DiverseAD: A Large-Scale Driving Dataset with Diverse Atmospheric Conditions

Haoyu Wang[1,2], Baorui Ma[2], Donglin Di[2], Suhang Xuan[1], Hao Li[2], and Shiliang Zhang[1(✉)]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, Beijing, China
[2] Li Auto, Beijing, China
haoyu.wang@stu.pku.edu.cn, mabaorui2014@gmail.com, donglin.ddl@gmail.com, xsh1217@stu.pku.edu.cn, lihao43@lixiang.com, slzhang.jdl@pku.edu.cn

Abstract. The strong generalization capability to diversified atmospheric conditions is crucial for autonomous driving models. However, models trained on existing datasets struggle with real-world complexity due to the homogeneous weather, limited maneuvers, and insufficient scale in training data. To overcome this issue, we introduce DiverseAD, a large-scale driving dataset comprising 150K scenes that feature a great diversity in atmospheric condition, road types, and driving actions. Based on this dataset, we further propose a novel end-to-end autonomous driving model robust across diverse atmospheric conditions. More specifically, an atmospheric-invariant feature learning mechanism is proposed, which spots and disentangles atmospheric-agnostic features from visual inputs by using driving intent and scene structure cues as stable anchors. Our method thus allows the extracted features to be more invariant to changes in atmospheric conditions. Experiments show that DiverseAD is superior to existing public datasets in diversity, hence is valuable for training and benchmarking. Extensive comparisons also illustrate the superior performance of our proposed method.

Keywords: Autonomous Driving Algorithm · Driving Dataset

1 Introduction

End-to-end autonomous driving has emerged as a promising paradigm that unifies perception, prediction, and planning within a single integrated framework. By directly optimizing the final driving objective, this holistic approach has the potential to deliver improved performance while reducing the error accumulation characteristic of modular pipelines. Formally, the task is defined as planning a safe and comfortable future driving trajectory for the ego-vehicle by processing a combination of historical data and rich sensory inputs from the present. This field has witnessed remarkable advancements, with pioneering

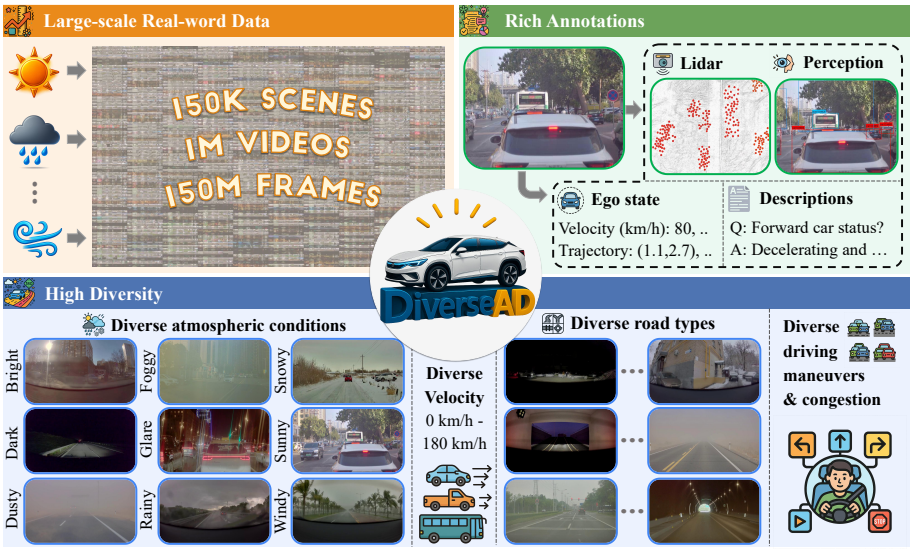


Fig. 1: Overview of DiverseAD. We introduce DiverseAD, a novel driving dataset that provides three key advantages: large-scale real-world data, high environmental diversity, and rich annotations, aiming to address the limitations of existing datasets and advance the development of more reliable autonomous driving systems.

studies [14, 15] establishing strong baselines for unified driving models. More recently, the research landscape has been further transformed by the emergence of novel paradigms, notably diffusion-based frameworks [21,51] and approaches that harness the powerful reasoning capabilities of Vision-Language Models (VLMs) to interpret complex traffic scenarios [17, 43, 46, 49]. A critical factor underpinning these technological advancements is the availability of high-quality datasets, which are fundamentally crucial for the effective pre-training and fine-tuning of these increasingly sophisticated models.

Despite achieving high performance on established benchmarks [6,7,9,34], we observe that these state-of-the-art models [43, 49] exhibit a marked degradation in performance when deployed in real-world scenarios characterized by diverse atmospheric and road conditions. This performance gap poses a significant challenge, making it difficult to fully apply these systems in complex, uncontrolled environments. A critical factor contributing to this fragility is the lack of diversity in the datasets used for training. Existing benchmarks [6, 7, 9, 34] often feature homogeneous settings, primarily capturing data in clear weather, on uniform road typologies such as highways or urban streets, and with a limited repertoire of common driving maneuvers. Furthermore, their scale is often insufficient to cover the long tail of challenging events. Consequently, models trained on such data develop a strong bias towards these over-represented scenarios and struggle to generalize when confronted with the multifaceted and unpredictable nature of real-world driving.

To address these limitations and facilitate the development of more reliable autonomous driving systems, we introduce DiverseAD, a new large-scale dataset that offers three key advantages, as shown in Fig. 1. i) **Large-scale real-world data**: By deploying test vehicles in real-world environments, we capture a vast and varied collection of complex driving data. The dataset amounts to 150K scenes, recorded by multi-view RGB cameras and comprising over 1M videos with more than 150M frames. ii) **High diversity**: DiverseAD features a comprehensive range of atmospheric conditions, including bright sunlight, darkness, dust, fog, glare, rain, snow, and wind. It also covers a wide array of road typologies, from multi-lane highways and urban viaducts to narrow rural roads, and includes a broad spectrum of driving maneuvers and traffic congestion to comprehensively reflect the full range of real-world driving situations. iii) **Rich annotations**: In addition to the multi-view RGB video streams, we provide synchronized LiDAR point clouds, perception data for surrounding environment, and detailed ego-vehicle state information, such as real-time trajectory and velocity. Crucially, the dataset is also enriched with natural language descriptions of each scene and corresponding VQA pairs, enabling novel research directions in Vison-Language-Action intelligence.

