

# Decoupling Scene Perception and Ego Status: A Multi-Context Fusion Approach for Enhanced Generalization in End-to-End Autonomous Driving

Jiacheng Tang<sup>1</sup>, Mingyue Feng<sup>2</sup>, Jiachao Liu<sup>2</sup>, Yaonong Wang<sup>2</sup>, Jian Pu<sup>1\*</sup>

<sup>1</sup>Fudan University

<sup>2</sup>Zhejiang Leapmotor Technology Co., Ltd.

jiachengtang21@m.fudan.edu.cn, jianpu@fudan.edu.cn

## Abstract

Modular design of planning-oriented autonomous driving has markedly advanced end-to-end systems. However, existing architectures remain constrained by an over-reliance on ego status, hindering generalization and robust scene understanding. We identify the root cause as an inherent design within these architectures that allows ego status to be easily leveraged as a shortcut. Specifically, the premature fusion of ego status in the upstream BEV encoder allows an information flow from this strong prior to dominate the downstream planning module. To address this challenge, we propose AdaptiveAD, an architectural-level solution based on a multi-context fusion strategy. Its core is a dual-branch structure that explicitly decouples scene perception and ego status. One branch performs scene-driven reasoning based on multi-task learning, but with ego status deliberately omitted from the BEV encoder, while the other conducts ego-driven reasoning based solely on the planning task. A scene-aware fusion module then adaptively integrates the complementary decisions from the two branches to form the final planning trajectory. To ensure this decoupling does not compromise multi-task learning, we introduce a path attention mechanism for ego-BEV interaction and add two targeted auxiliary tasks: BEV unidirectional distillation and autoregressive online mapping. Extensive evaluations on the nuScenes dataset demonstrate that AdaptiveAD achieves state-of-the-art open-loop planning performance. Crucially, it significantly mitigates the over-reliance on ego status and exhibits impressive generalization capabilities across diverse scenarios.

## Introduction

The modular design of planning-oriented autonomous driving (Hu et al. 2023) offers a novel paradigm for end-to-end models, yet it inherits the persistent challenge of causal confusion (Muller et al. 2005), where models learn spurious correlations from benchmark data. A critical manifestation is the tendency for models to ‘drive by inertia’ rather than ‘driving by sight’, a shortcut that leads to catastrophic failures in novel or long-tail scenarios where historical momentum fails to predict the requisite future action. Existing mitigation efforts have largely followed two paths: data-centric strategies, like balanced sampling (Chen et al. 2024a), which

\*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

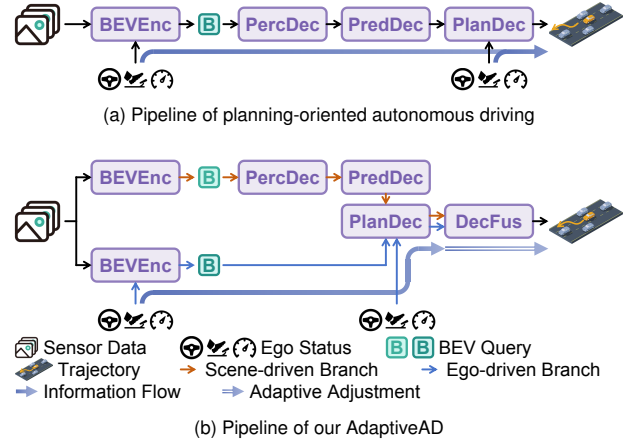


Figure 1: Ego-status shortcut and our proposed architectural solution. (a) In conventional architectures, ego status is coupled with scene context, creating a shortcut that allows the planning module to rely on kinematic state. (b) Our AdaptiveAD framework uses a dual-branch design to explicitly decouple scene-driven reasoning from ego-status influence. A scene-aware fusion module then adaptively integrates these complementary decision contexts to generate the final trajectory.

mitigate dataset biases but do not alter the model’s internal information flow; and regularization techniques, such as dropout and contrastive imitation learning (Cheng, Chen, and Chen 2024), which improve feature quality but risk exacerbating the difficulty of multi-task learning in complex end-to-end frameworks. These approaches focus on refining the inputs to the decision-making process, rather than restructuring the process itself.

Recent investigations have revealed that this decision-making inertia stems from an over-reliance on the vehicle’s own kinematic state, often termed ego status (Li et al. 2024b). For instance, when a high-speed vehicle must execute an emergency maneuver to avoid a sudden obstacle, a model ‘driving by inertia’ is prone to generating a fatal trajectory. While dataset biases, such as the prevalence of straight-driving scenarios in benchmarks like nuScenes (Caesar et al. 2020), certainly exacerbate this issue, we ar-

gue the root cause is a critical but often overlooked inherent design that allows ego status to be easily leveraged as a shortcut. In many state-of-the-art architectures, ego status is fused with perceptual features early in the processing pipeline, as illustrated in Figure 1a. This unfiltered access allows the planning module to depend directly on ego status, bypassing complex scene understanding.

To address this, we propose AdaptiveAD, an architecture that explicitly decouples scene perception from ego status via a dual-branch, multi-context fusion strategy, as illustrated in Figure 1b. A scene-driven branch, with ego status excluded during BEV encoding, generates a planning decision derived solely from environmental cues. Concurrently, a lightweight, ego-driven branch generates a trajectory based primarily on the vehicle’s kinematic state. These two complementary decision contexts—one encoding external scene constraints, the other the vehicle’s own kinematic inertia—are then adaptively integrated by a scene-aware fusion module to produce a final trajectory that is both dynamically feasible and contextually appropriate.

