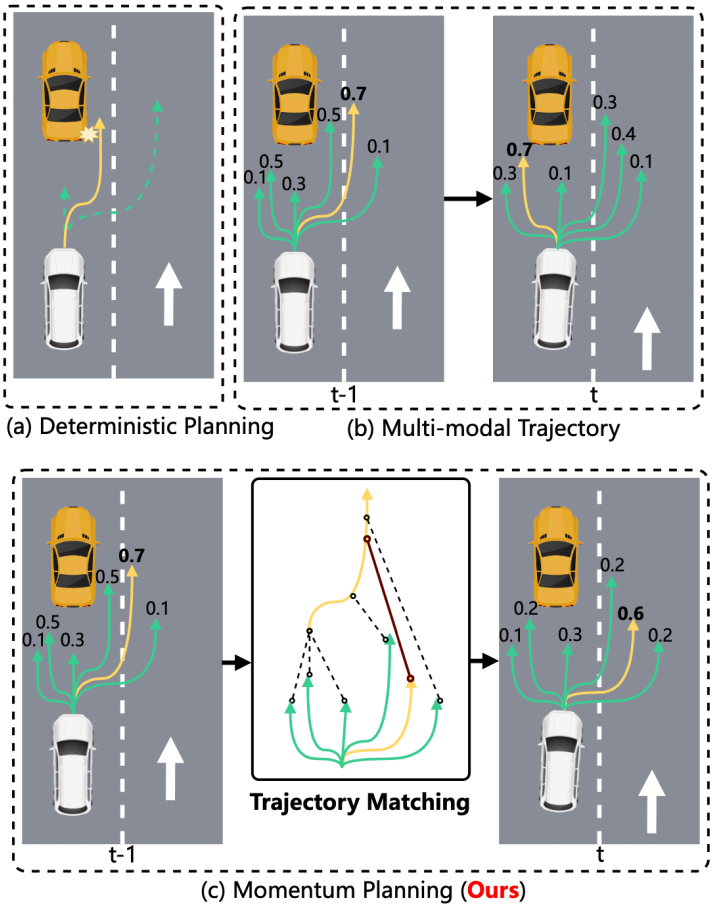

Figure 1. (a) **Deterministic Planning**[4][5][11][12] predicts deterministic trajectories, but lacks action diversity, posing safety risks. (b) **Multi-modal Trajectory Planning**[8][7][13] selects the highest-scoring trajectory among the multi-modal trajectories, yet fails to ensure stability and consistency, having risks in vehicle trembling. (c) **Momentum Planning** leverages the trajectory and perception momentum to enhance current planning through historical guidance to overcome temporal inconsistency.

To stabilize trajectory prediction, we draw inspiration from human driving behaviors and introduce the concept of momentum into autonomous driving. In physics, momentum reflects an object's tendency to maintain its velocity based on speed and direction[14]. Analogously, in driving, momentum captures the smooth, forward progression of movement informed by past trajectories and modulated by present conditions. As illustrated in Figure 1 (c), by explicitly integrating historical trajectories with current predictions, we aim to achieve smoother and more coherent planning outcomes. To this end, we propose an end-to-end Momentum-Aware Driving (MomAD) framework, which incorporates momentum awareness to deliver stable and responsive planning in driving scenarios. MomAD interprets momentum on two levels: (1) *trajectory momentum*: By aligning candidate multi-modal trajectory with prior predictions, abrupt shifts in ego vehicle's path can be minimized, ensuring consistent control and a more comfortable driving experience. (2) *perception momentum*: By aggregating historical context and attending to map elements and surroundings over time, the model broadens its perspective, capturing subtle agent intentions missed in single-frame observations. To implement these ideas, we introduce (1) *Topological Trajectory Matching (TTM)*: we first use this module to

minimize planning discrepancies across time steps by employing the Hausdorff Distance to identify multi-modal trajectory proposals that best align with past planning results. This approach ensures temporal coherence by preventing excessive deviation from previous trajectories. (2) *Momentum Planning Interactor (MPI)*: Since the selected trajectory may still be biased toward the current perception and sacrifice long-horizon considerations, we cross-attends the current best planning query with historical plan queries, which implicitly convey critical long-term ego-temporal, ego-agent, and ego-map information as key and value vectors. This interaction enriches the current query with long-horizon perception momentum, improving its context awareness. To enhance robustness against environmental noise and perception errors, we incorporate a Robust Instance Denoising Module in the perception stage. By introducing controlled perturbations during training, the model learns to denoise perception inputs, achieving resilience to dynamic changes and misdetections.

To evaluate the planning stability of MomAD, we propose a new Trajectory Prediction Consistency (TPC) metric to measure consistency between predicted and historical trajectories. Experiments demonstrate that MomAD can maintain long-term consistency ($\geq$ 3 seconds). Given that most scenes in nuScenes involve straight roads, which limit the assessment of temporal inconsistency, we curated a Turning-nuScenes validation set from turning scenarios within the nuScenes dataset to provide a more challenging evaluation, where our approach outperforms state-of-the-art end-to-end frameworks. For example, our MomAD reduces the collision rate by 26% and the TPC by 0.97m compared to SparseDrive[8] for a 6-second horizon prediction in the Turning-nuScenes validation set.

2. Related Work

End-to-end autonomous driving, which learns directly from raw sensor data to generate planning trajectories or driving commands, eliminates manual feature extraction[1][3]. End-to-end autonomous driving methods[15][16][17][18][19][20][21][13][4][5][22][11][23][24][25][26][27][28][29][30][31][32][33] have garnered increased attention. UniAD[4] effectively integrates information from various preceding tasks, including perception, prediction, and planning modules to assist in trajectory planning, and achieves significant performance improvements. VAD[5] models driving scenarios as fully vectorized representations and employs explicit instance-level planning constraints to enhance planning safety. However, they[4][5][12][25][24] adopt a deterministic approach to trajectory prediction, which fails to account for trajectory diversity and may introduce risks due to intermediate regression results. VADv2[7] proposes probabilistic planning to address the limitations of deterministic trajectory prediction, enabling multi-modal trajectory predictions. Building upon the multi-modal trajectory planning framework, SparseDrive[8] designs planning and motion prediction modules to achieve SOTA performance and efficiency on the nuScenes dataset. However, while multi-modal trajectory planning methods[7][8][13] have achieved SOTA performance, they overlook the temporal inconsistency caused by maximum score offsets. Regarding temporal inconsistency, existing methods only address temporal instance characteristics up to now, and they entirely overlook the issue of temporal planning consistency. In this work, we focus on this issue, aiming to address it using the concept of momentum planning.

3. Method

Framework Overview

Figure 2 presents an overview of the proposed MomAD system, which integrates sparse perception and momentum-aware planning. To capture key dynamic and static instances interacting with the ego vehicle, the sparse perception module builds upon the SparseDrive[8] to encode multi-view image features, which are aggregated into instance features $\mathbf{F}_t^{ins}$ for road agents and map elements at time step t. These features are obtained by sampling keypoints around the anchor boxes and polylines, feeding into the detection/tracking and online mapping blocks for accurate predictions. The core of MomAD is the joint motion and momentum-aware planning module, which comprises two main components: (1) Topological Trajectory Matching (Sec 3.1), which explicitly selects the candidate trajectory that best matches the prior path among all multi-modal trajectories to ensure temporal coherence; and (2) Momentum Planning Interaction (Sec 3.2), which expands the perceptive field by cross-attending the selected candidate trajectory's planning query with queries from the previous time step in the long-horizon query mixer. This approach provides a broader view of the surrounding environment and the intentions of other agents. The refined query is then processed

by the planning head to generate updated multi-modal trajectories. Since the planning module heavily relies on detection and map instance features, we introduce a robust instance denoising via perturbation module (Sec 3.3) within the sparse perception component during training. This ensures robustness by reducing sensitivity to noisy perception features, enhancing the stability of trajectory prediction and planning.