Leveraging the proposed DiverseAD dataset, we introduce an end-to-end autonomous driving framework built upon VLMs [2, 42]. To master the challenge of real-world driving under varying atmospheric conditions, we propose a novel Atmospheric-Invariant Feature Learning mechanism, which is designed to guide the model in learning visual features that are robust to environmental variations. Specifically, it utilizes the stable and intrinsic properties of the driving environment, which encompass the underlying scene structure and driving intent cues, as its foundational anchors [29]. By anchoring the learning process on these cues, our framework is compelled to learn atmospheric-agnostic visual representations, effectively disentangling the core driving-relevant information from transient atmospheric effects. This strategy enables our model to achieve reliable planning performance across the diverse atmospheric conditions present in real-world driving scenarios.

We conduct a comprehensive comparison of the proposed DiverseAD dataset against existing benchmarks. To validate our approach, we perform extensive experiments on both DiverseAD and the widely-used nuScenes dataset, which collectively demonstrate the efficacy of our proposed dataset and method. In summary, our contributions are as follows:

- We introduce DiverseAD, a large-scale end-to-end autonomous driving dataset encompassing diverse atmospheric conditions.
- We propose the Atmospheric-Invariant Feature Learning mechanism that enables our model to achieve robust and reliable performance across diverse and challenging environmental conditions.
- Through extensive experiments, we validate the effectiveness of both the proposed dataset for training generalizable models and our accompanying framework.

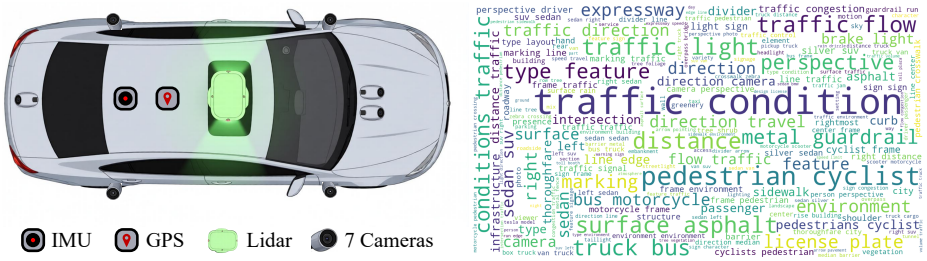
Table 1: The quantitative comparison of our dataset with previously available autonomous driving datasets. "Perc." indicates that this dataset lacks ego-vehicle trajectory annotations and can only be used for perception tasks in autonomous driving scenarios. "Plan." indicates that the annotations of vehicle trajectories can be used for end-to-end autonomous driving tasks.

Dataset	Year	Cameras	Scene Size	Frames	Lidar	Atmospheric Conditions	Task
Foggy Zurich [31]	2018	1	-	3808	✗	Foggy	Perc.
Apolloscape [16]	2018	2	-	100K	✓	Normal	Perc.
A2D2 [11]	2020	6	-	40K	✓	Normal	Perc.
BDD100K [47]	2020	1	100k	-	✗	Rainy, foggy, snowy, dark	Perc.
STF [4]	2020	2	-	1.4M	✓	Rainy, foggy, snowy	Perc.
ONCE [24]	2021	7	-	7M	✓	Normal	Perc.
ACDC [32]	2021	1	-	8012	✗	Rainy, foggy, snowy, dark	Perc.
RADIATE [33]	2021	2	-	5h	✓	Rainy, foggy, snowy, dark	Perc.
K-Radar [26]	2022	4	58	35K	✓	Rainy, foggy, snowy, dark	Perc.
Boreas [5]	2022	1	53	-	✓	Rainy, snowy, dark	Perc.
nuScenes [6]	2019	6	1K	40K	✓	Normal	Plan.
Waymo E2E [6]	2019	8	4K	-	✗	Normal	Plan.
nuPlan [7]	2021	6	15K	-	✓	Normal	Plan.
SceNDD [27]	2022	6	68	-	✓	Normal	Plan.
Ours	2026	7	150K	150M	✓	Rainy, foggy, snowy, dark bright, dusty, glare, sunny, windy	Plan.

2 Related Work

End-to-End Autonomous Driving. End-to-end autonomous driving [14, 15, 17, 18, 21, 37, 46, 49] has gained significant attention for learning direct mappings from raw sensor inputs to vehicle control or planning outputs. This approach has progressed from simple behavioral cloning to sophisticated architectures that leverage structured intermediate representations. Central to modern methods is the use of Bird's-Eye-View (BEV) representations [22], which unify multi-camera inputs and enable scene reasoning. Building upon this, frameworks like UniAD [15] have established powerful unified models that integrate perception, prediction, and planning. To address the uncertainty and multimodality inherent in driving, recent approaches have employed advanced modeling techniques. Generative models [12, 21, 30, 51], such as GenAD [51], use generative schemes to produce realistic trajectories, while diffusion-based methods like DiffusionDrive [21] excel in modeling complex, multi-modal planning distributions. Additionally, the integration of Vision-Language Models (VLMs) [1,2,39], exemplified by frameworks like EMMA [17, 46], OmniDrive [43] and FSDrive [49], enhances scene comprehension and decision-making in complex or long-tail scenarios.

Autonomous Driving Dataset. The development of end-to-end autonomous driving systems, particularly for motion planning, is highly rely on high-quality real-world datasets. Such datasets [6, 7, 9–11, 13, 16, 23, 24, 36, 40, 41, 44, 45, 47] underpin the training of reliable models and serve as benchmarks for comparative evaluation. NAVSIM [9] and nuPlan dataset [7] introduce the closed-loop planning benchmark, enabling realistic system assessment, while datasets such as Lyft Level 5 [13] and Argoverse [44] have been instrumental to motion planning research through their extensive trajectory data and semantic maps. Multimodal



(a) Configuration diagram of the onboard sensors (b) Word clouds of scene description

