TAD-E2E: A Large-scale End-to-end Autonomous Driving Dataset

Chang Liu* Mingxu Zhu* Zheyuan Zhang* Linna Song Xiao Zhao
Qingliang Luo, Qi Wang, Chufan Guo, Kuifeng Su
ADLab, Tencent
{changeliu,mingxuzhu,lukezyzhang,linnaasong,shawzhao,
fineluo,kiwang,chufanguo,kuifengsu}@tencent.com

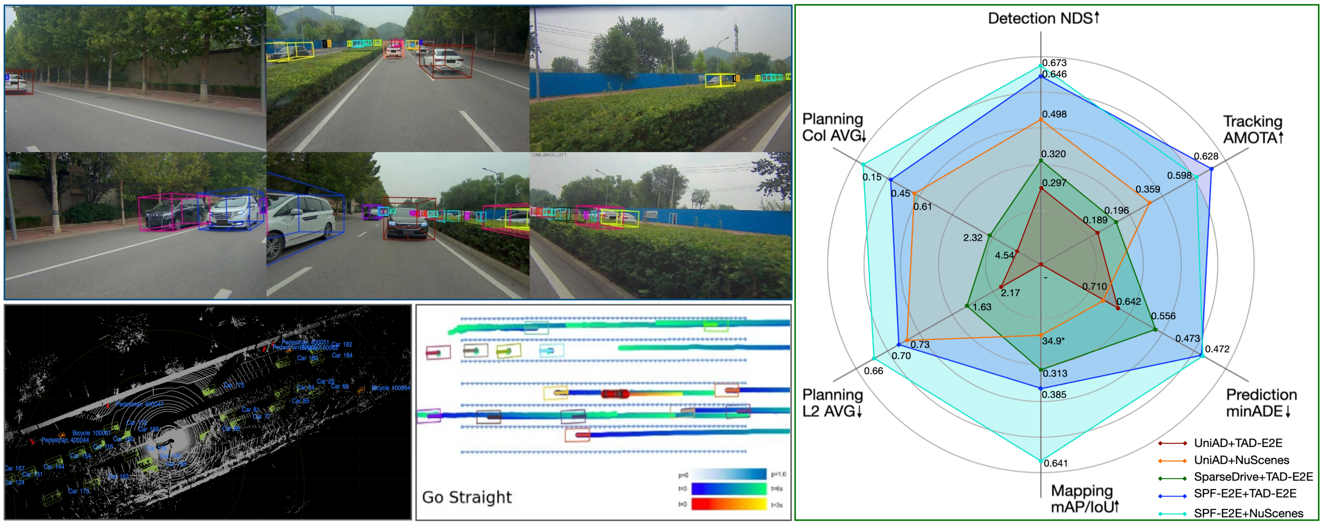


Figure 1. We introduce TAD-E2E, a large-scale dataset for end-to-end autonomous driving, featuring extensive sensor data and challenging urban scenarios. Experiments with state-of-the-art (SOTA) methods on this dataset show notable performance drops, highlighting its complexity. To address these challenges, we propose SparseFusion-E2E (SPF-E2E), a multimodal end-to-end neural network that serves as a strong baseline to facilitate further research.

Abstract

End-to-end autonomous driving technology has recently become a focal point of research and application in autonomous driving. State-of-the-art (SOTA) methods are often trained and evaluated on the nuScenes dataset. However, the nuScenes dataset, introduced in 2019 for 3D perception tasks, faces several limitations—such as insufficient scale, simple scenes, and homogeneous driving behaviors—that restrict the upper-bound development of end-to-end autonomous driving algorithms. In light of these issues, we propose a novel, large-scale real-world dataset specifically designed for end-to-end autonomous driving tasks, named TAD-E2E, which is 25x larger, 1.7x scene complexity over nuScenes, and features a highly diverse range of driving behaviors. We replicated SOTA methods on the TAD-E2E dataset and observed that these methods no longer performed well, as expected. Additionally, in response to the challenging scenarios presented in the TAD-E2E dataset, we devised a multimodal sparse end-to-end method that significantly outperforms SOTA methods. Ablation studies demonstrate the effectiveness of our method, and we analyze the contributions of each module. The dataset will be released in the near future.

1. Introduction

In recent years, the technology of autonomous driving has developed rapidly. As one of the forefront research directions, data-driven end-to-end (E2E) autonomous driving has demonstrated excellent performance. Compared to traditional modular approaches that involve mapping, localization, perception, prediction, and decision-making, end-to-end autonomous driving offers advantages in effectively reducing information transfer loss by enabling joint learning and optimization among various modules at the feature level. Due to the involvement of training and joint learning across various functional modules, end-to-end systems also strongly demand high-quality, large-scale, full-chain

*Equal contribution.

ground truth data for autonomous driving. Notably, the term "ground truth" here refers not only to the truth of a single module but to the comprehensive coverage of all modules under the same spatial and temporal conditions; i.e., it includes the 3D object information ground truth from the perception module, the temporal trajectory ground truth from the prediction module, the vector map ground truth from the mapping module, and the driving trajectory ground truth from the decision-making module.