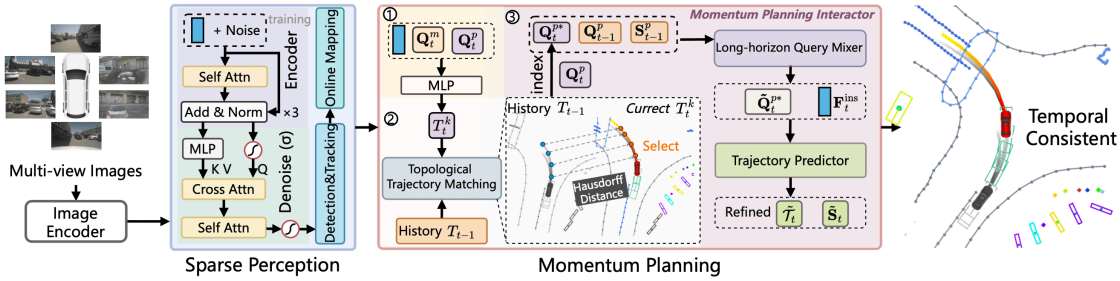

Figure 2. The overall architecture of MomAD. MomAD, as a multi-modal trajectory end-to-end autonomous driving method, first encodes multi-view images into feature maps, then learns a sparse scene representation through a robust instance denoising via perturbation module, and finally performs a momentum planning through Topological Trajectory Matching (TTM) module and Momentum Planning Interactor (MPI) module to accomplish planning tasks. Our approach addresses critical challenges of stability and robustness in dynamic driving conditions.

3.1. Topological Trajectory Matching (TTM)

The proposed TTM module is inspired by the continuity of human driving behavior, where the optimal trajectories T_t^* are influenced by the historical path T_{t-1} to maintain temporal consistency and stability. Let $\mathcal{T}_t = \{T_t^k\}_{k=1}^K$ denote a set of K multi-modal candidate trajectories generated at time step t, where each trajectory $T_t^k = \{(x_{t,i}^k, y_{t,i}^k)\}_{i=1}^{N_t}$ consists of N_t predicted waypoints. Typically, K is set to 6×3 to account for six trajectory proposals for each of three possible commands (left, right, and straight), and N_t is chosen as 6 or 12 representing 0.5s interval for a 3- or 6-second prediction horizon.

Trajectory Coordinate Transformation. Since the historical and current predicted trajectories are generated in the ego vehicle's coordinate system at different moments, it is essential to transform them into a common coordinate system for accurate matching. This transformation from moment t to $t-1$ is achieved as follows:

$$T_t^k \leftarrow \mathbf{R}_{t-1}^{-1}(T_t^k - \mathbf{\Gamma}_{t-1}), \forall k \in [K], \tag{1}$$

where $\mathbf{R}_{t-1}$ and $\mathbf{\Gamma}_{t-1}$ denote the rotation and translation matrix, respectively.

Trajectory Distance Measurement. Simple Euclidean distance is inadequate for capturing the global alignment of trajectories, as it only measures pointwise proximity and is highly sensitive to local variations. This limitation becomes especially apparent in complex scenarios such as turns or varying point densities, where close points may not represent the alignment of the entire trajectory path. To address these limitations, TTM employs the Hausdorff distance as a more robust metric for evaluating trajectory alignment. The Hausdorff distance captures both local and global trajectory structures by measuring the maximum deviation between two sets of points, effectively quantifying the worst-case alignment between the candidate and historical trajectories. For each candidate trajectory T_t^k, the Hausdorff distance to the historical trajectory T_{t-1} is computed as,

$$\begin{aligned} d_1 &= \sup_{p \in T_t^k} \inf_{h \in T_{t-1}} \|p - h\|, \\ d_2 &= \sup_{h \in T_{t-1}} \inf_{p \in T_t^k} \|h - p\|, \\ d_H(T_t^k, T_{t-1}) &= \max(d_1, d_2), \end{aligned} \tag{2}$$

where $p \in T_t^k$ and $h \in T_{t-1}$ represent waypoints in the candidate and historical trajectories, respectively. The Hausdorff distance considers the furthest point discrepancies between trajectories in both directions, ensuring that even minor global misalignments are captured. TTM then selects the trajectory $T_t^{k^*}$ that minimizes this distance:

doi.org/10.32388/2VUIDZ

$$k^* = \text{argmin}_{k\in[K]} d_H(T_t^k, T_{t-1}) \tag{3}$$

This selection enforces continuity, aligning with historical driving patterns and providing stable trajectory predictions that are less prone to sudden shifts.

3.2. Momentum Planning Interactor (MPI)

While TTM selects the most consistent trajectory $T_t^{k^*}$ based on historical alignment, $T_t^{k^*}$ is solely based on the current perception $\mathbf{F}_{ins}$, which may lack a comprehensive view of the environment and be sensitive to occlusions. Therefore, the MPI module, as illustrated in Figure 3, incorporates a *long-horizon query mixer* to enrich the selected planning query $\mathbf{Q}_t^{p*} \in \mathbb{R}^{D_q}$ of $T_t^{k^*}$ with historical planning query $\mathbf{Q}_{t-1}^{p} \in \mathbb{R}^{K\times D_q}$ and the associated planning scores $\mathbf{S}_{t-1}^{p} \in \mathbb{R}^{K}$, implicitly capturing a broader understanding of the surrounding context and other agents' intentions over time. Here, D_q represents the latent dimension of planning queries. This enriched planing query $\tilde{\mathbf{Q}}_t^{p*} \in \mathbb{R}^{D_q}$ will be combined with instance features to re-generate an improved trajectory $\tilde{\mathcal{T}}_t \in \mathbb{R}^{K\times N_t\times 2}$.

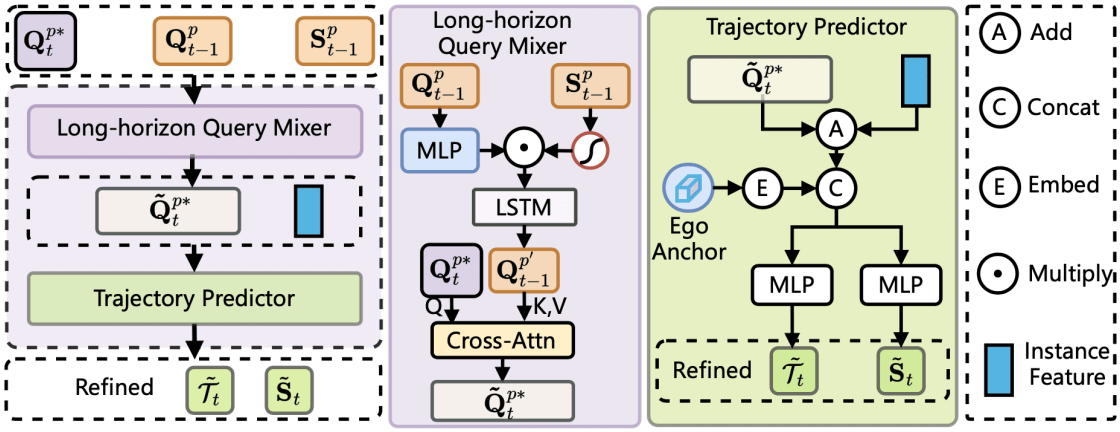

Figure 3. The illustration of **Momentum Planning Interactor (MPI)**. MPI cross-attends a selected planning query with historical queries to expand static and dynamic perception files, resulting in an enriched query that improves long-horizon trajectory generation and reduces collision risks.