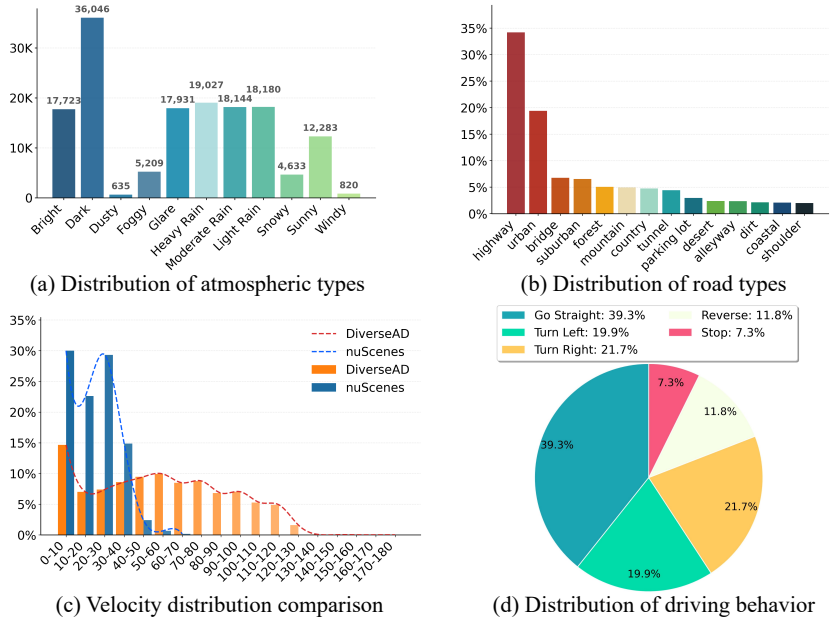
Fig. 2: (a) The key configuration diagram of the onboard sensors on the test vehicles when collecting our DiverseAD dataset. (b) Word clouds of scene description in DiverseAD, including key objects.

datasets such as nuScenes [6] and the Waymo Open Dataset [36] have become the de facto standards for training and evaluating end-to-end driving models. While these benchmarks have been instrumental, they exhibit limited environmental diversity, as most data are collected under ideal or near-ideal conditions. Datasets featuring adverse weather [4, 26, 31, 33], such as ACDC [32] and K-Radar [26], remain limited in atmospheric and road diversity, and their small scale with insufficient annotations constrains their applicability to motion planning. Consequently, there is a pressing need for a new benchmark emphasizing atmospheric and environmental diversity to advance the robustness and reliability of autonomous driving systems.

3 Dataset

3.1 Dataset Construction Process

Sensor Configuration. To facilitate data acquisition under diverse atmospheric conditions, we have equipped our test vehicle with a multi-sensor suite, the configuration of which is illustrated in Fig. 2. Firstly, the Inertial Measurement Unit (IMU) is mounted at the vehicle. By leveraging its internal accelerometer and gyroscope, the IMU measures the vehicle's linear velocity, angular velocity, and acceleration in three-dimensional space. This enables the estimation of the vehicle's pose and its short-term motion trajectory, thereby providing high-frequency motion information. Additionally, the Global Positioning System (GPS) receiver is installed within the vehicle to provide global coordinate positioning data. The GPS also plays a crucial role in timestamp alignment, ensuring temporal synchronization across data streams from the various sensors. Moreover, the Lidar system is integrated as the core sensor for environmental perception. It acquires high-precision 3D point cloud data by emitting laser pulses and measuring the return time, and detects surrounding obstacles, road boundaries, buildings, vehicles, pedestrians, etc., thereby constructing the high-definition (HD) map. Finally, the vehicle is equipped with a total of seven cameras to capture a complete



(a) Distribution of atmospheric types (b) Distribution of road types

(c) Velocity distribution comparison (d) Distribution of driving behavior

Fig. 3: Visualization of the dataset statistics, including distribution of atmospheric conditions, road types, velocity distribution comparison (km/h) with nuScenes dataset [6], and distribution of driving behavior in (a), (b), (c), and (d), respectively. Detailed statistics such as road congestion levels are provided in the supplementary material. Rainy conditions are classified into three levels based on rainfall intensity: light, moderate, and heavy rain.

ego-centric view of the surrounding scene from multiple perspectives. This setup includes two forward-facing cameras with distinct Fields of View (FoV) of 30 degrees and 120 degrees, respectively. These are supplemented by five additional cameras, each with an FoV of 120 degrees, positioned at the front-left, front-right, rear-left, rear-right, and rear of the vehicle.

Data Collection and Cleaning. To ensure a high degree of diversity in the collected DiverseAD dataset, our data acquisition vehicle is operated across a wide spectrum of scenarios. These scenarios encompass diverse road typologies (e.g., highways, bridges, and dirt roads), a range of atmospheric conditions, and varying vehicle speeds. While ensuring safety, a comprehensive set of driving maneuvers, such as acceleration, deceleration, left turns, and right turns, is executed to capture a rich variety of dynamic behaviors.

To ensure data quality, we systematically clean and filter the acquired multimodal raw data, including multiple RGB video streams, LiDAR point clouds, IMU and GPS signals. First, the data is uniformly named and structured for storage, establishing a metadata index containing information such as timestamps, poses, and environmental conditions. Then, temporal and spatial alignment is performed to ensure synchronization and calibration accuracy between multi-

ple sensors. Next, quality checks and anomaly removal are conducted on each modality of data: this included removing point clouds with sparse or severely distorted echoes, GPS data with signal drift, and noisy IMU sequences. Simultaneously, redundant or duplicate segments are eliminated through visual feature and trajectory similarity detection, and data integrity and temporal consistency are checked.

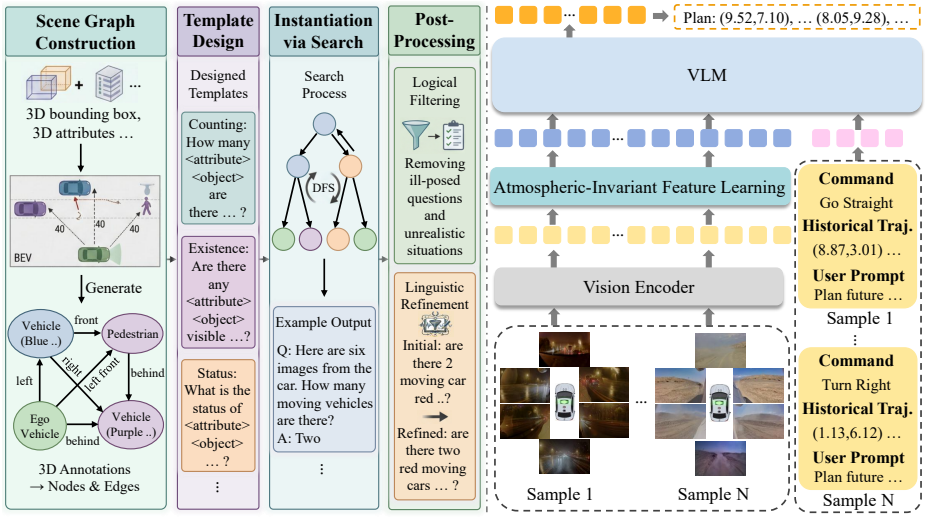
3.2 Dataset Statistics and Visualization