Introduced in 2019, the nuScenes dataset meets the above requirements for ground truth data structure. It is currently the most widely used dataset for open-loop real-world scenarios in state-of-the-art (SOTA) end-to-end autonomous driving methods. However, the nuScenes dataset was initially designed for 3D detection and tracking tasks related to perception. Although it also provides mapping and localization information, allowing it to support the development of end-to-end autonomous driving models, its overall construction and data collection methodology was not specifically tailored for end-to-end algorithms. Consequently, this limitation may hinder the exploration of upper bounds for end-to-end autonomous driving algorithms:

1. Scale Issues: nuScenes provides only approximately 40,000 annotated frames, which is insufficient for contemporary mainstream algorithm models based on BEV representation and Transformer modular connections. For example, other single-module datasets, such as the MSCOCO [20] for image recognition, contain 320,000 frames, and the Waymo [31] dataset for 3D perception includes 230,000 frames. Moreover, end-to-end autonomous driving requires even more extensive data volumes.

2. Sensor Issues: The nuScenes dataset was collected in early 2019, and the sensors used are now considered outdated relative to those employed in current autonomous vehicles. For instance, the LiDAR used in nuScenes only has 32 channels, whereas the advancement in LiDAR hardware has reached 64 channels or more. The enhancement of sensor hardware capabilities necessitates corresponding updates at the dataset level to support more advanced algorithm research.

3. Scene Difficulty: Overall, the scenes in the nuScenes dataset are relatively simple, with an average of 34.63 objects per frame and a ground truth frame rate of 2Hz. However, the short duration of trajectory sequences makes them insufficient for exploring algorithms in complex urban road scenarios.

4. Driving Behavior Issues: During the initial data collection to recognize obstacles, the routes and driving behaviors captured in the nuScenes dataset did not receive adequate attention. This has led to significant deviations in driving path coverage within the dataset, contributing to overfitting issues in recent end-to-end autonomous driving algorithm research.

Recent studies have indicated that on the nuScenes dataset, it is possible to achieve good results by inferring vehicle driving trajectories using only the vehicle state and multilayer perceptron (MLP) without relying on sensor data or perception modules [18, 38]. Our experiments also confirm this observation, which is unreasonable and indicates potential data issues when using nuScenes for decision-making learning.

Based on the above analysis and research challenges, we propose a novel, high-frequency, large-scale, real-world dataset named TAD-E2E, explicitly designed for end-to-end autonomous driving tasks. Compared to nuScenes, TAD-E2E features 25 times more data, 1.7 times greater scene complexity, and a diverse range of driving behaviors tailored for end-to-end autonomous driving tasks. Additionally, TAD-E2E provides high-quality ground truth for individual modules, supporting independent and joint research across submodules of autonomous driving, including 3D detection, multi-object tracking, trajectory prediction, mapping, and decision-making.

We replicated SOTA methods on the TAD-E2E dataset and observed a significant decline in performance, confirming the research value of more complex scene data and indicating that the existing methods for end-to-end autonomous driving still have upper bounds yet to be explored. Based on SOTA methods, we designed a multimodal sparse end-to-end autonomous driving network model that surpasses SOTA performance. This model will be provided as the baseline for the TAD-E2E dataset upon release. We expect that the TAD-E2E dataset will accelerate the progress of research on end-to-end autonomous driving algorithms and provide a platform for exploring the upper limits of end-to-end algorithm research.

In summary, the main contributions of this paper are:

- We proposed a large-scale end-to-end autonomous driving dataset, TAD-E2E, designed explicitly for complex urban scenarios, providing comprehensive, high-quality ground truth to support the expansion of research boundaries in autonomous driving algorithms. We conducted detailed quantitative and qualitative analyses to demonstrate the value of the TAD-E2E dataset.
- We replicated end-to-end autonomous driving SOTA methods on the TAD-E2E dataset, confirming the limitations of the nuScenes dataset in supporting trajectory planning for end-to-end autonomous driving.
- We introduced a baseline multimodal sparse end-to-end autonomous driving network that outperformed SOTA methods. We conducted extensive quantitative, qualitative, and ablation analyses to validate the effectiveness of our approach. Additionally, we analyzed the necessity and contribution of each module within the modular design of end-to-end autonomous driving methods.
- The dataset will be released in the near future.

Table 1. Comparison of Real Sensor-Based Datasets Related to End-to-End Autonomous Driving. ∗ E2E supported refers to the data support from raw sensor data to ego trajectory in the same frame.

Dataset	Year	Num Samples	Ojects Per Frame	Sequence Length	Annotation Frequency	Detection Supported	Tracking Supported	Prediction Supported	Planning Supported	Mapping Supported	E2E Supported∗
nuScenes [2]	2019	40k	34.63	20s	2Hz	✓	✓	✓	✓	✓	✓
Waymo [31]	2019	230k	52.17	20s	10Hz	✓	✓	✓	-	-	-
Argoverse2 [35]	2023	6M	75	15s	10Hz	✓	✓	✓	-	✓	-
ONCE [24]	2021	16k	26.06	-	2Hz	✓	-	-	-	-	-
KITTI [10]	2013	7.4k	10.72	-	10Hz	✓	-	-	-	-	-
Cityscapes [8]	2016	25k	-	-	-	✓	-	-	-	-	-
TAD-E2E **(ours)**	2025	1M	59.63	60s	10Hz	✓	✓	✓	✓	✓	✓