Notably, architectural decoupling may introduce challenges for multi-task learning. To support our design, we further propose several key innovations. Path attention mechanism models the fine-grained interactions between potential ego paths and the dense BEV feature map, to enhance both long-range dependency modeling and local detail capture. To further strengthen feature-behavior associations, we employ two auxiliary tasks. BEV unidirectional distillation mitigates potential motion blur in the perception features of the scene-driven branch arising from the absence of ego status, while autoregressive online mapping establishes a feedback loop from planning to mapping to improve multi-task learning efficacy. Evaluated on the nuScenes benchmark, AdaptiveAD not only achieves state-of-the-art open-loop planning performance but, more importantly, demonstrates significantly reduced reliance on the ego-status shortcut and superior generalization in complex scenarios.

Our primary contributions are outlined as follows:

- We identify and address the architectural flaw that enables the ego-status shortcut. We propose a multi-context fusion strategy that explicitly decouples scene- and ego-driven reasoning, offering a robust solution to causal confusion in end-to-end models.
- We introduce three key innovations to support our decoupling strategy: a path attention mechanism to enhance ego-BEV interaction, BEV unidirectional distillation to preserve perceptual quality, and autoregressive online mapping to improve multi-task consistency.
- Our proposed framework, AdaptiveAD, sets a new state of the art for open-loop planning on the nuScenes dataset, demonstrating superior robustness and generalization in diverse and challenging driving scenarios.

## Related Work

### End-to-End Autonomous Driving Architectures

The paradigm for autonomous driving has progressively shifted from modular pipelines, which are prone to error accumulation, towards integrated end-to-end systems.

These systems aim to jointly optimize perception, prediction, and planning within a single differentiable framework. A significant milestone in this domain is the development of planning-oriented models like UniAD (Hu et al. 2023), which leverage a unified query-based design to facilitate communication across different tasks. This approach has demonstrated that coordinating tasks with the ultimate goal of planning in mind can substantially improve performance over simple multi-task learning. However, while these architectures (Jiang et al. 2023) have advanced the field, their intricate data flow has inadvertently created new vulnerabilities, most notably the risk of learning spurious correlations from data biases (Li et al. 2024b).

### Causal Confusion and Ego-status Over-reliance

Causal confusion, particularly shortcut learning from ego status (Li et al. 2024b), has prompted diverse mitigation strategies. Data-level methods, such as re-sampling (Chen et al. 2024a) or augmentation (Katare et al. 2024), alleviate dataset biases but do not fundamentally alter the model’s internal information flow. Representation-level techniques, like dropout (Bansal, Krizhevsky, and Ogale 2018) and contrastive learning (Cheng, Chen, and Chen 2024), refine feature quality or restrict inputs to encourage perceptual reliance. While valuable, these approaches primarily address symptoms by refining inputs, rather than restructuring the decision-making process itself. In contrast, our work intervenes at the architectural level. We argue the root cause is the premature fusion of ego status within the BEV encoder, creating an information flow that allows the planner to bypass scene understanding (Li et al. 2024b; Yan et al. 2025). Instead of manipulating data or inputs, AdaptiveAD introduces a structural solution. By architecturally decoupling scene- and ego-driven reasoning before adaptive fusion, we directly suppress the ego-status shortcut, compelling the model to ground its decisions in environmental perception. This offers a more principled solution to causal confusion.

### Fusion Strategies and Feedback Mechanisms

Information fusion is central to robust autonomous systems. Multi-sensor fusion (Yang et al. 2022; Cai et al. 2023; Yan et al. 2023, 2024), which combines data from different modalities like cameras and LiDAR, is a well-established technique for enhancing perception. However, its impact on planning is often indirect (Ye et al. 2023). In contrast, our work explores the fusion of different decision contexts—a conceptually distinct approach aimed directly at improving the planning logic itself. Furthermore, enforcing causal consistency between a model’s planned actions and its perception of the world is crucial for robust decision-making. Methodologies employing autoregressive mechanisms to create feedback loops have become influential for this purpose, ensuring that predicted future states align with intended actions (Li et al. 2024a; Gao et al. 2024). Inspired by this principle, we introduce a targeted autoregressive online mapping task. This task establishes a feedback loop from planning to perception, enhancing the coherence and reliability of our multi-task learning architecture.