Detailed statistics of the proposed DiverseAD dataset can be found in Tab. 1 and Fig. 3. Specifically, the DiverseAD dataset presents three advantages over existing benchmarks. Firstly, it provides large-scale, real-world driving data, featuring 150K driving scenes captured by a seven-camera array, which totals 1 million videos and 150 million frames. Secondly, as shown in Fig. 3, the dataset boasts an exceptional degree of diversity. It covers nine types of atmospheric conditions (e.g., sunny, rainy, foggy, glare), fourteen road typologies (e.g., highways, tunnels, dirt roads), and a wide range of driving maneuvers and speeds (0-180 km/h) to ensure comprehensive coverage of realistic scenarios. "Bright" denotes a strong light source in daytime conditions, typically sunlight, whereas "glare" refers to intense light sources in low-light or nighttime environments, such as streetlights or the headlights and taillights of nearby vehicles, as shown in Fig. 1. We will provide visualizations of driving videos under different atmospheric conditions in the supplementary material. Finally, the dataset includes comprehensive annotations. In addition to multi-view RGB videos, it provides synchronized 3D LiDAR point clouds, ego-vehicle trajectories and speeds, as well as textual scene descriptions with corresponding Visual Question Answering (VQA) pairs, as illustrated in Fig. 1 and Fig. 2. To get this dataset, please fill dataset request form via `https://drive.google.com/file/d/1fZJshoUlhDIi4hS047iNeld-1a5KnILR/view?usp=sharing`, and contact `cshy2mvvton@outlook.com` with this form.

3.3 Perception and VQA Annotation Process

DiverseAD includes perception data of the surrounding environment, such as 3D lane line coordinates, vehicle 3D bounding boxes, and road boundary. For perception annotation process, we first extract image streams from cameras with synchronized LiDAR point clouds. Annotators then label objects by adding 3D bounding boxes based on point clouds, and label lane lines and road boundaries, while referring to images to verify categories and details. Then quality inspection is applied, including expert review and algorithmic checks, to further ensure the accuracy.

To construct the VQA pairs, we draw inspiration from previous work [28, 43] to create a robust and scalable pipeline. As shown in Fig. 4(a), the process involves the following key steps: (i) Scene Graph Construction: We first build a comprehensive scene graph utilizing existing 3D bounding box annotations. In this abstract representation, individual objects and their specific attributes serve



(a) VQA annotation pipeline for our DiverseAD dataset (b) Overview of our autonomous driving model

Fig. 4: (a) Our VQA annotation process for data includes four key steps. (b) Overview of our end-to-end autonomous driving model. During training, visual inputs from six camera views are utilized. For experiments on DiverseAD, following previous work [6], the closer forward-facing 30-degree camera is employed.

as nodes. To establish the edges, we automatically calculate the spatial relationships between objects by projecting their 3D bounding boxes onto a Bird's-Eye-View (BEV). (ii) Template Design: Based on this structured geometric data, we manually design a diverse set of question templates. These templates are carefully crafted to cover various reasoning tasks, including existence, counting, and status querying, etc. (iii) Instantiation via Search: We then programmatically populate these templates using a depth-first search algorithm. This algorithm assigns parameter values and deduces the ground-truth answers directly from the logical pathways of the scene graph. (iv) Automated Post-Processing: Finally, we apply a rigorous post-processing filtering stage, combining vision-language model [3] analysis with manual inspection. This step eliminates ill-posed questions or incorrect attribute combinations (e.g., "parked pedestrians") and refines unnatural linguistic expressions to ensure the generated text is fluent and logically sound.

4 Methodology

4.1 Formulation

End-to-End Autonomous Driving aims to directly map raw sensor data to future vehicle trajectories. Formally, given a set of N surrounding-view videos or images $\mathcal{V}_t = \{V_t^n\}_{n=1}^N$ at timestep t, along with corresponding visual features

h_t, the vehicle's historical trajectory $\mathcal{H}_t$, and high-level navigation commands $\mathcal{C}_t$, the goal is to train a model $\mathcal{M}$ that predicts a future trajectory in the BEV space. The predicted trajectory, $\hat{W}_t = \{\hat{w}_t^k\}_{k=1}^K$, consists of K waypoints, where each waypoint $\hat{w}_t^k = (\hat{x}_t^k, \hat{y}_t^k)$ represents a future position of the vehicle, *i.e.*,

$$\hat{W}_t = \mathcal{M}(\mathcal{V}_t, \mathcal{H}_t, \mathcal{C}_t). \tag{1}$$

Recent advancements have highlighted the potential of vision-language models for autonomous driving tasks, owing to their extensive world knowledge [17, 43, 46, 49]. Following this paradigm, the trajectory prediction task can be reformulated as a VQA task. The model processes the multi-modal inputs and generates the future trajectory as a sequence of discrete coordinate tokens. Specifically, the BEV space is quantized into a discrete grid, forming a vocabulary of possible waypoint locations. The model is then trained to predict a sequence of tokens corresponding to the ground-truth waypoint locations $W_t = \{w_t^k\}_{k=1}^K$. The learning process is supervised by a vanilla cross-entropy loss $\mathcal{L}_{traj}$ over the sequence of predicted waypoints, *i.e.*,

$$\mathcal{L}_{traj} = -\sum_{k=1}^{K} \log P(w_t^k | \mathcal{V}_t, \mathcal{H}_t, \mathcal{C}_t; \theta), \tag{2}$$

where θ denotes the model parameters.

Visual Feature Learning for Reliable Planning. While the direct end-to-end formulation is powerful, real-world driving scenarios present a significant challenge: the visual observations $\mathcal{V}_t$ are often corrupted by diverse and complex atmospheric conditions. These visual inputs are a composite of two distinct information types: i) **Task-Relevant Invariants:** Information critical for motion planning, such as the scene structure (e.g., road layout, lane markings) and the vehicle's driving intent. These elements are fundamentally invariant to atmospheric conditions. ii) **Task-Irrelevant Variants:** Superficial information that varies with the environment, such as changes in color palettes, reflections, and occlusions caused by rain, fog, or lighting conditions.