2. Related Work

2.1. Real World E2E-AD Datasets

There are numerous datasets related to autonomous driving, and for a complete AD datasets review, please refer to paper [22]. Here, we only consider commonly used or highly related to end-to-end autonomous driving [2, 8, 10, 24, 31, 35]. These datasets were generally created focusing on specific autonomous driving module tasks, such as 3D detection and tracking [2, 31, 35] and image recognition [8, 10], rather than for end-to-end (E2E) autonomous driving research. For modular design-based E2E autonomous driving research, datasets must possess several key elements: 1) 3D bounding box annotations for objects, 2) temporal and tracking ID annotations, 3) map information, and 4) trajectories of the collection vehicle. There is an urgent need in the industry for a dedicated, real-world dataset aimed at E2E autonomous driving, characterized by large-scale, complex scenarios and diverse ego vehicle behaviors, to support the rapidly evolving research in E2E autonomous driving algorithms. This is the motivation behind our proposed TAD-E2E dataset. Please refer to Table 1 for a complete dataset comparison. In its latest version, nuPlan [16] released a 128-hour curated subset containing full sensor data, theoretically enabling open-loop end-to-end autonomous driving research. However, we observe that few E2E-AD methodologies experiment on nuPlan open-loop subset, but mostly on nuScenes [13, 15, 32, 37, 39]. We hypothesize this disparity may stem from nuPlan established data usability (in SQL DB format). This observation motivates our proposal for a unified data interface that maintains backward compatibility with nuScenes, enabling seamless migration between benchmark environments.

2.2. E2E-AD Methods

The early development of end-to-end autonomous driving methods can be traced back to the PilotNet [1] approach proposed by Nvidia in 2016 and other early methods that bypassed intermediate tasks such as perception and prediction [6, 7, 28]. Recently, with advancements in Bird's Eye View (BEV) representation [27] and Transformer [34] modeling, research in E2E autonomous driving has rapidly progressed. ST-P3 [12] introduced an E2E method based on a modular loss function that utilizes BEV feature maps. VAD [15] proposed a vectorized representation for scene learning and instance-level planning. UniAD [13] presented a unified learning approach based on BEV representation and Transformer KQV, achieving strong performance across various submodule tasks. FusionAD [37] further integrated LiDAR point cloud data for multimodal learning, enhancing performance metrics. SparseDrive [32] and SparseAD [39] proposed a symmetric system design using sparse representations that do not rely on BEV feature expressions. Additionally, there are end-to-end methods based on large language models [25, 26, 29, 29, 33, 36], which incorporate natural language descriptions during dataset construction with the expectation that the model will learn interpretability in autonomous driving. Building upon SparseDrive and FusionAD, we propose an end-to-end model for multimodal sparse representation learning, improving the performance metrics of various modules on the nuScenes dataset and serving as the baseline method for the TAD dataset.

2.3. Open-Loop and Close-Loop E2E AD

End-to-end autonomous driving can be categorized into two main types: open-loop and closed-loop. Closed-loop E2E AD refers to testing the driving behavior scores of autonomous vehicles in a simulation environment [3, 5, 9, 11, 14]. The simulation system allows the surrounding environment to interact with the autonomous vehicle and change its behavior based on its actions, enabling the simulation of extreme scenarios such as collisions. On the other hand, open-loop E2E AD involves comparing the decision trajectories outputted by the autonomous driving model against expert trajectories collected in real-world scenarios, utilizing metrics such as L2 distance and collision rates. For a detailed introduction and analysis of the advantages and disadvantages of open-loop and closed-loop evaluations, please refer to [4]. This paper pertains to open-loop evaluations in real-world scenarios, with further contemplation and exploration of evaluation methodologies within open-loop contexts.

3. Dataset

3.1. Sensor Setup

The objective of this study is to provide ground truth data that supports end-to-end autonomous driving research in complex urban scenarios. Therefore, high-end collection equipment is essential. In contrast to the sensors used in the nuScenes dataset, the TAD-E2E collection vehicle employs higher-specification sensors, as outlined in Table 2. The collection vehicle utilizes a sensor configuration comprising a surrounding perception system with five LiDARs and six cameras, ensuring comprehensive perception nearly without blind spots. A schematic representation of the collection vehicle can be found in Figure 2. The LiDAR system features a higher-resolution configuration with 128 beams and 64 beams sensor, compared to nuScenes 32 beams sensor. The cameras have an enhanced 1920 x 1080 resolution sensor, compared to nuScenes 1600 x 900. These enhanced sensors result an overall improvements of 4.9x for LiDAR points and 1.44x for image pixels, ensuring a detailed captures and representations of complex scenarios. Additionally, the sensor data provided by TAD-E2E—including subsequent ground truth data—are delivered at a frequency of 10Hz, representing a 5x increase over nuScenes' 2Hz data, thereby enabling more robust time-sequential frame data essential for end-to-end autonomous driving studies. Other sensors for autonomous driving, e.g., IMU, GNSS, and wheel-speed-meter, are also utilized.