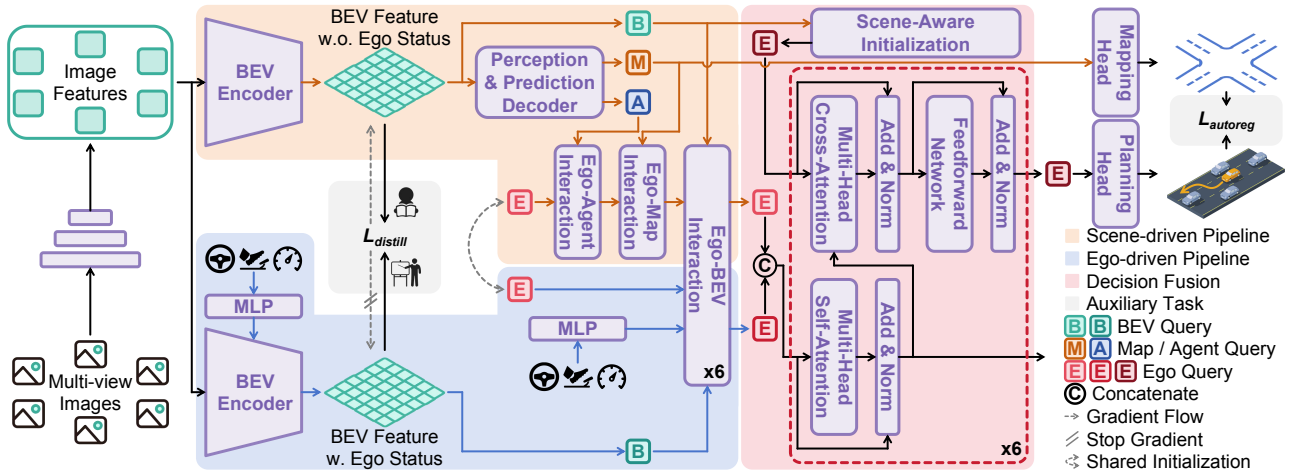


Figure 2: An overview of AdaptiveAD framework. Given a sequence of multi-view images, AdaptiveAD first extracts features using a shared backbone. The core of our framework is a dual-branch architecture that explicitly decouples information flow: one branch generates a scene-driven decision without ego-status influence, while a complementary branch produces an ego-driven decision. These distinct decision contexts are then adaptively integrated by a multi-context decision fusion module, which uses dense scene features as priors to generate the final trajectory. The integrity of this process is supported by two auxiliary tasks designed to enhance perceptual quality and enforce causal consistency.

## Method

The core strategy of AdaptiveAD is to mitigate ego-status over-reliance through a principled, three-part architectural design. Our framework is built upon the explicit decoupling of information streams, the adaptive fusion of their outputs, and the support of targeted regularization to ensure robust multi-task learning. This design, illustrated in Figure 2, systematically severs the shortcut pathway between ego status and planning. It begins by generating distinct scene-driven and ego-driven decisions in a dual-branch architecture, which are then intelligently integrated by a scenario-aware fusion module to produce a final, robust trajectory.

### Multi-context Decision Generation

The cornerstone of AdaptiveAD is its dual-branch architecture, which processes perceptual inputs in parallel streams to isolate ego-status influence. This design facilitates learning from both a pure, scene-centric representation and a conventional, ego-informed representation.

**Scene-driven Branch.** This branch is engineered to generate planning decisions based exclusively on environmental perception, free from the direct influence of ego status. Following VAD (Jiang et al. 2023), it comprises a BEV encoder, a vectorized scene decoder, and a decision generator that process multi-scale image features  $F$ . Our critical modification is removing the BEV query enhancement within the BEV encoder—a step that typically injects ego status and is a primary source of the shortcut. This removal yields a BEV feature,  $B_{woes} \in \mathbb{R}^{C \times H_{bev} \times W_{bev}}$ , representing the environment without ego-motion priors.

The subsequent vectorized scene decoder transforms this dense BEV into sparse agent queries  $A \in \mathbb{R}^{N_{agent} \times C}$  and map queries  $M \in \mathbb{R}^{N_{map} \times C}$ . The decision generator then

initializes a multimodal ego query  $E_{woes} \in \mathbb{R}^{N_{mode} \times C}$  and uses it to interact sequentially with agent queries  $A$ , map queries  $M$ , and BEV feature  $B_{woes}$  to generate a planning decision grounded solely in scene understanding.

**Planning-only Branch.** In contrast, the ego-driven branch mirrors a more conventional architecture by incorporating ego status to generate a complementary decision. This streamlined branch retains the BEV query enhancement operation, producing a motion-compensated BEV feature map,  $B_{wes} \in \mathbb{R}^{C \times H_{bev} \times W_{bev}}$ , informed by ego-kinematics. It bypasses the explicit scene decoder; its ego query,  $E_{wes} \in \mathbb{R}^{N_{mode} \times C}$ , interacts directly with  $B_{wes}$  to be updated into the final ego-driven decision. Reflecting a strong prior for motion extrapolation, the initial reference points for attention in this branch are predicted directly from the ego status.

**Path Attention.** Both branches leverage a novel path attention mechanism for the crucial interaction between the ego query and the BEV feature map. Standard deformable attention mechanisms learn to sample features from sparse, arbitrary locations, which is computationally efficient but lacks task-specific semantic grounding (Li et al. 2023). Path attention refines this concept by introducing trajectory-guided semantic sampling, as depicted in Figure 3.

Specifically, for a given planning modality, we first decode a preliminary trajectory. We then sample  $T$  reference points uniformly in time along this trajectory. Each reference point is assigned an independent attention head, which then learns to sample  $K$  local features in its vicinity. This design constrains the model to gather evidence along a semantically meaningful axis—the hypothesized future path—mimicking how a human driver visually scans their intended route. For the  $l$ -th layer, given the  $i$ -th modality’s ego query  $E^i \in \mathbb{R}^C$  and its  $T$  reference points  $P^i \in \mathbb{R}^{T \times 2}$  on the BEV grid, path

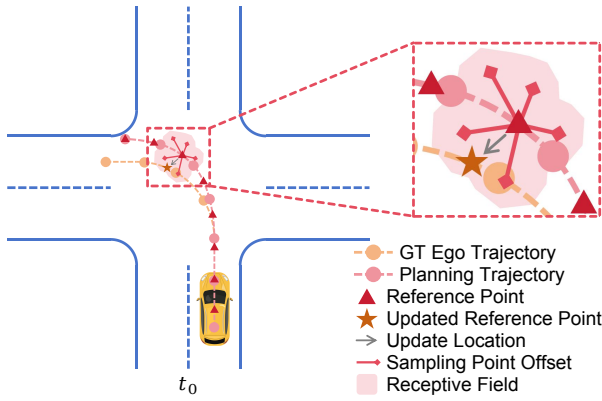


Figure 3: Diagram of path attention.

