Towards Viewpoint-Robust End-to-End Autonomous Driving with 3D Foundation Model Priors

Hiroki Hashimoto[1] Hiromichi Goto[2] Hiroyuki Sugai[3]
Hiroshi Kera[4,6] Kazuhiko Kawamoto[5]
[1,4,5]Chiba University, [2,3]SUZUCA.AI, [6]National Institute of Informatics
[1]hiroki.hashimoto@chiba-u.jp, [2]goto@suzuca.ai, [3]sugai@suzuca.ai
[4]kera@chiba-u.jp, [5]kawa@faculty.chiba-u.jp

Abstract

Robust trajectory planning under camera viewpoint changes is important for scalable end-to-end autonomous driving. However, existing models often depend heavily on the camera viewpoints seen during training. We investigate an augmentation-free approach that leverages geometric priors from a 3D foundation model. The method injects per-pixel 3D positions derived from depth estimates as positional embeddings and fuses intermediate geometric features through cross-attention. Experiments on the VR-Drive camera viewpoint perturbation benchmark show reduced performance degradation under most perturbation conditions, with clear improvements under pitch and height perturbations. Gains under longitudinal translation are smaller, suggesting that more viewpoint-agnostic integration is needed for robustness to camera viewpoint changes.

1. Introduction

End-to-end autonomous driving jointly optimizes the full pipeline from sensor input to trajectory planning within a single deep neural network. This approach avoids information loss and error accumulation inherent in conventional modular pipelines [3, 5–7, 10, 12, 17, 19, 23]. However, existing end-to-end autonomous driving models depend heavily on the camera viewpoints present in the training data, and their trajectory planning accuracy degrades under unseen camera viewpoints [3, 16]. When autonomous driving systems are deployed across different vehicle platforms, camera viewpoints inevitably vary because of differences in vehicle types and sensor configurations. Adapting existing models to such viewpoint changes typically requires additional data collection and retraining for each platform. Such adaptation is operationally expensive. Improving robustness to unseen camera viewpoints is therefore critical for scalable end-to-end autonomous driving systems.

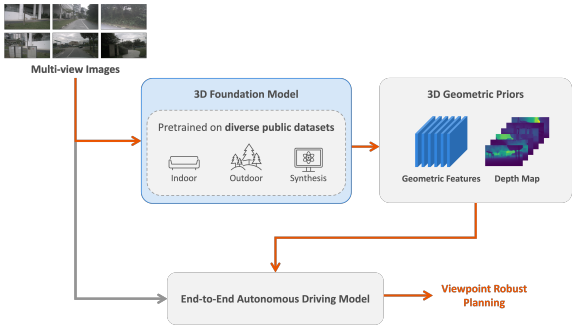

Figure 1. Overview of the proposed method. We extract geometric features and depth estimates from a 3D foundation model and integrate them into an end-to-end autonomous driving model to improve robustness against camera viewpoint changes.

To address this challenge, prior work has mainly improved viewpoint robustness by increasing viewpoint diversity during training. Stelzer et al. [16] show that sensor configuration mismatch degrades end-to-end driving performance, while multi-configuration training improves robustness. VR-Drive [3] improves generalization to camera viewpoint perturbations through novel view synthesis with 3D Gaussian Splatting (3DGS) [8]. However, viewpoint augmentation may not generalize well beyond the camera configurations covered during training. Augmentation-free approaches are therefore a promising direction for improving robustness to camera viewpoint changes.

In this work, we investigate an augmentation-free approach for improving robustness to camera viewpoint changes in trajectory planning by leveraging geometric priors from a 3D foundation model (Fig. 1). Recent 3D foundation models [13, 15, 18, 20] acquire generalizable 3D geometric knowledge through large-scale pretraining and have been applied to downstream driving tasks [19, 22, 23]. To incorporate such geometric knowledge into an end-to-end autonomous driving model, we introduce two modules

(Fig. 2). The first module, 3D Spatial Encoder, computes per-pixel 3D position from DA3 [13] depth estimates and camera parameters, and injects these positions as positional embeddings into image features. The second module, Geometric Feature Fusion, fuses DA3 intermediate features into image features via cross-attention. These two modules incorporate transferable 3D geometric cues into the driving model.