3.2. Synchronization & Calibration

All sensors on the vehicle are synchronized using a high-precision PTP (Precision Time Protocol) server, aligning their timestamps with that of the primary LiDAR (the 128-channel LiDAR located at the front left of the vehicle). Camera exposure is triggered based on the LiDAR scan reaching the camera center and serves as the camera's timestamp. The LiDAR timestamps are recorded at the completion of each full rotation of the LiDAR scan, and motion compensation is applied to account for the duration of the LiDAR scan using localization information.

The intrinsic parameters of the cameras are obtained through checkerboard calibration [40], and the extrinsic parameters between the cameras and the LiDAR are acquired through joint calibration using calibration boards within the same field of view. The point clouds from the five LiDARs are merged according to the coordinates of the primary LiDAR.

Sensor	Num	Specifications	Purpose
Camera	6	RGB image @ 1920x1080 resolution, 30Hz, FOV=100°.	Surround View
LiDAR-128	2	Spinning, 128 beams, 10Hz, 360°horizontal FOV @ 0.1°resolution, 40°vertical FOV, 0.3∼230m range @ ±3cm accuracy, with up to 6.9M points per second.	Surround View
LiDAR-64	3	Spinning, 64 beams, 10Hz, 360°horizontal FOV @ 0.6°resolution, 104.2°vertical FOV, 0.1∼60m range @ ±3cm accuracy, with up to 0.768M points per second.	Bind Spots Supplement

Table 2. Sensor Specifications. We utilize 6 x cameras and 5 x LiDARs with high-end specifications deployed in a 360-degree configuration, to provide rich and nearly complete perceptual data coverage without blind spots.

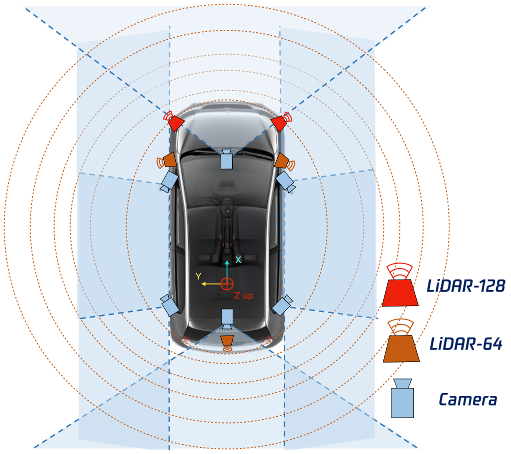


Figure 2. Sensor Deployment. We use 2xLiDAR-128, 3xLiDAR-64 and 6xFOV100°camera. Blue shadow areas show the field of view of cameras. Brown ellipse dots show the field of view of LiDARs.

3.3. Coordinates

This data collection system involves five coordinate systems: the image UV coordinate system, the camera coordinate system, the LiDAR coordinate system (after merging), the ego vehicle coordinate system, and the world coordinate system. Among these, the image UV coordinate system, camera coordinate system, and LiDAR coordinate system are provided based on calibration parameters. The ego vehicle coordinate system has its origin at the center of the rear axle, with the X-axis pointing forward, the Y-axis pointing to the left, and the Z-axis pointing upward. The world coordinate system is represented by an offset-adjusted UTM (Universal Transverse Mercator) coordinate system.

3.4. HD Map & Scenes

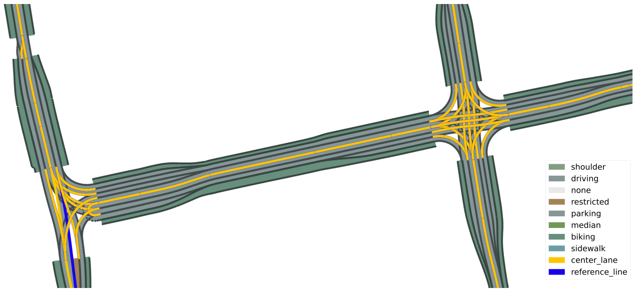


Figure 3. HD Map Sample Illustration.

We provide a high-definition map in the UTM coordinate system, meticulously annotated through advanced techniques to include critical map elements such as lane markings, boundary lines, etc. The map features a variety of intersections, T-junctions, and signal-controlled intersection data. Sample illustrations of the map can be seen in Figure 3. To capture valuable data relevant to complex L4 scenarios, we selected peak traffic periods in bustling urban areas of first-tier cities for data collection, including typical complex traffic scenarios such as subway stations, office buildings, congested roads, and major intersections.

3.5. Ground Truths

3.5.1. Overview

The ultimate goal of end-to-end autonomous driving tasks is to enable effective vehicle behavior decision-making. Theoretically, decision ground truth feedback could satisfy the training requirements for end-to-end networks. However, current SOTA E2E-AD methods still necessitate supervisory signals derived from modular ground truth, as seen in works like UniAD [13], BEVPlanner [18], and SparseDrive [32]. Relying solely on decision ground truth for supervised learning is an area worth exploring, such as 3D bounding boxes for perception, significantly reducing data costs. Nonetheless, given the current stage of advancements in SOTA end-to-end methods and conjunction with the limitations of the nuScenes dataset, we provide not only decision ground truth but also ground truth for individual modules to support modular designs of end-to-end algorithms. This dataset also facilitates the development of standalone module algorithms for autonomous driving tasks, such as perception-based BEV detection and tracking, BEV mapping, and multi-agent trajectory prediction. A visualization of a sample frame containing ground truth data for sensors(camera & LiDAR), detection, tracking, mapping, prediction, and decision-making is displayed in Figure 1.