attention is formulated as:

$$\text{PathAttn}(E^i, P^i, B) = \sum_{t=1}^T W_t \left[ \sum_{k=1}^K a^{i,t,k} W_t' B_{samp}^{i,t,k} \right], \quad (1)$$

$$B_{samp}^{i,t,k} = \text{BiLinear}(B, P^{i,t} + \Delta P^{i,t,k}). \quad (2)$$

where  $B$  is the BEV feature map,  $W_t' \in \mathbb{R}^{C_T \times C}$  and  $W_t \in \mathbb{R}^{C \times C_T}$  are projection matrices for the  $t$ -th head, and  $\text{BiLinear}(\cdot)$  performs bilinear sampling. The sampling offsets  $\Delta P^{i,t,k}$  and attention weights  $a^{i,t,k}$  are predicted from the ego query  $E^i$ . Crucially, the weights are normalized within each head ( $\sum_{k=1}^K a^{i,t,k} = 1$ ), allowing the mechanism to exploit feature separation across heads to efficiently model both long-range context (via the spread of reference points) and local detail (via the learned offsets).

### Multi-context Decision Fusion

Our dual branches produce two distinct decision contexts: a scene-driven  $E_{woes}$  for complex scenarios, and an ego-driven  $E_{wes}$  providing a stable prior for inertial motion. As shown in Figure 2, An adaptive fusion module then adaptively arbitrates between these complementary outputs.

First, to ground the final decision in the current environment, the fusion ego query  $E_{fusion} \in \mathbb{R}^{N_{mode} \times C}$  is initialized with scene awareness. We extract a global scene feature from the scene-driven BEV map, which is shared across all planning modalities:

$$E_{fusion}^{com} = \text{GAP}(B_{woes}), \quad (3)$$

where  $\text{GAP}(\cdot)$  denotes global average pooling. This is combined with learnable modality-specific embeddings to form the initial  $E_{fusion}$ .

Next, a stack of transformer-based fusion layers merges the contexts. A key challenge is the potential feature space misalignment between  $E_{woes}$  and  $E_{wes}$ . To address this, each fusion layer first performs a context alignment step. The two decision queries are concatenated and processed by a multi-head self-attention (MHSA) block:

$$E_{multi} = \text{Concat}(\text{Proj}_{woes}(E_{woes}), \text{Proj}_{wes}(E_{wes})), \quad (4)$$

$$E'_{multi} = \text{LayerNorm}(E_{multi} + \text{MHSA}(E_{multi})). \quad (5)$$

This self-attention mechanism is critical, as it allows for rich inter-context interactions ( $E_{woes} \rightarrow E_{wes}, E_{wes} \rightarrow E_{woes}$ ) and intra-context refinement ( $E_{woes} \rightarrow E_{woes}, E_{wes} \rightarrow E_{wes}$ ). This step produces an aligned and enriched multi-context representation  $E'_{multi} \in \mathbb{R}^{2N_{mode} \times C}$ . The final decision query  $E_{fusion}$  then attends to this representation via multi-head cross-attention to synthesize the final output, adaptively weighting the scene-driven and ego-driven information based on the demands of the current scenario.

### Auxiliary Tasks for Regularization

Our dual-branch architecture is regularized by two targeted auxiliary tasks to preserve decoupled representations and enforce multi-task consistency.

**BEV Unidirectional Distillation.** The scene-driven branch, by design, lacks ego-motion compensation, which can lead to motion blur in its BEV features ( $B_{woes}$ ), particularly for dynamic agents. To counteract this, we introduce a BEV unidirectional distillation task. We treat the motion-compensated BEV from the ego-driven branch,  $B_{wes}$ , as a ‘teacher’ and  $B_{woes}$  as a ‘student’. This teacher-student paradigm is a proven method for knowledge transfer. The distillation loss is a composite objective:

$$L_{distill} = \alpha L_{distill}^{DF} + \beta L_{distill}^{IK} + \gamma L_{distill}^{IC}. \quad (6)$$

Here,  $L_{distill}^{DF}$  is a dense feature distillation loss with an agent-guided reweighting mechanism to focus on foreground objects.  $L_{distill}^{IK}$  and  $L_{distill}^{IC}$  are inter-keypoint and inter-channel distillation losses that encourage the student to learn feature correlations across agent keypoints and channels, narrowing the semantic gap between the two BEV representations. To ensure the teacher’s stability, gradients are not propagated through  $B_{wes}$  during this loss computation.

**Autoregressive Online Mapping.** End-to-end models jointly performing mapping and planning suffer from conflicting optimization goals, where the model might ambiguously shift map elements versus the planning trajectory. To resolve this, we introduce an auxiliary task inspired by driving world models, enforcing causal consistency by predicting future states based on actions.