Long-horizon Query Mixer. To achieve a robust perception of temporal momentum, the query mixer allows cross-attention between the selected candidate trajectory's planning query with multi-modal planning queries from the previous time step. The historical planning queries $\mathbf{Q}_{t-1}^{p}$ and associated scores $\mathbf{S}_{t-1}^{p}$ are combined through element-wise interaction and processed with an LSTM to simulate temporal evolution:

$$\mathbf{Q}_{t-1}^{p'} = \text{LSTM}(\sigma(\mathbf{S}_{t-1}^{p}) \circ \text{MLP}(\mathbf{Q}_{t-1}^{p})), \tag{4}$$

where $\sigma(\cdot)$ indicates the sigmoid function, $\text{MLP} : \mathbb{R}^{D_q} \mapsto \mathbb{R}^{D_q}$ the linear transform and $\circ$ the element-wise product. The $LSTM$ processes this interaction, producing a surrogate multi-modal query $\mathbf{Q}_{t-1}^{p'} \in \mathbb{R}^{K\times D_q}$ that captures the temporal evolution of planning queries. To aggregate historical information, the current planning query $\mathbf{Q}_t^{p*}$ is used as a query in a cross-attention module. The result $\tilde{\mathbf{Q}}_t^{p*}$ incorporates long-term spatiotemporal context, which is further combined with the planning instance features and the encoded ego-anchor position information to inform the subsequent trajectory predictor:

$$\begin{aligned} \tilde{\mathbf{Q}}_t^{p*} &= \text{Attention}(\mathbf{Q}_t^{p*}, \mathbf{Q}_{t-1}^{p'}, \mathbf{Q}_{t-1}^{p'}). \\ \tilde{\mathcal{T}}_t, \tilde{\mathbf{S}}_t &= \text{PlanHead}(\tilde{\mathbf{Q}}_t^{p*}, \mathbf{F}_t^{ins}), \end{aligned} \tag{5}$$

The PlanHead module then generates refined $\tilde{\mathcal{T}}_t, \tilde{\mathbf{S}}_t$. The best trajectory T_t^* is then selected based on the highest scores among multi-modal outputs. Importantly, while the best trajectory is chosen based on the multi-modal trajectory scores, unlike previous selections, the current multi-modal trajectories now fully consider the temporal consistency. This approach provides a stable, temporally-aware planning solution that is robust to occlusions and noise, significantly improving trajectory stability and control in complex driving environments.

3.3. Robust Instance Denoising via Perturbation

Our trajectory prediction and refinement rely heavily on the instance features $\mathbf{F}_t^{ins}$ of road agents and map elements provided by the sparse perception module. However, due to detector instability and the dynamically changing map, these instance features may be noisy, potentially introducing errors in downstream planning. To enhance the stability of planning against such noisy inputs, we introduce controlled noise perturbations during training and employ a lightweight encoder-decoder transformer block (see Figure 2) to learn effective denoising. This approach enables the model to distinguish between essential and extraneous features, reducing the impact of perception noise on trajectory predictions. During test-time inference, this denoising capability allows the trajectory predictor to be resilient to fluctuations in instance features. As a result, the model can produce smoother, more stable trajectories even in challenging scenarios with occlusions, temporary obstacles, or misdetections.

4. Experiments

4.1. Experimental Setup

Datasets. We conducted extensive experiments on the widely adopted nuScenes dataset[34] to evaluate tasks including detection, online mapping and planning in an open-loop setting. The nuScenes dataset comprises 1,000 driving scenes, with 700 and 150 sequences allocated for training and validation. Each scene spans around 20 seconds and contains roughly 40 key-frames annotated at 2Hz, where each sample includes six images captured by surrounding cameras covering 360° FOV horizontally and point clouds collected by both LiDAR and radar sensors. Since most planning tasks in nuScenes focus predominantly on go-straight commands, we curate a challenging subset of turning scenarios to form the **Turning-nuScenes** dataset, aimed at verifying the temporal consistency of predicted trajectories across time steps. Planning samples for turns are selected by setting the threshold between 3s and 0.5s of 'gt_ego_fut_trajs' to 25. The turning nuScenes validation dataset constitutes only one-tenth of the full nuScenes validation set, including 17 scenes with 680 samples. We use **Bench2Drive**[35], a closed-loop evaluation protocol under CARLA Leaderboard 2.0 for end-to-end autonomous driving. It provides an official training set, where we use the base set (1000 clips) for fair comparison with all the other baselines. We use the official 220 routes for evaluation.

Evaluation Metrics for Planning. For planning evaluation, we adopt the commonly used L2 Displacement Error (L2) and Collition Rate to assess planning performance. The calculation of the L2 error follows VAD[5] and the collision rate is aligned with SparseDrive[8]. However, the mainstream planning metrics cannot faithfully reveal the stability of predicted trajectories. Therefore, we introduce a novel metric, Trajectory Prediction Consistency (TPC), to measure the disparity between current predicted trajectories and historical predicted trajectories, allowing for a more comprehensive assessment of the consistency of trajectories. With coordinates transformed, the TPC between the current predicted trajectory T_t^{Pred} and historical one T_{t-1}^{Pred} is defined as,

$$TPC = \frac{1}{N_T} \sqrt{\sum_{i=1}^{N} \left(\left(T_t^{Pred} - T_{t-1}^{Pred} \right)^2 \cdot T_{GT}^{Mask} \right)}, \tag{6}$$

where N_T is the total number of GT trajectories in the validation set, and T_{GT}^{Mask} is the mask for trajectories exceeding the overlapped time period of two trajectories. Our TPC metric evaluates whether autonomous vehicles adhere to predicted trajectories, ensuring continuity across frames. Notably, the TPC metric provides a statistical perspective on the dataset-wide evaluation rather than at the individual sample level.

4.2. Main Results

To conduct a comprehensive comparison, we have undertaken an exhaustive analysis of two distinct metrics presented in Table 1, which are derived from UniAD[4] and VAD[5]. It is noteworthy that, aside from Table 1, all other tables rely on the VAD[5] evaluation metrics for their assessments.

$Method$	$Input$	$Backbone$	$L2(m)\downarrow$				$Col.Rate(\%)\downarrow$				$TPC(m)\downarrow$				$FPS\uparrow$
			1s	2s	3s	$Avg.$	1s	2s	3s	$Avg.$	1s	2s	3s	$Avg.$	
$UniAD^{\dagger}$[4]	$Camera$	$ResNet101$	0.48	0.96	1.65	1.03	**0.05**	0.17	0.71	0.31	0.45	0.89	1.54	0.96	1.8 ($A100$)
$VAD^{\dagger}$[5]	$Camera$	$ResNet50$	0.54	1.15	1.98	1.22	0.10	0.24	0.96	0.43	0.47	0.83	1.43	0.91	-
$SparseDrive^{\dagger *}$[8]	$Camera$	$ResNet50$	0.44	0.92	1.69	1.01	0.07	0.19	0.71	0.32	0.39	0.77	1.41	0.85	**9.0** ($RTX4090$)
$MomAD$ ($Ours$)†	$Camera$	$ResNet50$	**0.43**	**0.88**	**1.62**	**0.98**	0.06	**0.16**	**0.68**	**0.30**	**0.37**	**0.74**	**1.30**	**0.80**	7.8 ($RTX4090$)
$UniAD$[4]	$Camera$	$ResNet101$	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61	0.41	0.68	0.97	0.68	1.8 ($A100$)
VAD[5]	$Camera$	$ResNet50$	0.41	0.70	1.05	0.72	0.03	0.19	0.43	0.21	0.36	0.66	0.91	0.64	-
$SparseDrive$[8]	$Camera$	$ResNet50$	**0.29**	0.58	0.96	0.61	**0.01**	**0.05**	**0.18**	**0.08**	**0.30**	0.57	0.85	0.57	**9.0** ($RTX4090$)
$MomAD$ ($Ours$)	$Camera$	$ResNet50$	0.31	**0.57**	**0.91**	**0.60**	**0.01**	**0.05**	0.22	0.09	**0.30**	**0.53**	**0.78**	**0.54**	7.8 ($RTX4090$)

Table 1. Planning results on the $nuScenes$ validation dataset. † denotes evaluation protocol used in $UniAD$[4]. * denotes results reproduced with the official checkpoint. As Ref. [36] states, we deactivate the ego status information for a fair comparison.

4.2.1. Planning Results

NuScenes. As shown in Table 1, MomAD achieves an L2 error of 0.60m, a collision rate of 0.09%, and a TPC of 0.54m, respectively. Compared to SOTAs like UniAD[4], VAD[5] and SparseDrive[8], our method shows SOTA performance in planning results. It is worth noting that we have made significant improvements in TPC, which directly proves our effectiveness in timing consistency. It is worth noting that we achieved significant improvements in TPC at 0.30m, 0.53m, and 0.78m at 1s, 2s, and 3s on the nuScenes dataset, directly demonstrating our effectiveness in temporal consistency. Additionally, our MomAD is straightforward and achieves an FPS of 7.8, a slightly slower than SparseDrive. In summary, our MomAD effectively utilizes the smoothing advantage of Momentum and has a significant effect on improving temporal consistency.