3.5.2. Planning GT

The decision ground truth refers to the accurate representation of the driving path provided for autonomous driving decision-making algorithms. This data is generated through a combination of high-precision vehicle localization and the organization of temporal data. The high-precision localization is achieved using a sensor fusion algorithm that integrates LiDAR, Inertial Measurement Units (IMU), Global Navigation Satellite System (GNSS), and wheel speed sensors, yielding robust and high-precision localization module.

The temporal aspect of the data is structured into clips of 60 seconds each, with localization signals recorded at a frequency of 100Hz, yielding a total of 6000 frames of temporal localization data per sequence, providing sufficient information for the alignments of sensor data and modular ground truth data. Driving behavior data is collected by multiple professional autonomous driving safety operators. Additionally, to ensure diversity, we resampled the distribution to cover typical and critic driving senarios, including different weather conditions (sunny, rainy, night), driving maneuvers (u-turn, lane change), and safety-critical scenarios (non-protected left turn, traffic light, interactions with other objects). Please refer to Table 3 for details.

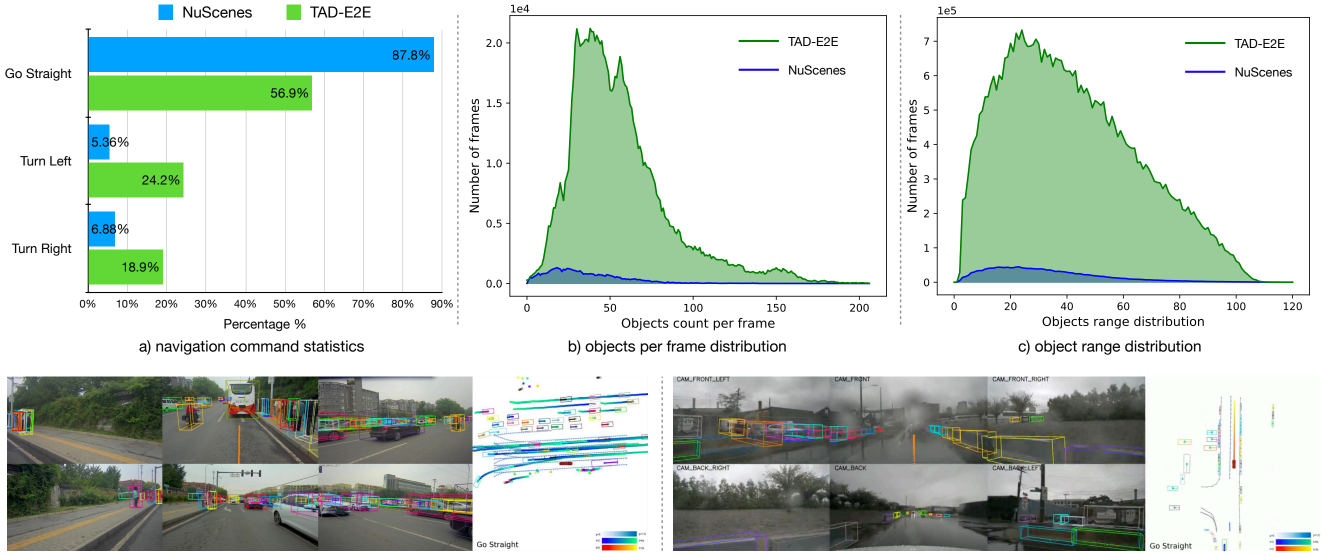
Table 3. Distribution resampling over driving scenes.

Scenes	Sunny	Rainy	Night	U-Turn	Non-protected left turn	Traffic Light	Lane Change	Interactions
Scene Counts	1152	238	276	138	152	175	274	335

3.5.3. Detection, Tracking and Prediction GT

The ground truth for perception refers to the 3D bounding box annotations of surrounding objects within the vehicle's coordinate system. Ideally, this would involve comprehensive manual annotation. However, the vast amount of data in this dataset makes complete manual annotation impractical in terms of both time and cost. Therefore, we adopted a methodology consistent with previous works (such as nuPlan [3], SAM [17], and H-V2X [21]) that employs a cold start approach combined with offline pre-annotation using large models.

In the cold start phase, we performed high-quality manual annotations on approximately 10% of the total dataset. For the offline pre-annotation, we utilized a hybrid training model that integrates both the nuScenes annotated data and the manually annotated 10% of the TAD-E2E dataset. We implemented a multi-model detection framework leveraging two distinct BEVFusion architectures [19, 23] with large backbones, followed by Weighted Boxes Fusion (WBF) [30] for optimal proposal aggregation. This ensemble approach achieved 93.38% 3D mAP@IoU=0.5 and 95.24% BEV mAP@IoU=0.5 on our hold-out validation set. Sequential tracking IDs are performed using offline EKF-based tracking algorithms and further form tracking and prediction ground truths. All pre-annotated data underwent a further human verification process to ensure accuracy and quality.



a) navigation command statistics

b) objects per frame distribution

c) object range distribution

d) TAD-E2E average scene showcase

e) NuScenes average scene showcase

Figure 4. Statistics illustration. a) Navigation command statistics comparison. b) Objects per sample distribution comparison. c) Objects range distribution comparison. d) and f) Scene complexity comparison.