Our task enforces that the perceived map should be consistent whether the ego vehicle follows the predicted trajectory or the ground-truth trajectory, as depicted in Figure 4. Let  $\hat{P}_E, P_E \in \mathbb{R}^{T \times 2}$  be the predicted and ground-truth ego trajectories, and  $\hat{P}_M, P_M \in \mathbb{R}^{N_{map} \times N_{point} \times 2}$  be the predicted and ground-truth map instances. For each future timestep  $\tau$ , we define an ‘autoregressive key region’ as the intersection of the ego vehicle’s perception bounding box if it were to follow  $\hat{P}_E$  versus  $P_E$ . We then apply a masked L1 loss,  $L_{autoreg}^{MAP}$ , only on the map instances within this region:

$$L_{autoreg}^{MAP} = \frac{1}{T} \sum_{\tau=1}^T \frac{1}{\|\mathcal{M}\|_1 + \epsilon} \left\| (\hat{P}_M - P_M) \odot \mathcal{M} \right\|_1, \quad (7)$$

where  $\mathcal{M} = \text{Mask}(P_M, \hat{R}_E^\tau \cap R_E^\tau)$  is the mask for key instances in the overlapping perception region, with  $\epsilon$  being

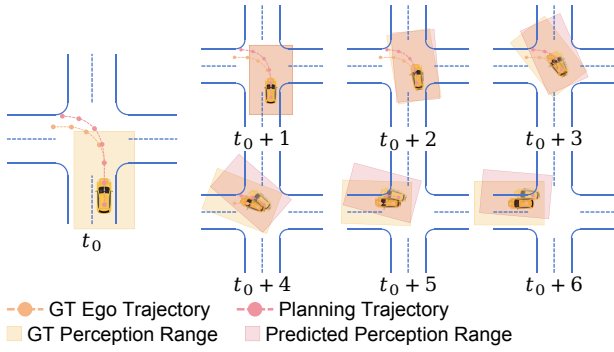


Figure 4: Diagram of autoregressive online mapping.

a smoothing term. This ensures the model learns a consistent mapping outcome conditioned on its planned actions. To handle cases with no overlap and provide a stable gradient, we add a Gaussian Wasserstein distance loss (Yang et al. 2021),  $L_{autoreg}^{GWD}$ , on the perception bounding boxes themselves. The total loss is:

$$L_{autoreg} = \delta L_{autoreg}^{MAP} + \lambda L_{autoreg}^{GWD}. \quad (8)$$

This autoregressive supervision establishes a feedback loop from planning to mapping, mitigating task conflict and promoting a more coherent world representation.

## Experiments

### Experimental Setup

**Datasets and Evaluation Metrics.** Our primary evaluation is conducted on the nuScenes (Caesar et al. 2020) dataset, a standard benchmark for vision-centric autonomous driving. Following established protocols (Jiang et al. 2023), we train on the train set and report results on the val set. We use the primary open-loop planning metrics: L2 Displacement Error (L2) to measure trajectory accuracy and Collision Rate (CR) to assess safety (Jiang et al. 2023). To further probe model robustness, we conduct supplementary experiments on the NAVSIM (Dauner et al. 2024) non-reactive benchmark and the Bench2Drive (Jia et al. 2024) closed-loop simulator, using their official metrics: PDM Score (PDMS) (Dauner et al. 2024), Driving Score (DS), and Success Rate (SR) (Jia et al. 2024). Unless otherwise specified, all reported results are based on the nuScenes dataset. And all FPS measurements are conducted on one NVIDIA GeForce RTX 3090 GPU, with the exception of UniAD (Hu et al. 2023) and PPAD (Chen et al. 2024b).

**Implementation Details.** AdaptiveAD is implemented in PyTorch, leveraging the open-source MMDetection3D framework. Our configuration largely follows VAD, predicting a 3-second trajectory from 2 seconds of historical data within a 60m x 30m perception range. The model includes 6 ego-BEV interaction layers and 6 multi-context fusion layers. Loss weights for auxiliary tasks ( $\alpha, \beta, \gamma, \delta, \lambda$ ) are set to (0.01, 0.1, 0.01, 0.01, 0.01). We train for 60 epochs on 32 NVIDIA A100 GPUs using the AdamW optimizer and a CosineAnnealing scheduler, with a batch size of 2 per GPU.

Method	L2 (m) ↓				Collision (%) ↓				FPS
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	
UniAD	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61	1.8
VAD §	0.31	0.58	0.94	0.61	0.17	0.25	0.43	0.28	3.4
PPAD	0.31	0.56	0.87	0.58	0.08	0.12	0.38	0.19	2.6
SparseDrive †	0.30	0.58	0.95	0.61	0.01	0.05	0.23	0.10	5.2
BridgeAD †	0.28	0.55	0.92	0.58	<b>0.00</b>	<b>0.04</b>	0.20	<b>0.08</b>	3.1
FusionAD	-	-	-	1.03	0.25	0.13	0.25	0.21	-
Ours	<b>0.23</b>	<b>0.43</b>	<b>0.74</b>	<b>0.47</b>	0.05	0.12	<b>0.18</b>	0.12	3.0

Table 1: Open-loop planning performance. § denotes re-implement result. † denotes that the auxiliary task of predicting ego status was used during training.