The model $\mathcal{M}$ trained solely with $\mathcal{L}_{traj}$ may capture spurious correlations between visual variations and driving behaviors, resulting in poor generalization under unseen conditions. Achieving reliable motion planning requires learning visual features $\bar{h}_t$ disentangled from atmospheric variations and sensitive only to task-relevant invariants, *i.e.*,

$$\bar{h}_t = f(\mathcal{V}_t, \mathcal{H}_t, \mathcal{C}_t; \theta), \tag{3}$$

where f donates the learning process of $\bar{h}_t$.

To this end, we introduce two auxiliary constraints that explicitly structure the visual feature space. These constraints encourage the model to learn robust visual representations by anchoring the learning process f to two invariant factors: driving intent and scene structure. We refer to the corresponding processes as Driving-Intent Anchored Learning and Scene-Structure Anchored Learning.

The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{traj} + \lambda_{int}\mathcal{L}_{int} + \lambda_{str}\mathcal{L}_{str}, \tag{4}$$

where $\mathcal{L}_{int}$ and $\mathcal{L}_{str}$ are the proposed constraints, and λ_{int} and λ_{str} are respective balancing weights. The overall training process is shown in Fig. 4(b).

4.2 Atmospheric-Invariant Feature Learning

Driving-Intent Anchored Learning. This learning objective guides the model to produce similar feature representations for driving scenarios that share the similar underlying driving intent, regardless of superficial visual differences. We found that driving intent is more faithfully represented by the driving patterns of a trajectory (e.g., turning, accelerating) rather than its absolute position.

Let a mini-batch $\mathcal{B}$ consist of N samples. For each sample $i \in \mathcal{B}$, we have a ground-truth trajectory $W_i \in \mathbb{R}^{K\times 2}$ and its corresponding L2-normalized feature representation $z_i \in \mathbb{R}^{L\times D}$ from the vision encoder. L is the number of tokens and D is corresponding token dimension. First, we compute the sequence of instantaneous displacement vectors $v_k^i = w_{k+1}^i - w_k^i$ for each trajectory, which are then normalized to yield a sequence of unit direction vectors $\hat{v}_k^i$, effectively isolating the trajectory's directional pattern from its absolute position.

Next, we define the pairwise intent similarity, S_{int}^{ij}, between two trajectories W_i and W_j as the average cosine similarity of their corresponding direction vector sequences:

$$S_{int}^{ij} = \frac{1}{K-1}\sum_{k=1}^{K-1} \hat{v}_k^i \cdot \hat{v}_k^j. \tag{5}$$

We linearly map this score, which ranges from $[-1, 1]$, to a non-negative weight $w_{int}^{ij} \in [0, 1]$ via $w_{int}^{ij} = (S_{int}^{ij} + 1)/2$.

Finally, the driving-intent anchored learning is formulated to minimize the weighted KL-divergence between the trajectory-derived weights and the feature similarity distribution, *i.e.*,

$$\mathcal{L}_{int} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} w_{int}^{ij}\log\left(\frac{\exp(z_i^\top z_j/\tau)}{\sum_{k=1,k\neq i}^{N}\exp(z_i^\top z_k/\tau)}\right), \tag{6}$$

where $z_i^\top z_j$ is the cosine similarity of the visual features, and τ is a temperature hyperparameter.

Scene-Structure Anchored Learning. While the driving-intent anchored learning enforces invariance to dynamics and intent, the model $\mathcal{M}$ must also learn to be robust to domain-variant visual changes such as rain, fog, or lighting conditions. These variations act as a "style" that corrupts the visual input without altering the underlying "structure" of the driving scene, *e.g.* the road layout, building positions, and static obstacles.

To achieve this, we introduce the scene-structure anchored learning. The intuition is to leverage the robust semantic features from a pre-trained and powerful vision foundation model as an expert, which provides an atmosphere-agnostic representation of the scene's structure. By compelling visual features z to align with the similarity distribution of these expert features, we effectively distill the knowledge of scene structure invariance. For this purpose, we employ the pre-trained DINOv3 [35] as our scene structure expert, denoted as Φ.

For each sample i with visual inputs $\mathcal{V}_i$, we first extract a scene structure representation z_i^{Φ} using the frozen expert: $z_i^{\Phi} = \Phi(\mathcal{V}_i)$. After L2-normalization, we compute the pairwise scene structure similarity $S_{str}^{ij} = (z_i^{\Phi})^{\top} z_j^{\Phi}$. This similarity is then converted into a soft weight w_{str}^{ij} using a Gaussian kernel:

$$w_{str}^{ij} = \exp\left(-\frac{(1 - S_{str}^{ij})^2}{\sigma_{str}^2}\right), \tag{7}$$

where σ_{str} is a kernel bandwidth hyperparameter.

The Scene-Structure Anchored Learning objective then uses these expert-derived weights to guide the primary visual feature learning process, *i.e.*,

$$\mathcal{L}_{str} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} w_{str}^{ij} \log\left(\frac{\exp(z_i^{\top} z_j/\tau)}{\sum_{k=1,k\neq i}^{N}\exp(z_i^{\top} z_k/\tau)}\right). \tag{8}$$

By minimizing $\mathcal{L}_{str}$, we regularize the feature space, promoting robustness against visual domain shifts.

Based on the proposed Atmospheric-Invariant Feature Learning mechanism, the overall training objective is shown in Eq. (4). During inference, the model can directly predict future trajectories based on historical trajectories and current observations without additional overhead.

5 Experiments

5.1 Experimental Settings

Datasets. We first evaluate our model on the DiverseAD dataset, comparing its performance with publicly available methods [43, 46, 49] for trajectory planning. DiverseAD is partitioned into training and testing subsets with an 8:2 ratio. Following the same settings used in the previous methods, given the historical trajectory of the past two seconds and the RGB images of the surrounding 6 views at the moment, the model predicts the trajectory for the next 3 seconds. Comparison with more publicly available methods are provided in the supplementary material. Furthermore, to demonstrate the effectiveness of our proposed DiverseAD dataset, we also compare the model trained on it with previous methods [15, 43, 51, 52] on the nuScenes dataset [6] following the same settings.