3.5.4. Mapping GT

The mapping ground truth, derived from lanes, boundaries, etc., in the HD MAP, can be accurately projected onto images using vehicle localization and sensor calibration, as shown in Figure 5. Unlike some industry practices, we retained the complete mapping ground truth without filtering for occlusions at the image level, thereby supporting research into mapping predictions in occluded conditions.

3.6. Statistics

We statistically compare TAD-E2E and nuScenes in Figure 4:

- **Figure a)** illustrates navigation command distributions (left turns, right turns, straight movements), categorized by ego-vehicle positional changes over a 3-second horizon. NuScenes exhibits significant homogeneity: 87.8% straight driving versus 5.36% left and 6.88% right turns. In contrast, TAD-E2E demonstrates balanced maneuver diversity, with left/right turns constituting 43.1% of scenarios.
- **Figure b)** quantifies object density (annotated objects per frame), serving as an indicator of environmental complexity. TAD-E2E displays markedly higher complexity than NuScenes, evident in both quantity and mean.
- **Figure c)** analyzes object distance distributions relative to the ego vehicle. TAD-E2E contains a notably higher proportion of medium-to-long range objects compared to NuScenes, with absolute counts substantially exceeding those in NuScenes.
- **Figure d) and f)** present average cases (based on object density) for TAD-E2E and NuScenes, respectively. TAD-E2E features a significantly more complex and challenging autonomous driving scene.

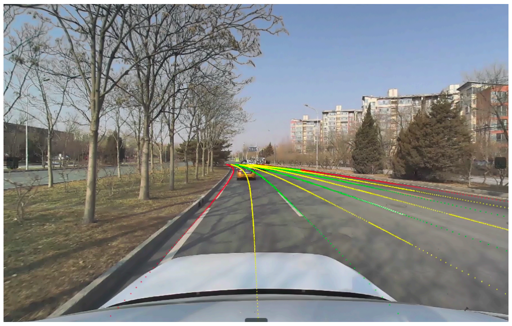


Figure 5. Map ground truth illustration: HD Map project to camera view.

3.7. Privacy

For privacy and data security concerns, all faces and license plates in the dataset have been processed with mosaic anonymization.

4. Experiments

This chapter quantitatively and qualitatively analyzes the value contributions of the proposed TAD-E2E dataset and presents an end-to-end algorithm model that surpasses state-of-the-art (SOTA) performance.

4.1. Baseline Methods

To validate the effectiveness of the proposed TAD-E2E dataset, we selected two representative SOTA end-to-end methods and conducted experimental replication on the TAD-E2E dataset.

Table 4. Quantitative Metrics Comparison of E2E-AD Methods on nuScenes and TAD-E2E datasets. *UniAD evaluate mapping performance using IoU rather than mAP. AD-MLP does not have perception module and thus has no related scores.

Dataset	No.	Method	Detection↑		Tracking↑	Mapping↑	Prediction↓	Planning L2(m)↓				Planning Col(%)↓			
			mAP	NDS	AMOTA	mAP	minADE	1s	2s	3s	avg	1s	2s	3s	avg
nuScenes	a	AD-MLP [38]	–	–	–	–	–	0.20	0.26	0.41	0.29	0.17	0.18	0.24	0.19
	b	UniAD [13]	0.380	0.498	0.359	–*	0.710	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61
	c	SparseDrive [32]	0.415	0.526	0.373	0.556	0.634	0.32	0.63	1.04	0.69	0.01	0.10	0.33	0.14
	d	**SparseFusion (ours)**	0.618	0.673	0.598	0.641	0.472	0.34	0.66	1.06	0.66	0.08	0.11	0.25	0.15
TAD-E2E	e	AD-MLP [38]	–	–	–	–	–	3.34	6.78	10.42	4.37	9.79	11.23	13.47	11.50
	f	UniAD [13]	0.214	0.297	0.189	–*	0.642	1.12	2.02	3.21	2.17	2.56	4.43	6.65	4.54
	g	SparseDrive [32]	0.226	0.320	0.196	0.313	0.556	0.84	1.62	2.42	1.63	0.45	2.27	4.25	2.32
	h	**SparseFusion (ours)**	0.659	0.646	0.628	0.385	0.473	0.35	0.68	1.06	0.70	0.22	0.35	0.79	0.45

4.1.1. Methods Description

1. AD-MLP [38]: AD-MLP proposes a trajectory decision-making algorithm that completely eliminates the need for camera data and perception algorithm modules. By solely utilizing the ego vehicle's state information (including velocity, acceleration, and historical trajectory) along with navigation commands, the MLP network performs reasoning to generate trajectory decisions for subsequent moments. AD-MLP serves as an effective baseline method that appropriately reflects the driving difficulty inherent in the dataset.