Turning-nuScenes. As noted in[36], the nuScenes dataset features many straight routes, which limits the assessment of end-to-end methods. The simplicity of these paths can mask a model's true performance in complex scenarios. To address this, we evaluated our MomAD on the Turing-nuScenes validation set, as shown in Table 2. SparseDrive[8], a SOTA end-to-end method utilizing multi-modal trajectories, performs well across various scenarios but struggles with driving stability during turns. In contrast, MomAD exhibits superior consistency in trajectory predictions, as indicated by the TPC metric. Overall, MomAD not only delivers effective trajectory predictions under standard conditions but also maintains reliability amid dynamic changes and complex environments.

Method	$L2(m)\downarrow$				$Col.Rate(\%)\downarrow$				$TPC(m)\downarrow$			
	1s	2s	3s	$Avg.$	1s	2s	3s	$Avg.$	1s	2s	3s	$Avg.$
$SparseDrive$[8]	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. Planning results on the $Turning-nuScenes$ validation dataset. $SparseDrive$[8] is a SOTA end-to-end multi-modal trajectory planning method. *Italic* indicates improvement. We follow the VAD[5] evaluation metric.

Long Trajectory Prediction. Accurate long trajectory prediction is vital for enhancing the stability of autonomous driving, is useful to evaluate molels' ability to address temporal inconsistency issues in multi-modal trajectory planning. As shown in Table 3, we compared SparseDrive and MomAD in 4-6s long trajectory prediction on the nuScenes and Turning-nuScenes dataset, demonstrating a significant performance improvement. Specifically, in nuScenes dataset, compared with SparseDrive, MomAD experienced a decrease of 0.09m (5.14%), 0.34m (14.66%), and 0.50m (16.95%) in the 4s, 5s, and 6s of L2 error, a decrease of 0.04%, 0.11%, and 0.20% in the 4s, 5s, and 6s of collision rate, and a decrease of 0.14m (10.53%), 0.21m (12.65%), and 0.38m (19.10%) in the 4s, 5s, and 6s of TPC, respectively. Futhermore, in Turning-nuScenes dataset, compared with SparseDrive, MomAD experienced a decrease of 0.27m (13.04%), 0.64m (23.62%), and 0.85m (25.30%) in the 4s, 5s, and 6s of L2 error, a decrease of 0.06%, 0.14%, and 0.26% in the 4s, 5s, and 6s of collision rate, and a decrease of 0.17m (11.04%), 0.73m (31.60%), and 0.97m (32.45%) in the 4s, 5s, and 6s of TPC, respectively. We can observe that MomAD significantly improves trajectory predictions at farther distances, with a magnitude improvement at 6s. In summary, our MomAD has improved the performance of long trajectory predictions, which further proves that MomAD can effectively alleviate the problem of temporal inconsistency.

Split	Method	$L2(m)\downarrow$			$Col.Rate(\%)\downarrow$			$TPC(m)\downarrow$		
		4s	5s	6s	4s	5s	6s	4s	5s	6s
$nuScenes$	$SparseDrive$[8]	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*
$T-nuScenes$	$SparseDrive$[8]	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. 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 VAD[5] evaluation metric.

Bench2Drive. We have included evaluations on the challenging closed-loop results on Bench2Drive dataset, as shown in Table 4, w which covers 44 interactive scenes (e.g., cut-ins, overtaking, detours) and 220 routes across diverse weather conditions and locations. Our MomAD improves success rate by **16.3%** and **8.4%** over the VAD multi-modal variant and SparseDrive, and enhances the Comfortness score (trajectory smoothness) by **7.2%** and **5.3%**, demonstrating its effectiveness.

Method	Open-loop Metric	Closed-loop Metric			
	Avg. L2 ↓	DS ↑	SR (%) ↑	Effi ↑	Comf ↑
VAD	0.91	42.35	15.00	157.94	46.01
VAD_{mmt}	0.89	42.87	15.91	158.12	47.22
MomAD (Euclidean)	0.87	44.22	16.91	161.77	48.70
MomAD	**0.85**	**45.35**	**17.44**	**162.09**	**49.34**
SparseDrive*	0.87	44.54	16.71	170.21	48.63
MomAD (Euclidean)	0.84	46.12	17.45	173.35	50.98
MomAD	**0.82**	**47.91**	**18.11**	**174.91**	**51.20**

Table 4. Open-loop and Closed-loop results on Bench2Drive (V0.0.3) under base training set. 'mmt' refers multi-modal trajectory variant of VAD and * the re-implementation.

4.2.2. Perception and Motion Prediction Results

Sparse representation is efficient but suffers from instability issues caused by the variability of instance features. To address these issues, we have enhanced the instance features using the Encoder and Denoise (σ) module (denoted as ED) within the sparse perception framework, ensuring end-to-end stability for autonomous driving. As shown in Table 5, our MomAD perception module includes 3D object detection, multi-object tracking, and online mapping tasks. For 3D object detection, MomAD achieves 42.3% mAP and 53.1% NPS, improving the mAP by 0.5% and the NDS by 0.6% compared to the baseline SparseDrive[8]. For multi-object tracking, MomAD achieves an AMOTA of 39.1%, surpassing the baseline SparseDrive by 0.5%. For online mapping, compared to 55.1% mAP for the baseline SparseDrive, our MomAD achieves 55.9% mAP, improving the mAP by 0.8%. For motion prediction, our MomAD outperforms SparseDrive[8] and UniAD[4], achieves better motion prediction performance by considering the influence of an ego vehicle on other agents. In detail, Our MomAD achieves a 0.61m minADE, 0.98 minFDE, and 13.7% MR, and 0.499 EPA, respectively.

Method	3D Object Detection							Multi-Object Tracking				Online Mapping				Motion Prediction			
	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓	AMOTA ↑	AMOTP ↓	Recall ↑	IDS ↓	mAP ↑	AP_{ped} ↑	AP_d ↑	AP_b ↑	mADE ↓	mFDE ↓	MR ↓	EPA ↑
UniAD[4]	0.380	0.498	0.684	0.277	0.383	0.381	0.192	0.359	1.320	0.467	906	-	-	-	-	0.71	1.02	0.151	0.456
VAD†[5]	0.312	0.435	0.610	0.288	0.541	0.534	0.228	-	-	-	-	47.6	40.6	51.5	50.6	-	-	-	-
SparseDrive[8]	0.418	0.525	0.566	0.275	0.552	0.261	0.190	0.386	1.254	0.499	886	55.1	49.9	57.0	58.4	0.62	0.99	0.136	0.482
MomAD (Ours)	**0.423**	**0.531**	**0.561**	**0.269**	**0.549**	**0.258**	**0.188**	**0.391**	**1.243**	**0.509**	**853**	**55.9**	**50.7**	**58.1**	**58.9**	**0.61**	**0.98**	**0.137**	**0.499**

Table 5. Perception and motion results on the nuScenes validation dataset. † indicates the results are reproduced with the official checkpoint. AP_d denotes $AP_{divider}$. AP_b denotes $AP_{boundary}$. $mADE$ denotes $minADE$. $mFDE$ denotes $minFDE$.

4.3. Ablation Study

Roles of 'ED' module in Sparse Perception. Sparse representation end-to-end methods yield efficient computation but unstable metrics. Further details are available in the Appendix. As shown in Table 6, the Encoder and Denoise (σ) module (ED) within sparse perception enhances the instance features, significantly impacts the overall pipeline. By introducing Gaussian noise and employing techniques, the robustness of instance features is improved, particularly when training with Noisy at 0.1. Our findings suggest that a controlled level of noise can enhance the end-to-end capabilities of sparse methods during training, offering insights for the community on sparse end-to-end methods.

ED	MP	NS	Detection		Tracking	Online Mapping		Motion	Planning (Avg.)		
			$mAP\uparrow$	$NDS\uparrow$	$AMOTA\uparrow$	$mAP\uparrow$	$AP_{ped}\uparrow$	$mADE\downarrow$	$L2\downarrow$	$Col.\downarrow$	$TPC\downarrow$
		0.0	0.407	0.521	0.381	55.0	49.3	0.63	0.62	0.14	0.56
	✔	0.0	0.405	0.520	0.380	55.1	49.5	0.63	0.61	0.13	0.55
✔		0.1	0.420	0.530	0.390	55.8	50.5	0.58	0.61	0.12	**0.55**
✔	✔	0.0	0.417	0.528	0.386	55.4	50.6	0.63	0.61	0.11	**0.55**
✔	✔	0.05	0.421	0.529	0.388	55.6	**50.8**	0.62	0.61	0.11	**0.54**
✔	✔	0.1	**0.423**	**0.531**	**0.391**	**55.9**	50.7	**0.61**	**0.60**	**0.09**	**0.54**
✔	✔	0.2	0.418	0.520	0.388	54.4	49.2	0.63	0.62	0.18	0.58
✔	✔	0.3	0.412	0.518	0.383	54.0	48.8	0.65	0.64	0.22	0.61

Table 6. Ablation studies of the sparse perception module in MomAD on the nuScenes validation split. The Encoder and Denoise (σ) module is denoted as ED. MP represents Momentum planning. NS is the Gaussian noise factor controlling the noise level. Noise is applied during training only. We follow the VAD[5] evaluation metric.