Evaluation Metrics. For experiments on the DiverseAD dataset, we first evaluate the L2 displacement error under different atmospheric conditions. Due to

Table 2: Quantitative comparison on the DiverseAD dataset, including results on all atmospheric conditions. * indicates that fine-tuning is performed on our DiverseAD dataset. Avg. represents the average result of the metric. For the VQA experiment, we compare with previous VLM-based methods finetuned on our dataset. Due to page limitation, comparison with more publicly available methods are provided in the supplementary material. The best results are highlighted in **bold**.

Method	Reference	All - L2 (m) ↓				VQA Acc. (%) ↑
		1s	2s	3s	Avg.	
UniAD [15]	CVPR 23	15.40	31.11	46.99	31.17	No VQA functionality
UniAD* [15]	CVPR 23	1.04	2.06	3.23	2.11	No VQA functionality
VAD [18]	ICCV 23	15.01	30.02	44.97	30.00	No VQA functionality
VAD* [18]	ICCV 23	0.72	1.67	2.69	1.69	No VQA functionality
OpenEMMA [46]	arXiv 24	26.01	34.66	40.92	33.86	-
OpenEMMA* [46]	arXiv 24	0.71	1.60	2.67	1.66	60.28
OmniDrive [43]	CVPR 25	15.20	30.49	45.86	30.52	-
OmniDrive* [43]	CVPR 25	0.61	1.45	2.50	1.52	64.51
FSDrive [49]	NeurIPS 25	23.71	26.25	26.54	25.50	-
FSDrive* [49]	NeurIPS 25	0.59	1.38	2.34	1.44	65.33
Ours	-	**0.53**	**1.21**	**1.99**	**1.24**	**72.68**

page limitation, evaluations of more metrics are provided in the supplementary material. For experiments on the nuScenes dataset, following previous work [52], we use L2 displacement error and collision rate, and calculate them following ST-P3 [14]. Implementation details are provided in the appendix.

5.2 Comparison on DiverseAD Dataset

Tab. 2 and Tab. 3 presents the quantitative comparison under different atmospheric conditions on the DiverseAD dataset. For the previous methods, we test the performance of the initial models, and the performance after fine-tuning on our DiverseAD dataset. In the supplementary materials, we also test the performance of using a state-of-the-art image restoration model [48, 50] to preprocess image inputs before sending them into the vision encoder. We found that previous methods experience substantial performance degradation under diverse atmospheric conditions, such as in dark environments. In contrast, our proposed method demonstrates remarkable reliability, maintaining high performance across a wide range of challenging atmospheric environments.

5.3 Comparison on nuScenes Dataset

Tab. 4 shows the comparison on the nuScenes dataset. No additional ego status besides historical trajectories is used in the evaluation to ensure a fair comparison. To demonstrate the effectiveness of the proposed DiverseAD dataset, we adopt the model architecture as previous work [43,49] and train it on DiverseAD. The results show that introducing the DiverseAD dataset effectively improves the model's performance on the end-to-end planning task.

Table 3: Quantitative comparison on the DiverseAD dataset, including nine atmospheric conditions. Due to page limitation, VQA results under different atmospheric conditions are provided in the supplementary material.

Method	Bright - L2 (m) ↓				Dark - L2 (m) ↓				Dusty - L2 (m) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD [15]	28.48	28.83	29.62	28.98	50.37	52.62	55.91	52.97	38.55	39.72	41.49	39.92
UniAD* [15]	1.03	2.05	3.20	2.09	1.09	2.17	3.39	2.22	0.86	1.74	2.70	1.77
VAD [18]	21.31	25.82	30.58	25.90	38.54	47.69	57.07	47.77	31.81	39.29	47.07	39.37
VAD* [18]	0.70	1.64	2.65	1.66	0.86	1.88	2.96	1.90	0.60	1.42	2.27	1.43
OpenEMMA [46]	23.66	29.19	32.27	28.37	48.42	63.26	72.46	61.38	36.68	50.64	62.06	49.79
OpenEMMA* [46]	0.71	1.58	2.65	1.65	0.74	1.70	2.96	1.80	0.53	1.23	2.14	1.30
OmniDrive [43]	12.70	25.50	38.45	25.55	23.38	46.88	70.42	46.89	19.43	39.11	58.93	39.16
OmniDrive* [43]	0.61	1.55	2.83	1.66	0.66	1.56	2.76	1.66	0.47	1.07	1.86	1.14
FSDrive [49]	22.72	24.52	25.55	24.26	26.97	32.86	34.10	31.31	19.08	22.23	24.50	21.94
FSDrive* [49]	0.63	1.46	2.47	1.52	0.66	1.48	2.65	1.60	0.47	1.03	1.75	1.09
Ours	**0.50**	**1.18**	**2.00**	**1.23**	**0.56**	**1.24**	**2.09**	**1.30**	**0.45**	**0.97**	**1.70**	**1.04**
Method	Foggy - L2 (m) ↓				Glare - L2 (m) ↓				Rainy - L2 (m) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD [15]	27.60	28.09	28.81	28.17	29.68	30.71	31.99	30.79	32.41	34.15	36.37	34.31
UniAD* [15]	0.89	1.72	2.75	1.79	1.27	2.52	3.95	2.58	1.00	1.99	3.12	2.04
VAD [18]	16.19	20.12	24.35	20.22	22.07	26.62	31.35	26.68	31.58	38.51	45.66	38.58
VAD* [18]	0.61	1.41	2.29	1.44	0.89	2.02	3.27	2.06	0.68	1.59	2.56	1.61
OpenEMMA [46]	17.58	23.15	26.62	22.45	20.39	26.12	30.63	25.71	35.26	46.43	53.65	45.11
OpenEMMA* [46]	0.60	1.38	2.36	1.45	0.89	2.01	3.04	1.98	0.68	1.52	2.60	1.60
OmniDrive [43]	11.35	22.82	34.45	22.87	13.13	26.56	39.89	26.53	17.69	35.39	53.14	35.41
OmniDrive* [43]	0.53	1.30	2.66	1.36	0.74	1.84	2.92	1.83	0.57	1.31	2.29	1.39
FSDrive [49]	11.09	12.69	13.79	12.53	19.95	20.88	21.05	20.63	35.04	37.68	37.83	36.85
FSDrive* [49]	0.50	1.21	2.12	1.28	0.65	1.72	2.61	1.66	0.56	1.26	2.16	1.33
Ours	**0.44**	**1.02**	**1.70**	**1.05**	**0.58**	**1.57**	**2.39**	**1.51**	**0.51**	**1.14**	**1.90**	**1.19**
Method	Snowy - L2 (m) ↓				Sunny - L2 (m) ↓				Windy - L2 (m) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD [15]	28.54	29.32	30.32	29.39	28.01	28.65	29.70	28.78	29.17	30.04	31.15	30.13
UniAD* [15]	1.15	2.26	3.54	2.32	0.94	1.97	3.15	2.02	0.98	1.93	3.03	1.98
VAD [18]	17.02	20.59	24.37	20.66	19.75	24.17	28.86	24.26	21.50	25.50	29.64	25.55
VAD* [18]	0.76	1.79	2.89	1.81	0.63	1.56	2.57	1.59	0.74	1.69	2.73	1.72
OpenEMMA [46]	17.17	22.34	25.84	21.78	15.80	21.24	25.33	20.79	22.12	27.99	30.96	27.02
OpenEMMA* [46]	0.70	1.53	2.49	1.57	0.63	1.45	2.54	1.54	1.18	2.50	3.63	2.44
OmniDrive [43]	10.61	21.34	32.25	21.40	12.48	25.03	37.76	25.09	11.36	22.86	34.34	22.85
OmniDrive* [43]	0.68	1.54	2.62	1.61	0.59	1.38	2.39	1.45	0.53	1.32	2.23	1.36
FSDrive [49]	12.66	14.19	15.06	13.97	17.59	19.11	19.83	18.84	17.25	18.60	19.50	18.45
FSDrive* [49]	0.67	1.48	2.46	1.54	0.59	1.37	2.36	1.44	0.49	1.22	2.10	1.27
Ours	**0.61**	**1.32**	**2.09**	**1.34**	**0.54**	**1.18**	**2.03**	**1.25**	**0.45**	**1.12**	**1.88**	**1.15**