Method	Nav.	L2 (m) ↓				Collision (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD	ST	0.32	0.59	0.95	0.62	0.20	0.30	0.48	0.33
	LR	0.50	0.88	1.35	0.91	<b>0.00</b>	<b>0.11</b>	0.44	0.18
	LR ‡	0.40	0.83	1.52	0.92	0.07	0.22	0.84	0.38
Ours	ST	<b>0.23</b>	<b>0.43</b>	<b>0.74</b>	<b>0.47</b>	<b>0.06</b>	<b>0.11</b>	<b>0.17</b>	<b>0.11</b>
	LR	<b>0.34</b>	<b>0.61</b>	<b>0.95</b>	<b>0.63</b>	<b>0.00</b>	0.15	<b>0.34</b>	<b>0.16</b>
	LR ‡	<b>0.26</b>	<b>0.55</b>	<b>1.09</b>	<b>0.63</b>	<b>0.04</b>	<b>0.11</b>	<b>0.68</b>	<b>0.28</b>

Table 2: Scene generalization ability. Navigation commands are labeled as ‘ST’ (straight) and ‘LR’ (left/right turn) (Li et al. 2024b). ‡ split follows the Turning-nuScenes dataset’s official protocol (Song et al. 2025).

Method	Velo. Noise	L2 (m) ↓				Collision (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
VAD	-	0.31	0.58	0.94	0.61	0.17	0.25	0.43	0.28
	×0.0	3.53	5.61	7.48	5.54	0.53	2.15	3.85	2.18
	×0.5	1.90	3.07	4.17	3.05	0.23	0.52	1.15	0.63
	×1.5	1.91	3.21	4.53	3.22	0.22	0.96	1.98	1.05
	100m/s	8.94	14.92	20.93	14.93	7.47	6.20	5.11	6.26
Ours	-	0.23	0.43	0.74	0.47	0.05	0.12	0.18	0.12
	×0.0	2.62	4.18	5.43	4.08	0.20	1.07	2.13	1.13
	×0.5	1.52	2.45	3.26	2.41	0.09	0.16	0.49	0.25
	×1.5	1.66	2.76	3.81	2.74	0.16	0.60	1.53	0.76
	100m/s	2.89	5.04	7.24	5.06	3.25	5.42	5.92	4.86

Table 3: Reliance on ego status on nuScenes. We assign different levels of perturbation to the ego-velocity.

Method	Velo. Noise	NAVSIM PDMS ↑	Bench2Drive	
			DS ↑	SR(%) ↑
VAD	-	81.2	44.35	16.91
	×0.0	51.5	14.63	3.45
	×0.5	78.5	41.22	13.18
	×1.5	67.9	22.78	7.27
	100m/s	22.1	6.14	0.00
Ours	-	86.4	49.47	19.23
	×0.0	61.4	19.91	5.82
	×0.5	84.6	45.81	16.44
	×1.5	76.6	29.88	9.54
	100m/s	30.7	9.42	0.00

Table 4: Reliance on ego status on NAVSIM and Bench2Drive.

ID	D. B. S. A.	L2 (m) ↓				Collision (%) ↓				FPS
		1s	2s	3s	Avg.	1s	2s	3s	Avg.	
1	- - - -	0.28	0.54	0.90	0.57	0.10	0.20	0.35	0.22	3.4
2	✓ - - -	0.30	0.58	0.98	0.62	0.10	0.14	0.20	0.15	3.0
3	✓ ✓ - -	0.28	0.53	0.92	0.58	0.03	0.05	0.17	0.08	3.0
4	✓ ✓ ✓ -	0.25	0.49	0.82	0.52	0.06	0.13	0.18	0.12	3.0
5	✓ ✓ ✓ ✓	<b>0.23</b>	<b>0.43</b>	<b>0.74</b>	<b>0.47</b>	<b>0.05</b>	<b>0.12</b>	<b>0.18</b>	<b>0.12</b>	3.0

Table 5: Ablation studies for innovative components. ‘D.’ (dual-branch), ‘B.’ (BEV unidirectional distillation), ‘S.’ (scene-aware initialization), and ‘A.’ (autoregressive online mapping) are incrementally added to the baseline.

Mech.	L2 (m) ↓				Collision (%) ↓				FPS
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	
D.A.	0.23	0.45	0.77	0.48	0.07	0.13	0.25	0.15	3.0
P.A.	<b>0.23</b>	<b>0.43</b>	<b>0.74</b>	<b>0.47</b>	<b>0.05</b>	<b>0.12</b>	<b>0.18</b>	<b>0.12</b>	3.0

Table 6: Effectiveness of path attention.

Method	L2 (m) ↓				Collision (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
+ P.A.	0.43	0.65	0.96	0.68	0.57	0.55	0.54	0.55
SparseDrive	0.29	0.55	0.91	0.58	0.01	0.02	0.13	0.06
+ A.O.M.	0.27	0.49	0.82	0.53	0.01	0.03	0.14	0.06

Table 7: Effectiveness of plugins on open-loop planning.

## Main Results

**Open-loop Planning Performance.** As shown in Table 1, AdaptiveAD sets a new state of the art in planning accuracy on the nuScenes benchmark with the lowest average L2 error. It also achieves a highly competitive collision rate (Zhang et al. 2025), significantly outperforming our primary baseline, VAD, by reducing the average L2 error by 22% and the collision rate by 57%. Notably, this performance is achieved with negligible impact on latency, maintaining an inference speed of 3.0 FPS.

**Scene Generalization Ability.** The nuScenes dataset is heavily skewed towards simple, straight-driving scenarios (~75%), which can mask model weaknesses. To dissect our performance, we evaluate generalization on complex turning maneuvers (LR) versus simple straight driving (ST), as shown in Table 2. While both models perform well when driving straight, VAD’s performance degrades significantly in turning scenarios. In contrast, AdaptiveAD demonstrates far more consistent performance, substantially outperforming VAD in both L2 error and CR during turns. This result highlights AdaptiveAD’s superior ability to rely on scene understanding when ego-status priors become unreliable.