Roles of 'MP' module in Planing. As Li et al.[36] have stated, most end-to-end autonomous driving methods perform poorly in turning scenarios. As shown in Table 7, to better evaluate the planning performance of end-to-end methods in turning scenarios, our MomAD is evaluated on the Turning-nuScenes validation dataset rather than only on the full nuScenes validation dataset. Specifically, under the premise of executing 'ED', at $t = 1$, providing a 0 value to the MP module does not improve performance. We have tried to change the MLP operation of the planning to a more complex operation, but it does not enhance the results. However, when $t = 2$, historical queries and results are used, L2 (Avg) reaches 0.76m, Col. (Avg) reaches 0.32 %, and TPC reaches 0.63m, which represents a significant improvement. In addition, when $t = 3$, more frames are fused, the improvement has actually decreased, which may be due to the uncertainty introduced by the departure of historical features, but there is still an overall improvement.

ED	MP	NS	t	$L2(m)\downarrow$			$Col.Rate(\%)\downarrow$			$TPC(m)\downarrow$		
				2s	**3s**	**Avg.**	**2s**	**3s**	**Avg.**	**2s**	**3s**	**Avg.**
				0.77	1.46	0.86	0.17	0.98	0.40	0.70	1.33	0.79
	✔		2	0.71	1.27	0.77	0.14	0.83	0.34	0.55	1.07	0.65
✔		0.1		0.77	1.45	0.86	0.18	0.97	0.39	0.69	1.33	0.79
✔	✔	0.1	1	0.78	1.48	0.88	0.18	0.98	0.39	0.70	1.35	0.81
✔	✔	0.1	2	**0.70**	**1.24**	**0.76**	**0.13**	**0.79**	**0.32**	**0.54**	**1.05**	**0.63**
✔	✔	0.1	3	0.72	1.27	0.78	0.14	0.84	0.35	0.56	1.09	0.66

Table 7. Impact of history frames in MomAD on the Turning-nuScenes validation set. t denotes the frame number, where 1 indicates the history is empty (represented by 0), 2 signifies that the historical result corresponds to the previous 1 frame, and 3 indicates that the historical result pertains to the previous 2 frames. We follow the VAD[5] evaluation metric.

Roles of different sub-modules in 'MP' module. As shown in Table 8, we conducted an in-depth analysis of the internal mechanism of momentum planning. We found that simply using the native 'Add' operation to regenerate the planning results can achieve a good improvement, with L2 (Avg.), Col. (Avg.), and TPC (Avg.) decreasing by 0.04m, 0.04%, and 0.12m, respectively. However, the 'Add' operation alone does not fully utilize historical features. Our Long-horizon Query Mixer has achieved the current optimal performance. Overall, historical results are very important for current outcomes, and their reasonable utilization can maximize the performance of end-to-end planning.

QM	Add	TP	$L2(m)\downarrow$			$Col.Rate(\%)\downarrow$			$TPC(m)\downarrow$		
			2s	**3s**	**Avg.**	**2s**	**3s**	**Avg.**	**2s**	**3s**	**Avg.**
			0.77	1.46	0.86	0.17	0.98	0.40	0.70	1.33	0.79
	✔	✔	0.76	1.38	0.82	0.14	0.88	0.36	0.62	1.21	0.67
✔		✔	**0.70**	**1.24**	**0.76**	**0.13**	**0.79**	**0.32**	**0.54**	**1.05**	**0.63**

Table 8. Ablation studies of the impact of the different modules in MP on the Turning-nuScenes validation dataset. QM denotes Long-horizon Query Mixer, TP denotes Trajectory Predictor. Add refers to the addition operation between the historical planning query $\mathbf{Q}_{t-1}^{p}$ and the selected planning query $\mathbf{Q}_{t}^{p*}$. We follow the VAD[5] evaluation metric.

4.4. Visualization

As shown in Figure 4, we showcase multi-frame qualitative comparisons of end-to-end solutions, including UniAD[4], VAD[5], SparseDrive[8], and the proposed MomAD. In a representative turning scenario, the MomAD approach demonstrates superior long-term awareness of surrounding vehicles, reducing the likelihood of collisions. Additionally, it generates smoother ego-vehicle trajectories (shown in yellow and blue) that closely align with the ground-truth trajectory (in red). This highlights its strong temporal consistency and lower TPC scores. Additional visualizations for various driving commands are provided in the Appendix.

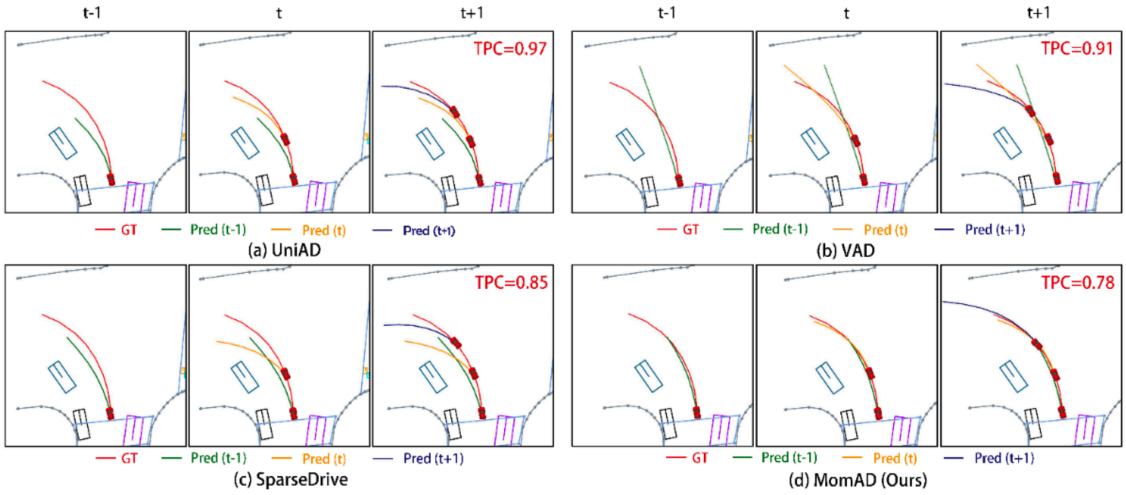

Figure 4. Visualization results of MomAD compared with UniAD, VAD and SparseDrive across multiple frames. The proposed MomAD achieves temporal consistency whichever from the predicted trajectory compared with ground truth (GT) or from the TPC metric.

5. Conclusion and Future Work

The proposed MomAD framework addresses key challenges in planning stability and robustness for end-to-end autonomous driving systems. By leveraging trajectory momentum and perception momentum, MomAD stabilizes trajectory predictions through Topological Trajectory Matching (TTM) and Momentum Planning Interactor (MPI), ensuring temporal coherence and enriching long-horizon context. Evaluations on nuScenes and the curated Turning-nuScenes validation set demonstrate its superior performance in reducing collision rates and improving trajectory consistency compared to state-of-the-art methods. While MomAD improves temporal consistency in long-horizon trajectory prediction, a gap remains due to mode collapse induced by the standard teacher-forcing approach to trajectory regression, limiting trajectory diversity. Future work will explore diffusion models and speculative decoding to enhance trajectory diversity while balancing efficiency.

Appendix

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

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

doi.org/10.32388/2VUIDZ

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.

$$\mathrm{dict}_{sa}^{sc}[sample\_token] = scene\_token \tag{1}$$

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 2,

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

   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 Figure A1.