5.4 Ablation Study

Effect of Driving-Intent Anchored Learning.

Tab. 5 presents a comprehensive quantitative ablation study evaluating the driving-intent anchored learning module. Specifically, the first row reports the performance when the intent loss $\mathcal{L}_{int}$ is entirely omitted during training. The second row explores an alternative similarity metric, where the Average Displacement Error (ADE) is employed to measure the proximity of driving intents among samples within a mini-batch, rather than our proposed direction-based approach. The empirical results underscore the superiority of using direction vectors for intent similarity modeling and confirm that driving-intent anchored

Table 4: Quantitative comparison of the end-to-end planning task with previous methods on the nuScenes [6] dataset. The best results are highlighted in **bold** and the second-best results are underlined. Comparison with more methods are provided in the supplementary material.

Method	Reference	L2 (m) ↓				Collision Rate (%) ↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3 [14]	ECCV 22	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [15]	CVPR 23	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
OccNet [38]	ICCV 23	1.29	2.13	2.99	2.13	0.21	0.59	1.37	0.72
VAD [18]	ICCV 23	0.41	0.70	1.05	0.72	0.07	0.18	0.43	0.23
BEV-Planner [20]	CVPR 24	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59
PPAD [8]	ECCV 24	0.31	0.56	0.87	0.58	0.08	0.12	0.38	0.19
GenAD [51]	ECCV24	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19
LAW [19]	ICLR 25	0.26	0.57	1.01	0.61	0.14	0.21	0.54	0.30
OmniDrive [43]	CVPR 25	**0.15**	0.36	0.70	0.40	0.06	0.27	0.72	0.35
World4Drive [52]	ICCV 25	0.23	0.47	0.81	0.50	0.02	0.12	**0.33**	**0.16**
FSDrive [49]	NeurIPS 25	0.28	0.52	0.80	0.53	**0.01**	0.11	0.57	0.23
Ours	-	0.22	**0.35**	**0.58**	**0.39**	**0.01**	**0.06**	0.41	**0.16**

learning significantly enhances the representation capacity of the model. To provide further qualitative evidence, a visualization of the learned features $\bar{h}_t$ is provided in the supplementary material, demonstrating their enhanced discriminative power.

Effect of Scene-Structure Anchored Learning. Tab. 6 presents the quantitative analysis of the effect of scene-structure anchored learning. The first row reports the results when the $\mathcal{L}_{str}$ constraint is removed during training, while the second row corresponds to the Gram-matrix-based similarity computed from DINOv3 [35] features among different samples. The results demonstrate that incorporating scene-structure anchoring with DINOv3 features more effectively guides the model to learn visual representations closely aligned with underlying scene structures. We also include a visual comparison of visual features with and without this mechanism in the supplementary material.

Comparison with Contrastive Learning Form. To evaluate the impact of the loss formulation, we implement a contrastive learning variant that utilizes a cosine similarity threshold of 0.5 for pair construction. Experimental results in Tab. 7 show that replacing our objective with this contrastive loss results in an increase in the average L_2 distance—from 1.24 to 1.33 for scene structure and to 1.38 for driving intent. This performance decay underscores the suboptimality of using hard contrastive pairs in this context. Ultimately, the comparison demonstrates that our soft-weighted KL terms are better suited for the trajectory prediction task, as they more effectively regularize the latent space by accounting for varying degrees of similarity.

Analysis of Scene-Structure Expert Selection. Tab. 8 presents the ablation study for the proposed scene structure expert. To evaluate its performance, we conduct a comparative analysis against state-of-the-art vision foundation models, specifically CLIP [29] and DINOv2 [25], as alternative backbones for structural feature extraction. The experimental results reveal that our design con-

Table 5: Quantitative analysis for the effect of the proposed driving-intent anchored learning.

Model	All - L2 (m) ↓			
	1s	2s	3s	Avg.
w/o Intent	0.62	1.37	2.18	1.39
Intent with ADE	0.62	1.35	2.18	1.38
Ours	**0.53**	**1.21**	**1.99**	**1.24**

Table 6: Quantitative analysis for the effect of the proposed scene-structure anchored learning.

Model	All - L2 (m) ↓			
	1s	2s	3s	Avg.
w/o Struc.	0.63	1.39	2.23	1.42
Struc. with Gram	0.60	1.32	2.13	1.35
Ours	**0.53**	**1.21**	**1.99**	**1.24**

Table 7: Comparison with contrastive learning (CL) on drving intent and scene struture.

Model	All - L2 (m) ↓			
	1s	2s	3s	Avg.
CL on Intent	0.59	1.33	2.21	1.38
CL on Struc.	0.58	1.29	2.13	1.33
Ours	**0.53**	**1.21**	**1.99**	**1.24**

Table 8: Analysis of the scene structure expert selection, comparing with CLIP and DINOv2.