**Reliance on Ego Status.** To directly test our core hypothesis, we inject varying levels of noise into the ego-velocity input during inference. As detailed in Table 3, VAD’s performance collapses catastrophically under noisy ego status, with L2 error increasing by over 800% when velocity is zeroed out. AdaptiveAD, while still affected, exhibits

markedly greater resilience, with significantly smaller performance degradation across all noise levels. This demonstrates that our multi-context fusion strategy successfully reduces the model’s hazardous over-dependence on ego status. We confirm this finding in more realistic settings on NAVSIM and the closed-loop CARLA-based Bench2Drive (Table 4), where AdaptiveAD consistently maintains higher and more stable driving scores under perturbation.

## Ablation Studies

We conduct a series of ablation studies, summarized in Table 5, to systematically validate the contribution of each component in AdaptiveAD.

**Component Contributions.** Our baseline (ID-1), a scene-driven branch that retains ego-status enhancement, performs reasonably well. Introducing our dual-branch and fusion module but without the auxiliary regularizers (ID-2) leads to a drop in L2 accuracy, as the scene-driven branch suffers from motion blur without ego-motion compensation. The addition of BEV unidirectional distillation (ID-3) is critical; it recovers the L2 performance while dramatically reducing the collision rate by over 60%. This is because distillation enhances the perceptual quality of the scene-driven BEV features ( $B_{woes}$ ), as qualitatively verified in Figure 6. Next, incorporating scene-aware initialization (ID-4) further improves trajectory accuracy by ~10%, enabling the model to adaptively weight the two decision contexts based on scene complexity. Finally, adding our autoregressive online mapping task (ID-5) yields the full AdaptiveAD model, which achieves the best overall performance by improving trajectory quality without compromising safety. As shown in Figure 7, this auxiliary task also accelerates model convergence by mitigating the optimization conflict between the mapping and planning heads.

**Effectiveness of Path Attention.** We compare our proposed path attention (P.A.) against a standard deformable attention (D.A.) mechanism (Zhu et al. 2020) for the ego-BEV interaction module. As shown in Table 6, path attention achieves superior long-term planning accuracy and safety with identical computational overhead. This confirms that guiding attention sampling along the hypothesized future trajectory is a more effective strategy for this task.

**Generalizability of Components.** To demonstrate the broader utility of our contributions, we integrate path attention (P.A.) and autoregressive online mapping (A.O.M.) as plug-in modules into two other SOTA models, UniAD (Hu et al. 2023) and SparseDrive (Sun et al. 2024). As reported in Table 7, both components provide consistent performance improvements, underscoring their general applicability beyond the AdaptiveAD framework.

## Qualitative Results

Figure 5 provides a qualitative comparison in a challenging obstacle avoidance scenario. The baseline model, likely over-relying on its prior for forward motion, fails to perceive the stopped vehicle and plans a collision course. In contrast,



Figure 5: Qualitative comparison of scene generalization ability. In this challenging scenario, our AdaptiveAD demonstrates significantly superior perception capabilities compared to VAD, providing more reliable obstacle-avoidance paths.

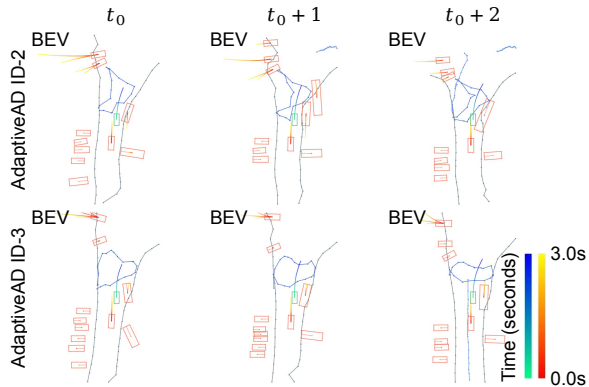


Figure 6: Qualitative evaluation of BEV unidirectional distillation.

AdaptiveAD’s robust scene understanding allows it to correctly identify the obstacle and generate a safe and comfortable avoidance maneuver. Visualizations of two additional state-of-the-art models are provided in Supplementary Material, Section C.

## Conclusion

In this work, we addressed the over-reliance on ego status in end-to-end autonomous driving, diagnosing it as a potential architectural flaw rather than a mere dataset bias. We introduced AdaptiveAD, a framework that remedies this issue through a principled decoupling of scene-driven and ego-driven reasoning, followed by an adaptive fusion module. AdaptiveAD achieves state-of-the-art open-loop planning performance on the nuScenes benchmark. Crucially, our experiments demonstrate this performance improvement does not arise from overfitting to benchmark statistics, but rather from a significant suppression of the ego-state short-

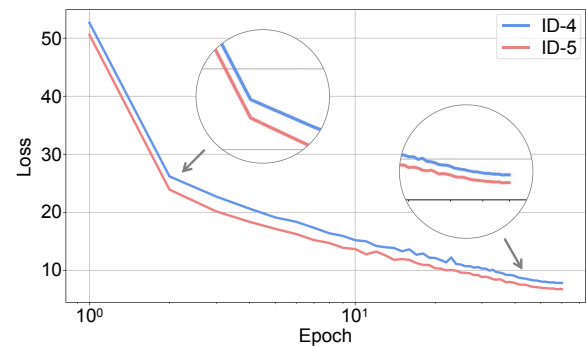


Figure 7: Impact of autoregressive online mapping on model convergence speed.

cut and substantially improved robustness in complex scenarios, a finding corroborated across both open- and closed-loop evaluations.