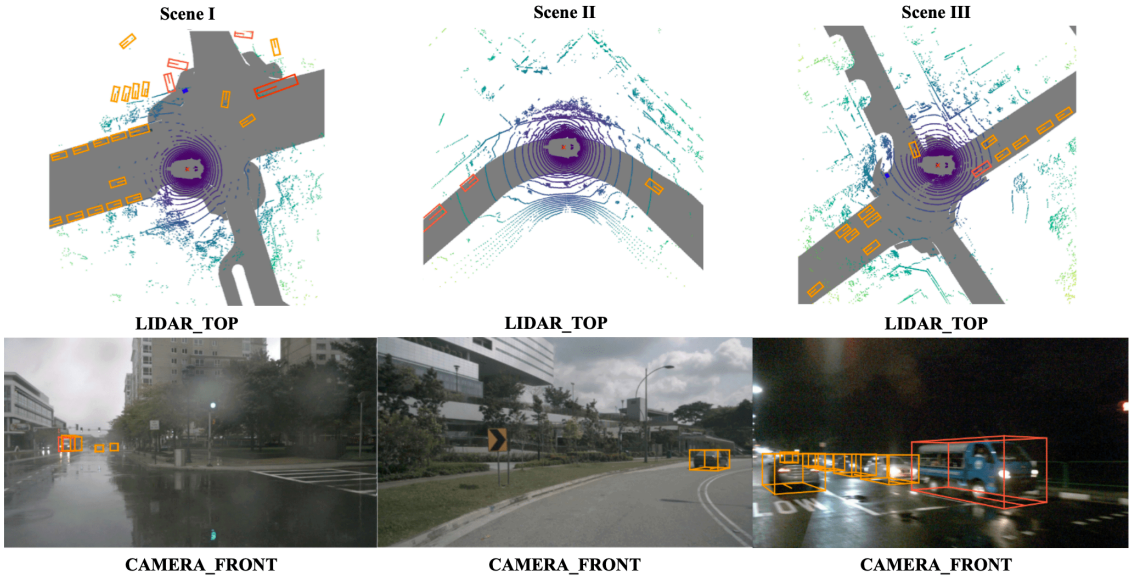

Figure A1. Visualization of turning scenarios in the Turning-nuScenes dataset. "LIDAR_TOP" represents the visualization of the corresponding scene from BEV. While "CAMERA_FRONT" refers to the images captured by the front camera of the ego vehicle in the respective scene.

A.3. Implementation Details

The training process of MomAD is divided into two stages following SparseDrive[8]. 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[37] as backbone network and the input image size is 256×704. 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[8] 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[38] 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[8], 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 A2 and A3 in the main by including UniAD[4] and VAD[5] to provide additional experimental data. As shown in Tables A1 and A2, 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[8] 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.

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[8]	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 A1. Planning results on the $Turning\text{-}nuScenes$ validation dataset. $UniAD$ [4] and VAD [5] are SOTA end-to-end deterministic planning methods. $SparseDrive$ [8] is a SOTA end-to-end multi-modal trajectory planning method. *Italic* indicates improvement. We follow the ST-P3[12] evaluation metric.

Split	Method	$L2(m)\downarrow$			$Col.\ Rate(\%)\downarrow$			$TPC(m)\downarrow$		
		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
	$SparseDrive$[8]	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*
$T\text{-}nuScenes$	$UniAD$[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
	$SparseDrive$[8]	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 A2. Long trajectory planning results on the $nuScenes$ and $Turning\text{-}nuScenes$ validation sets. We train models for 10 epochs for 6s-horizon prediction. $T\text{-}nuScenes$ indicates the challenging $Turning\text{-}nuScenes$. We follow the ST-P3[12] evaluation metric.

$Scene$	$Method$	$Detection$		$Tracking$	$Mapping$	$Motion$	$Planning$		
		$mAP\uparrow$	$NDS\uparrow$	$AMOTA\uparrow$	$mAP\uparrow$	$mADE\downarrow$	$L2\downarrow$	$Col.\downarrow$	$TPC\downarrow$
$Clean$	$SparseDrive$	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**

Table A3. Robustness analysis on $nuScenes-C$[39].

A.5. Detailed Result Analysis on Robustness

As shown in Table A3, we furthur evaluated MomAD on `nuScenes-C`[39], 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.

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 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 Figure A2, 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 Figure A3, 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.
3. **Trajectory Prediction across Multiple Frames.** As shown in Figure A4, 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.

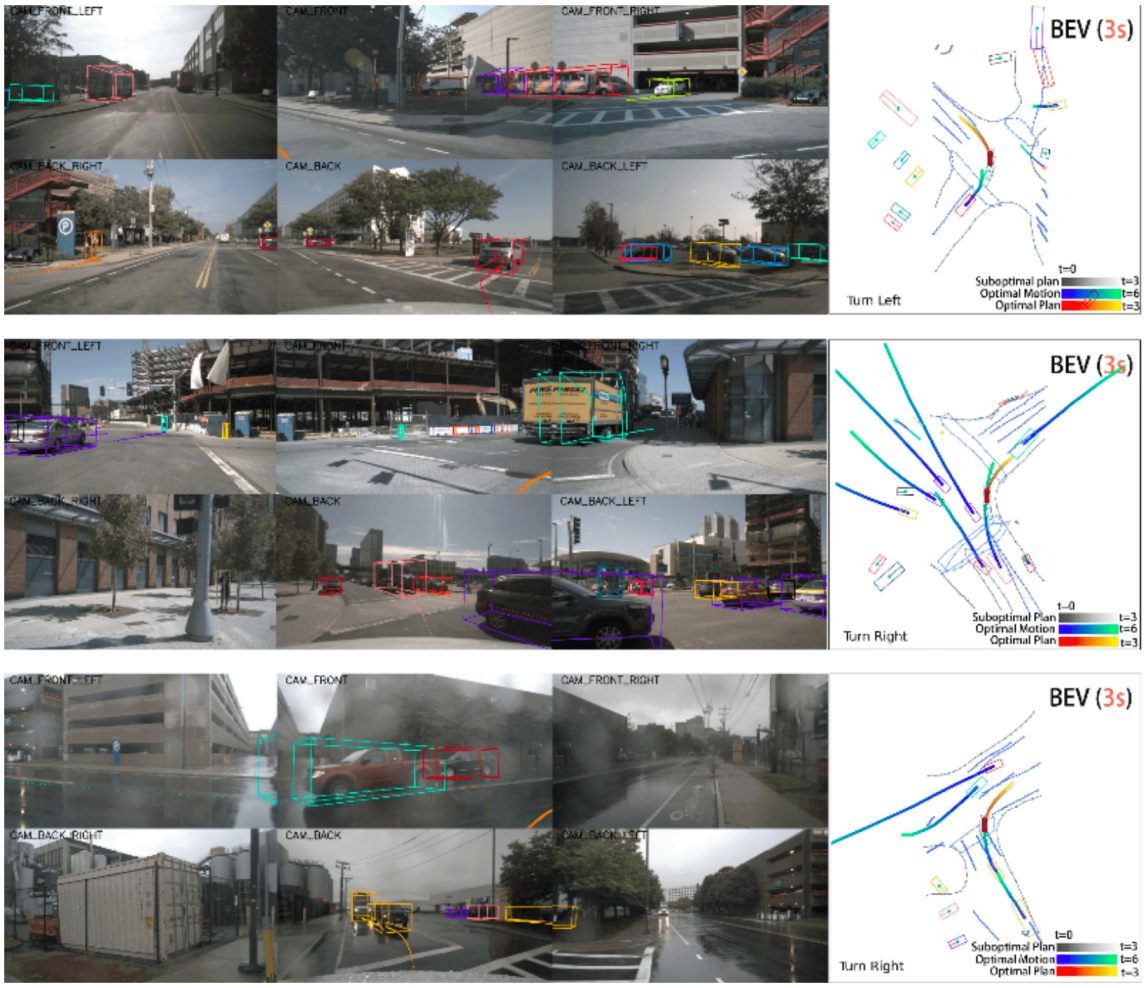

Figure A2. Visualization results (Planning for **3s** Trajectory Prediction). We visualize results for detection, online mapping, motion prediction, and planning. MomAD demonstrates stable and temporally consistent planning across various complex turning scenarios, especially in crowded environments. 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.