2. UniAD [13]: Proposed in 2023, this influential end-to-end autonomous driving model effectively integrates various autonomous driving modules (perception, prediction, mapping, decision-making) using BEV representation and Transformer KQV patterns. It achieves excellent experimental results at both the decision level and within each individual module.

3. SparseDrive [32]: This recently introduced method does not rely on BEV representations and focuses on a sparse scene approach for end-to-end autonomous driving. It features a symmetric system architecture and surpasses SOTA performance metrics across various module indicators.

4.1.2. Metrics

The evaluation metrics adopted are consistent with those used in UniAD [13] and SparseDrive [32]. Due to space limitations, we selected the most representative evaluation metrics for each module, with detailed results provided in Table 4.

4.2. Proposed Method: SparseFusion-E2E

To further investigate whether the simplicity of scenes in nuScenes limits the upper bounds of end-to-end algorithms, we propose the SparseFusion-E2E method, SPF-E2E for short, inspired by SparseDrive [32] and FusionAD [37]. This method combines the advantages of both approaches and designs a multimodal fusion sparse end-to-end network integrating LiDAR and camera data. A schematic diagram of the network model can be seen in Figure 6. We trained and evaluated SPF-E2E on both nuScenes and TAD-E2E datasets, with quantitative results presented in items c and f of Table 4.

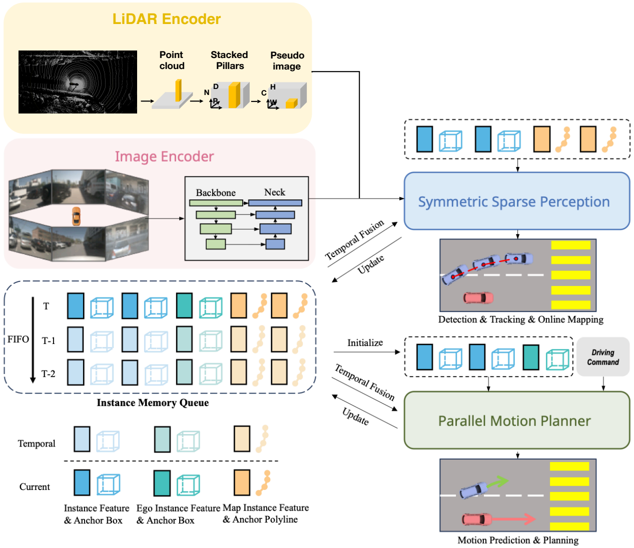

LiDAR Encoder
Point cloud
Stacked Pillars
Pseudo image
Image Encoder
Backbone
Neck
Symmetric Sparse Perception
Detection & Tracking & Online Mapping
Temporal Fusion
Update
FIFO
T
T-1
T-2
Instance Memory Queue
Initialize
Driving Command
Parallel Motion Planner
Motion Prediction & Planning
Temporal
Current
Instance Feature & Anchor Box
Ego Instance Feature & Anchor Box
Map Instance Feature & Anchor Polyline

Figure 6. SparseFusion-E2E model architecture. We incorporate LiDAR modality based on SparseDrive.

4.2.1. Training Details

The hyperparameters for model training on the TAD-E2E dataset are mainly consistent with those used in SparseDrive. The difference lies in our camera's operating frequency of 10 Hz, allowing us to predict a 30-frame trajectory for the ego vehicle and a 60-frame trajectory for surrounding vehicles. Additionally, in the temporal fusion module, we combined features from the previous 20 frames. We set a confidence threshold of 0.2 for object detection and only backpropagated gradients for targets exceeding this threshold. The model was trained with a batch size of 64 on eight L40 (48 GB) GPU cards for approximately ten days.

4.3. Analysis

4.3.1. TAD-E2E v.s. nuScenes

From the pairwise experimental comparisons (a-e, b-f, c-g, d-h), it is evident that the same algorithm model exhibits significant differences in quantitative performance under different datasets. The quantitative scores of various modules on the TAD-E2E dataset are worse than those on nuScenes, confirming our analysis that the TAD-E2E dataset is more challenging and complex, thus providing a higher research ceiling for end-to-end autonomous driving algorithms.

4.3.2. Value of LiDAR-Camera Fusion Over Pure Vision

Comparisons between experiments (c-d and g-h) indicate that the inclusion of LiDAR significantly enhances the performance metrics of various modules (except for the planning module in nuScenes, as analyzed in Section 4.3.3). This underscores the current necessity and value of LiDAR sensors in complex urban scenarios while also highlighting the considerable room for improvement in pure vision methods.

4.3.3. E2E Planning Issue on nuScenes

In addition, from the comparison of experiments (c-d), we observe that improvements from upper modules does not translate to planning module. This finding aligns with observations in AD-MLP [38] and BEVPlanner [18], where good metrics were achieved using only the decision module without perception modules. However, the same issue does not exist in TAD-E2E, where AD-MLP no longer works well.