Beyond a specific implementation, our work champions a broader principle: the explicit decoupling and adaptive fusion of distinct reasoning contexts is a powerful paradigm for building more robust and generalizable driving models. This architectural philosophy opens several promising avenues for future research. For instance, the modularity of our decoupled design provides a clear path for integration with other advanced systems, such as generative world models, to further enhance causal reasoning and scene understanding. Furthermore, it offers a structured approach to enhancing training efficiency and extending the framework to handle multimodal planning. By promoting a more robust and causally sound approach to decision-making, this research contributes to the development of safer, more scalable autonomous driving systems.

## References

- Bansal, M.; Krizhevsky, A.; and Ogale, A. 2018. Chauffeur-net: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*.
- Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 11621–11631.
- Cai, H.; Zhang, Z.; Zhou, Z.; Li, Z.; Ding, W.; and Zhao, J. 2023. BEVFusion4D: Learning LiDAR-Camera Fusion Under Bird’s-Eye-View via Cross-Modality Guidance and Temporal Aggregation. *arXiv preprint arXiv:2303.17099*.
- Chen, W.; Yang, K.; Yu, Z.; Shi, Y.; and Chen, C. P. 2024a. A survey on imbalanced learning: latest research, applications and future directions. *Artif. Intell. Review*, 57(6): 137.
- Chen, Z.; Ye, M.; Xu, S.; Cao, T.; and Chen, Q. 2024b. Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In *Eur. Conf. Comput. Vis.*, 239–256. Springer.
- Cheng, J.; Chen, Y.; and Chen, Q. 2024. Pluto: Pushing the limit of imitation learning-based planning for autonomous driving. *arXiv preprint arXiv:2404.14327*.
- Dauner, D.; Hallgarten, M.; Li, T.; Weng, X.; Huang, Z.; Yang, Z.; Li, H.; Gilitschenski, I.; Ivanovic, B.; Pavone, M.; et al. 2024. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Adv. Neural Inform. Process. Syst.*, 37: 28706–28719.
- Gao, S.; Yang, J.; Chen, L.; Chitta, K.; Qiu, Y.; Geiger, A.; Zhang, J.; and Li, H. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*.
- Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 17853–17862.
- Jia, X.; Yang, Z.; Li, Q.; Zhang, Z.; and Yan, J. 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Adv. Neural Inform. Process. Syst.*, 37: 819–844.
- Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 8340–8350.
- Katare, D.; Noguero, D. S.; Park, S.; Kourtellis, N.; Janssen, M.; and Ding, A. Y. 2024. Analyzing and mitigating bias for vulnerable classes: Towards balanced representation in dataset. *arXiv preprint arXiv:2401.10397*.
- Li, Q.; Jia, X.; Wang, S.; and Yan, J. 2024a. Think2Drive: Efficient Reinforcement Learning by Thinking with Latent World Model for Autonomous Driving (in CARLA-V2). In *Eur. Conf. Comput. Vis.*, 142–158. Springer.
- Li, T.; Jia, P.; Wang, B.; Chen, L.; Jiang, K.; Yan, J.; and Li, H. 2023. Laneseqnet: Map learning with lane segment perception for autonomous driving. *arXiv preprint arXiv:2312.16108*.
- Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024b. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 14864–14873.
- Muller, U.; Ben, J.; Cosatto, E.; Flepp, B.; and Cun, Y. 2005. Off-road obstacle avoidance through end-to-end learning. *Adv. Neural Inform. Process. Syst.*, 18.
- Song, Z.; Jia, C.; Liu, L.; Pan, H.; Zhang, Y.; Wang, J.; Zhang, X.; Xu, S.; Yang, L.; and Luo, Y. 2025. Don’t Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 22432–22441.
- Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; and Zheng, S. 2024. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*.
- Yan, J.; Liu, Y.; Sun, J.; Jia, F.; Li, S.; Wang, T.; and Zhang, X. 2023. Cross modal transformer: Towards fast and robust 3d object detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 18268–18278.
- Yan, Y.; Liu, B.; Ai, J.; Li, Q.; Wan, R.; and Pu, J. 2024. PointSSC: A cooperative vehicle-infrastructure point cloud benchmark for semantic scene completion. In *IEEE Int. Conf. Robot. Autom.*, 17027–17034. IEEE.
- Yan, Y.; Zhou, Z.; Gao, X.; Li, G.; Li, S.; Chen, J.; Pu, Q.; and Pu, J. 2025. Learning Spatial-Aware Manipulation Ordering. In *Adv. Neural Inform. Process. Syst.*
- Yang, X.; Yan, J.; Ming, Q.; Wang, W.; Zhang, X.; and Tian, Q. 2021. Rethinking rotated object detection with gaussian wasserstein distance loss. In *Int. Conf. Mach. Learn.*, 11830–11841. PMLR.
- Yang, Z.; Chen, J.; Miao, Z.; Li, W.; Zhu, X.; and Zhang, L. 2022. Deepinteraction: 3d object detection via modality interaction. *Adv. Neural Inform. Process. Syst.*, 35: 1992–2005.
- Ye, T.; Jing, W.; Hu, C.; Huang, S.; Gao, L.; Li, F.; Wang, J.; Guo, K.; Xiao, W.; Mao, W.; et al. 2023. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*.
- Zhang, B.; Song, N.; Jin, X.; and Zhang, L. 2025. Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 6854–6863.
- Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.