doi.org/10.32388/2VUIDZ

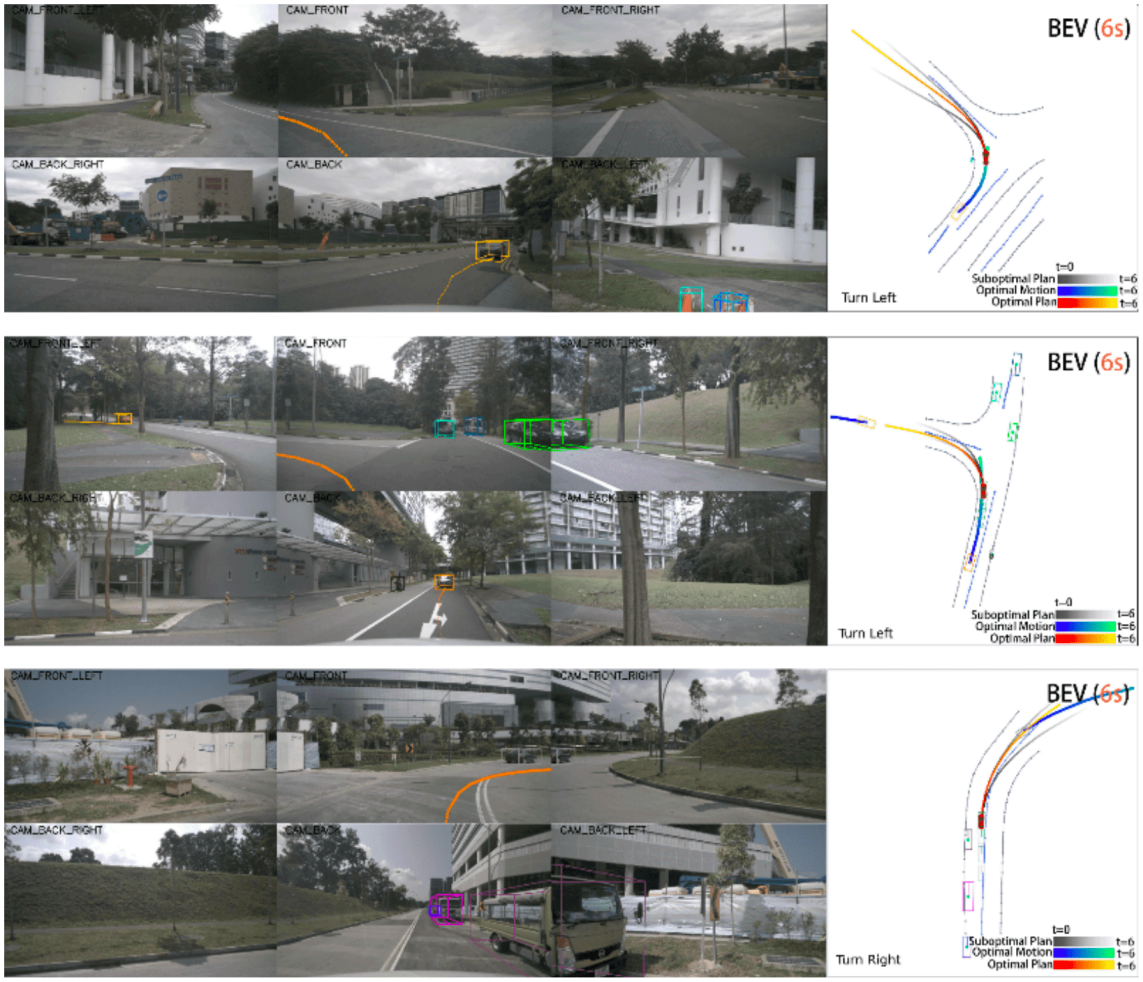

Figure A3. Visualization results (Planning for **6s** Trajectory Prediction). Long-horizon trajectories often face greater temporal consistency issues. We present 6-second trajectory prediction results to demonstrate how MomAD addresses these inconsistencies. Despite the increased challenge of long-horizon 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.

doi.org/10.32388/2VUIDZ

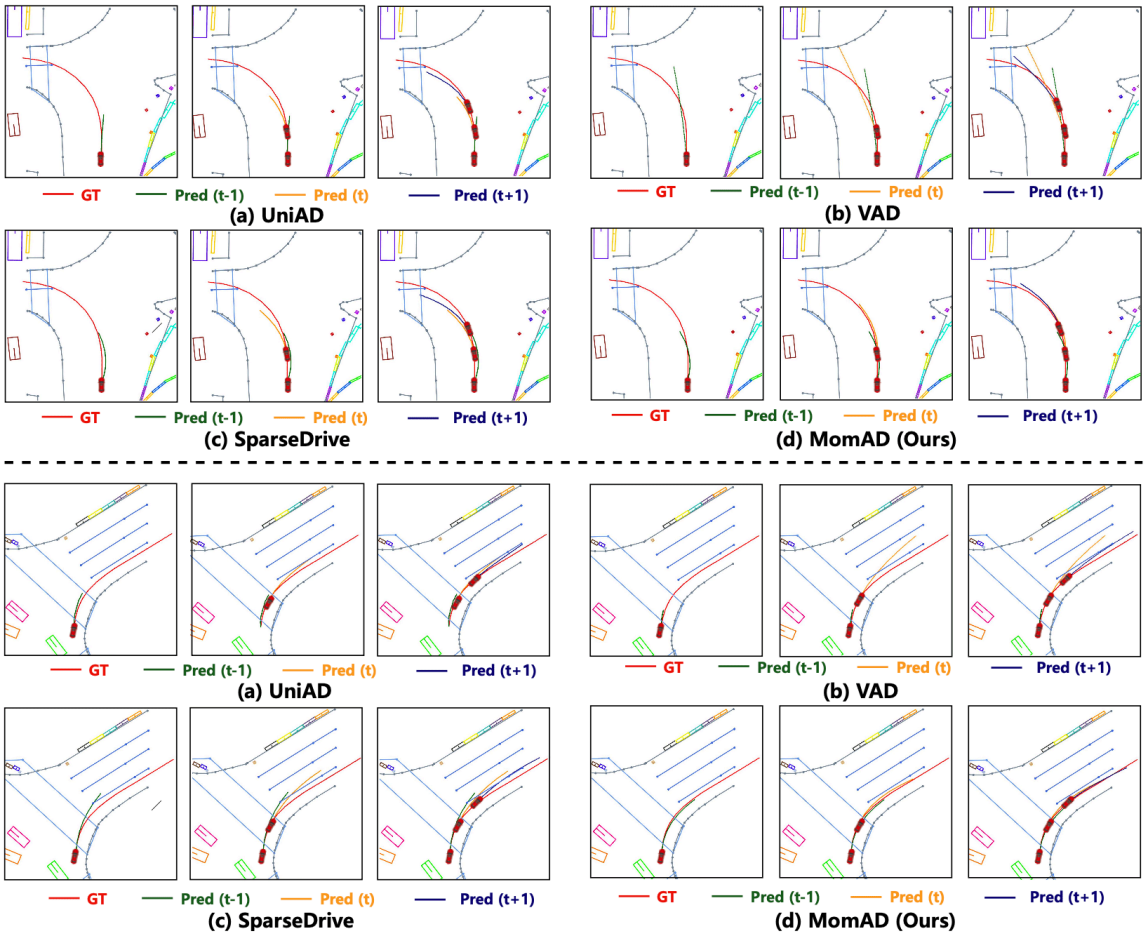

Figure A4. More visualization results of MomAD with SOTA methods across multiple frames.

Acknowledgements

We sincerely appreciate the helpful discussions provided by Wenchao Sun, Bo Jiang, Bencheng Liao from Horizon Robotics. This work was supported by the National Key R&D Program of China (2018AAA0100302).

References

1. a, b Song Z, Liu L, Jia F, Luo Y, Jia C, Zhang G, Yang L, Wang L (2024). "Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook". IEEE Transactions on Intelligent Transportation Systems. pages 1–30. doi:10.1109/TITS.2024.3439557. ISSN 1558-0016.
2. ^ Wang L, Zhang X, Song Z, Bi J, Zhang G, Wei H, Tang L, Yang L, Li J, Jia C, et al. Multi-modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy. IEEE Transactions on Intelligent Vehicles. 2023.
3. a, b Chen L, Wu P, Chitta K, Jaeger B, Geiger A, Li H (2024). "End-to-end autonomous driving: Challenges and frontiers". IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024. Published by IEEE.
4. a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u Hu Y, Yang J, Chen L, Li K, Sima C, Zhu X, Chai S, Du S, Lin T, Wang W, Lu L, Jia X, Liu Q, Dai J, Qiao Y, Li H. "Planning-oriented autonomous driving." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:17853-17862.