Model	All - L2 (m) ↓			
	1s	2s	3s	Avg.
Using CLIP	0.59	1.32	2.20	1.37
Using DINOv2	0.56	1.25	2.06	1.29
Ours	**0.53**	**1.21**	**1.99**	**1.24**

sistently outperforms these general-purpose visual models in capturing driving-related spatial layouts. This superiority demonstrates that our specialized scene structure expert is more effective at encoding the complex geometric constraints and topological relationships inherent in autonomous driving environments.

6 Conclusion

We present DiverseAD, a pioneering large-scale autonomous driving dataset featuring 150,000 unique scenarios. DiverseAD spans a vast spectrum of atmospheric conditions, complex road topologies, and heterogeneous driving behaviors, effectively addressing the diversity gaps inherent in current benchmarks. Leveraging this rich data resource, we propose a novel end-to-end autonomous driving framework grounded in Vision-Language Models (VLMs). A key innovation of our approach is the Atmospheric-Invariant Feature Learning mechanism, designed to extract robust representations across volatile environmental conditions. Comprehensive experiments on both DiverseAD and the nuScenes benchmark demonstrate that our dataset significantly enhances model generalization, while our framework sets a new standard for driving performance in challenging real-world settings.

Acknowledgments

This work was supported in part by under Grant 2023-JCJQ-LA-001-088, in part by the Natural Science Foundation of China under Grant U20B2052, Grant 61936011, in part by under 2025ZD1601300, in part by the Okawa Foundation Research Award, in part by the Ant Group Research Fund, and in part by the Kunpeng&Ascend Center of Excellence, Peking University.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11682–11692 (2020)
5. Burnett, K., Yoon, D.J., Wu, Y., Li, A.Z., Zhang, H., Lu, S., Qian, J., Tseng, W.K., Lambert, A., Leung, K.Y., et al.: Boreas: A multi-season autonomous driving dataset. The International Journal of Robotics Research **42**(1-2), 33–42 (2023)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
7. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021)
8. Chen, Z., Ye, M., Xu, S., Cao, T., Chen, Q.: Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In: European Conference on Computer Vision. pp. 239–256. Springer (2024)
9. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. Advances in Neural Information Processing Systems **37**, 28706–28719 (2024)
10. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
11. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., et al.: A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320 (2020)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., Ondruska, P.: One thousand and one hours: Self-driving motion prediction dataset. In: Conference on Robot Learning. pp. 409–418. PMLR (2021)
14. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
15. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)

16. Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., Yang, R.: The apolloscape dataset for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 954–960 (2018)
17. Hwang, J.J., Xu, R., Lin, H., Hung, W.C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., et al.: Emma: End-to-end multimodal model for autonomous driving. arXiv preprint arXiv:2410.23262 (2024)
18. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
19. Li, Y., Fan, L., He, J., Wang, Y., Chen, Y., Zhang, Z., Tan, T.: Enhancing end-to-end autonomous driving with latent world model. arXiv preprint arXiv:2406.08481 (2024)
20. Li, Z., Yu, Z., Lan, S., Li, J., Kautz, J., Lu, T., Alvarez, J.M.: 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. pp. 14864–14873 (2024)
21. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12037–12047 (2025)
22. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. arXiv preprint arXiv:2205.13542 (2022)
23. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
24. Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al.: One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037 (2021)
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
26. Paek, D.H., Kong, S.H., Wijaya, K.T.: K-radar: 4d radar object detection for autonomous driving in various weather conditions. Advances in Neural Information Processing Systems **35**, 3819–3829 (2022)
27. Prabu, A., Ranjan, N., Li, L., Tian, R., Chien, S., Chen, Y., Sherony, R.: Scendd: A scenario-based naturalistic driving dataset. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). pp. 4363–4368. IEEE (2022)
28. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4542–4550 (2024)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
31. Sakaridis, C., Dai, D., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: Proceedings of the european conference on computer vision (ECCV). pp. 687–704 (2018)
32. Sakaridis, C., Dai, D., Van Gool, L.: Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10765–10775 (2021)
33. Sheeny, M., De Pellegrin, E., Mukherjee, S., Ahrabian, A., Wang, S., Wallace, A.: Radiate: A radar dataset for automotive perception in bad weather. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–7. IEEE (2021)
34. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: European conference on computer vision. pp. 256–274. Springer (2024)
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
36. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
37. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8795–8801. IEEE (2025)
38. Tong, W., Sima, C., Wang, T., Chen, L., Wu, S., Deng, H., Gu, Y., Lu, L., Luo, P., Lin, D., et al.: Scene as occupancy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8406–8415 (2023)
39. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
40. Wang, H., Tang, H., Di, D., Zhang, Z., Zuo, W., Gao, F., Ma, S., Zhang, S.: Mosa: Motion-coherent human video generation via structure-appearance decoupling. arXiv preprint arXiv:2508.17404 (2025)
41. Wang, H., Zhang, Z., Di, D., Zhang, S., Zuo, W.: Mv-vton: Multi-view virtual try-on with diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7682–7690 (2025)
42. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
43. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the computer vision and pattern recognition conference. pp. 22442–22452 (2025)
44. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
45. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are vlms ready for autonomous driving? an empirical study from the reliability, data

and metric perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6585–6597 (October 2025)
46. Xing, S., Qian, C., Wang, Y., Hua, H., Tian, K., Zhou, Y., Tu, Z.: Openemma: Open-source multimodal model for end-to-end autonomous driving. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1001–1009 (2025)
47. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
48. Zamfir, E., Wu, Z., Mehta, N., Tan, Y., Paudel, D.P., Zhang, Y., Timofte, R.: Complexity experts are task-discriminative learners for any image restoration (2024)
49. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)
50. Zhang, Z., Wang, H., Liu, S., Wang, X., Lei, L., Zuo, W.: Self-supervised high dynamic range imaging with multi-exposure images in dynamic scenes. arXiv preprint arXiv:2310.01840 (2023)
51. Zheng, W., Song, R., Guo, X., Zhang, C., Chen, L.: Genad: Generative end-to-end autonomous driving. In: European Conference on Computer Vision. pp. 87–104. Springer (2024)
52. Zheng, Y., Yang, P., Xing, Z., Zhang, Q., Zheng, Y., Gao, Y., Li, P., Zhang, T., Xia, Z., Jia, P., et al.: World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28632–28642 (2025)