4.3.4. Modular Contribution Analysis

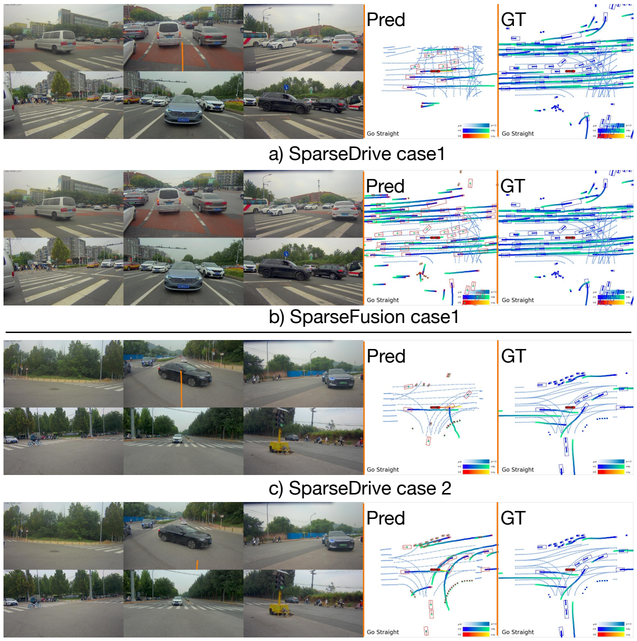
In end-to-end autonomous driving research, evaluating the necessity of ground truth supervision for individual modules holds critical industrial relevance due to its direct impact on data acquisition costs and scalability. Through systematic analysis of the SparseFusion-E2E framework (see Table 5), we demonstrate that each module's supervision is indispensable for achieving viable planning performance. Notably, the perception module proves foundational: its removal results in complete training failure, underscoring its irreplaceable role in feature learning. Consequently, developing methods that eliminate reliance on perception-specific ground truth represents a high-value research direction for cost-efficient, scalable autonomous driving systems.

Table 5. Modular Necessity and Contribution Ablation.

Detection	Tracking	Prediction	Mapping	Planning	L2 AVG↓	Col AVG↓
-	-	-	-	✓	-	-
-	-	-	✓	✓	-	-
✓	-	-	-	✓	0.88	0.66
✓	✓	-	✓	✓	0.73	0.53
✓	✓	✓	-	✓	0.89	0.47
✓	✓	✓	✓	✓	0.70	0.45

4.4. Qualilative Analysis

We conducted a qualitative visual analysis of the pure vision SparseDrive and SparseFusion-E2E models on the TAD-E2E dataset in Figure 7. In case 1, it can be observed that SparseDrive shows significant deviations in detecting the positions of vehicles directly ahead, along with missed detections of occluded vehicles. This results in the ego vehicle planning a trajectory with speed, posing a collision risk. In case 2, due to camera truncation of the field of view, SparseDrive fails to detect a vehicle truncated on the left side and provides incorrect trajectory predictions for the vehicle directly ahead, which leads to a collision risk. In summary, our SparseFusion-E2E model yielded accurate object detection and trajectory prediction results while also planning more reasonable trajectories.



a) SparseDrive case1

b) SparseFusion case1

c) SparseDrive case 2

d) SparseFusion case2

Figure 7. Qualitative comparison of SparseDrive and SparseFusion-E2E on the TAD-E2E dataset: (Left) Surround-view camera inputs, (Middle) Model prediction results, (Right) Ground-truth annotations.

5. Conclusion

This paper introduces TAD-E2E, a large-scale autonomous driving dataset tailored for complex urban environments in end-to-end autonomous driving. TAD-E2E surpasses existing datasets used in state-of-the-art (SOTA) methods by offering a grander scale, more objects, intricate scenes, and diverse driving behaviors. We quantitatively and qualitatively compared TAD-E2E with nuScenes, highlighting its advantages. Our experiments replicated SOTA methods on TAD-E2E, revealing performance declines. Furthermore, we developed a multimodal sparse end-to-end method that outperformed existing approaches. Ablation analysis was conducted to evaluate each module's contribution to the algorithm. Future plans include expanding TAD-E2E to cover more cities and integrating language descriptions to enhance research in the VLM-E2E domain.

References

[1] Mariusz Bojarski. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 3

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 3

[3] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 3, 5

[4] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4

[5] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12878–12895, 2022. 3

[6] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4693–4700. IEEE, 2018. 3

[7] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9329–9338, 2019. 3

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[9] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[11] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36, 2024. 3

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

[13] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 3, 5, 7

[14] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *arXiv preprint arXiv:2406.03877*, 2024. 3

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

[16] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, et al. Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. *arXiv preprint arXiv:2403.04133*, 2024. 3

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[18] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. 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*, pages 14864–14873, 2024. 2, 5, 8

[19] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 5

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[21] Chang Liu, Mingxu Zhu, and Cong Ma. H-v2x: A large scale highway dataset for bev perception. In *European Conference on Computer Vision*, pages 139–157. Springer, 2025. 5

[22] Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. *IEEE Transactions on Intelligent Vehicles*, 2024. 3

[23] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 5

[24] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 3

[25] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 3

[26] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 3

[27] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 3

[28] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021. 3

[29] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 3

[30] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117, 2021. 5

[31] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2, 3

[32] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 3, 5, 7

[33] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 3

[34] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[35] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 3

[36] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 3

[37] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*, 2023. 3, 7

[38] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023. 2, 7, 8

[39] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. 3

[40] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 4