5. a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, ^, ^Jiang B, Chen S, Xu Q, Liao B, Chen J, Zhou H, Zhang Q, Liu W, Huang C, Wang X (2023). "Vad: Vectorized scene representation for efficient autonomous driving". In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340--8350.
6. ^Prakash A, Chitta K, Geiger A. "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving." In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. doi:10.1109/cvpr46437.2021.00700.
7. a, b, c, d, eChen S, Jiang B, Gao H, Liao B, Xu Q, Zhang Q, Huang C, Liu W, Wang X (2024). "Vadv2: End-to-end vectorized autonomous driving via probabilistic planning". arXiv preprint arXiv:2402.13243. Available from: https://arxiv.org/abs/2402.13243.
8. a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, ^, ^Sun W, Lin X, Shi Y, Zhang C, Wu H, Zheng S (2024). "SparseDrive: End-to-End Autonomous Driving via Sparse Scene Representation". arXiv preprint arXiv:2405.19620.
9. ^Aradi S. (2020). "Survey of deep reinforcement learning for motion planning of autonomous vehicles". IEEE Transactions on Intelligent Transportation Systems. **23** (2): 740–759.
10. ^Claussmann L, Revilloud M, Gruyer D, Glaser S (2020). "A Review of Motion Planning for Highway Autonomous Driving". IEEE Transactions on Intelligent Transportation Systems. **21** (5): 1826–1848. doi:10.1109/TITS.2019.2913998.
11. a, bYe T, Jing W, Hu C, Huang S, Gao L, Li F, Wang J, Guo K, Xiao W, Mao W, Zheng H, Li K, Chen J, Yu K (2023). "FusionAD: Multi-modality Fusion for Prediction and Planning Tasks of Autonomous Driving". arXiv. arXiv:2308.01006 [cs.CV].
12. a, b, c, dHu S, Chen L, Wu P, Li H, Yan J, Tao D (2022). "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning". In: European Conference on Computer Vision. Springer. pp. 533–549.
13. a, b, cWang J, Zhang X, Xing Z, Gu S, Guo X, Hu Y, Song Z, Zhang Q, Long X, Yin W (2024). "HE-Drive: Human-Like End-to-End Driving with Vision Language Models". arXiv. arXiv:2410.05051 [cs.CV].
14. ^Ji X, Liu Y, Liu YS, Zhang JH, Zhao Y (2021). "Large-momentum effective theory". Reviews of Modern Physics. **93** (3): 035005.
15. ^Wu P, Jia X, Chen L, Yan J, Li H, Qiao Y (2022). "Trajectory-guided Control Prediction for End-to-end Autonomous Driving: A Simple yet Strong Baseline". In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2022. p. 6119-6132. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/286a371d8a0a559281f682f8fbf89834-Paper-Conference.pdf.
16. ^Codevilla F, Santana E, Lopez AM, Gaidon A (2019). "Exploring the Limitations of Behavior Cloning for Autonomous Driving". In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.
17. ^Codevilla F, Mfcller M, Lf3pez A, Koltun V, Dosovitskiy A (2018). "End-to-End Driving Via Conditional Imitation Learning." In: 2018 IEEE International Conference on Robotics and Automation (ICRA), 4693-4700. doi:10.1109/ICRA.2018.8460487.
18. ^Zhang Z, Liniger A, Dai D, Yu F, Van Gool L (2021). "End-to-End Urban Driving by Imitating a Reinforcement Learning Coach". In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15222-15232.
19. ^Codevilla F, Müller M, López A, Koltun V, Dosovitskiy A. End-to-end driving via conditional imitation learning. In: 2018 IEEE international conference on robotics and automation (ICRA). IEEE; 2018. p. 4693–4700.
20. ^Zeng W, Luo W, Suo S, Sadat A, Yang B, Casas S, Urtasun R. "End-To-End Interpretable Neural Motion Planner." In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun. doi:10.1109/cvpr.2019.00886.
21. ^Mao J, Qian Y, Ye J, Zhao H, Wang Y (2023). "Gpt-driver: Learning to drive with gpt". arXiv preprint arXiv:2310.01415. **2023**. Available from: arXiv:2310.01415.
22. ^Zheng W, Song R, Guo X, Chen L (2024). "Genad: Generative end-to-end autonomous driving". arXiv preprint arXiv:2402.11502.
23. ^Zhang Y, Qian D, Li D, Pan Y, Chen Y, Liang Z, Zhang Z, Zhang S, Li H, Fu M, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. arXiv preprint arXiv:2403.19098. 2024.
24. a, bDoll S, Hanselmann N, Schneider L, Schulz R, Cordts M, Enzweiler M, Lensch H (2024). "DualAD: Disentangling the Dynamic and Static World for End-to-End Driving". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024:14728-14737.

25. [a], [b]Chen Z, Ye M, Xu S, Cao T, Chen Q (2025). "Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving". In: European Conference on Computer Vision. Springer; 2025. p. 239–256.
26. [A]Guo M, Zhang Z, He Y, Wang K, Jing L (2024). "End-to-end autonomous driving without costly modularization and 3d manual annotation". arXiv preprint arXiv:2406.17680.
27. [A]Yang Z, Song N, Li W, Zhu X, Zhang L, Torr PHS (2024). "DeepInteraction++: Multi-Modality Interaction for Autonomous Driving". arXiv preprint arXiv:2408.05075.
28. [A]Jiang B, Chen S, Liao B, Zhang X, Yin W, Zhang Q, Huang C, Liu W, Wang X (2024). "Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving". arXiv. arXiv:2410.22313 [cs.CV].
29. [A]Jiang B, Chen S, Wang X, Liao B, Cheng T, Chen J, Zhou H, Zhang Q, Liu W, Huang C (2022). "Perceive, interact, predict: Learning dynamic and static clues for end-to-end motion prediction". arXiv preprint arXiv:2212.02181.
30. [A]Su H, Wu W, Yan J (2024). "DiFSD: Ego-Centric Fully Sparse Paradigm with Uncertainty Denoising and Iterative Refinement for Efficient End-to-End Autonomous Driving". arXiv preprint arXiv:2409.09777.
31. [A]Cheng J, Chen Y, Mei X, Yang B, Li B, Liu M. "Rethinking imitation-based planners for autonomous driving." In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE; 2024. p. 14123-14130.
32. [A]Li Z, Li K, Wang S, Lan S, Yu Z, Ji Y, Li Z, Zhu Z, Kautz J, Wu Z, et al. "Hydra-MDP: End-to-end Multimodal Planning with Multi-target Hydra-Distillation". arXiv preprint arXiv:2406.06978. 2024.
33. [A]Hu P, Huang A, Dolan J, Held D, Ramanan D (2021). "Safe local motion planning with self-supervised freespace forecasting". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12732–12741.
34. [A]Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, Krishnan A, Pan Y, Baldan G, Beijbom O (2020). "nuscenes: A multimodal dataset for autonomous driving". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11621--11631.
35. [A]Jia X, Yang Z, Li Q, Zhang Z, Yan J (2025). "Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving". Advances in Neural Information Processing Systems. **37**: 819–844.
36. [a], [b], [c]Li Z, Yu Z, Lan S, Li J, Kautz J, Lu T, Alvarez JM (2024). "Is ego status all you need for open-loop end-to-end autonomous driving?" In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14864–14873.
37. [A]He K, Zhang X, Ren S, Sun J (2016). "Deep residual learning for image recognition". In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.
38. [A]Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2018). "Focal Loss for Dense Object Detection". arXiv. arXiv:1708.02002 [cs.CV].
39. [a], [b]Dong Y, Kang C, Zhang J, Zhu Z, Wang Y, Yang X, Su H, Wei X, Zhu J. "Benchmarking Robustness of 3D Object Detection to Common Corruptions." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023:1022-1032.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.