Evaluation on the VR-Drive viewpoint perturbation benchmark [3] shows that the proposed method reduces performance degradation under most perturbation conditions, with particularly clear gains under pitch and height perturbations. Performance gains are smaller under longitudinal translation, indicating that some viewpoint changes remain challenging. These results suggest the importance of a more viewpoint-agnostic integration design for fully leveraging 3D foundation models.

2. Related Work

2.1. Viewpoint Robustness in End-to-End Autonomous Driving

Existing end-to-end autonomous driving methods can be broadly grouped into perception-based approaches [3, 5–7, 12, 17] and latent world model-based approaches [10, 19, 23]. Perception-based methods [3, 5–7, 12, 17] achieve high planning performance by learning auxiliary perception tasks such as 3D object detection and map construction, but they require large amounts of high-quality 3D annotations. Latent world model-based methods [10, 19, 23] learn future scene latent representations through self-supervised learning and require less manual annotation, making them suitable for scalable data-driven training.

For robustness to camera viewpoint changes, Stelzer et al. [16] investigate the impact of sensor configurations on end-to-end driving performance in the CARLA simulator [4]. Their study shows that mismatched sensor configurations between training and testing degrade performance, while multi-configuration training improves robustness. VR-Drive [3] integrates feed-forward 3DGS-based novel view synthesis as an auxiliary task into an end-to-end framework. During training, images rendered from randomly sampled camera extrinsics are used together with the original views, increasing viewpoint diversity and improving robustness to camera viewpoint perturbations. VR-Drive also provides a camera viewpoint perturbation benchmark based on nuScenes [1].

While these approaches improve robustness by augmenting training viewpoints, our work takes an augmentation-free approach based on geometric priors from a 3D foundation model.

2.2. 3D Foundation Models for Autonomous Driving

Recently, 3D foundation models have been developed through large-scale pretraining for 3D geometric tasks such as depth estimation and scene reconstruction [13, 15, 18, 20]. Among them, DA3 [13] uses DINOv2 [14] as its backbone and takes camera intrinsic and extrinsic parameters as camera tokens. This token-based design enables geometrically consistent depth estimation across multiple views and strong generalization to diverse viewpoints and domains.

3D foundation models have also been applied to downstream autonomous driving tasks. For object detection, DetAny3D [22] integrates UniDepth-pretrained 2D features with SAM [9] features and achieves robust zero-shot monocular 3D object detection, including under unseen camera viewpoints. For trajectory planning, World4Drive [23] constructs 3D positional encodings from Metric3D [20] depth estimates, while WorldRFT [19] fuses intermediate VGGT [18] features via cross-attention. These methods improve planning performance, but they do not explicitly address robustness to camera viewpoint changes. Prior work has not leveraged the geometric knowledge of 3D foundation models to improve such robustness in trajectory planning.

3. Method

3.1. Problem Setting

Let $\mathbf{I}_t = \{I_t^1, I_t^2, \dots, I_t^N\}$ denote the set of N surround-view images captured at time step t. Each camera i has an intrinsic matrix $\mathbf{K}^i \in \mathbb{R}^{3\times3}$ and extrinsic parameters ($\mathbf{R}^i \in \mathbb{R}^{3\times3}$, $\mathbf{d}^i \in \mathbb{R}^3$). We denote the full camera configuration by

$$\mathcal{C} = \{\mathbf{K}^i, \mathbf{R}^i, \mathbf{d}^i\}_{i=1}^N. \tag{1}$$

An end-to-end autonomous driving model F takes a temporal sequence of K frames, $\{\mathbf{I}_{t-K+1}, \dots, \mathbf{I}_t\}$, along with the camera configuration $\mathcal{C}$, and predicts future ego-vehicle waypoints:

$$\hat{\tau} = F(\{\mathbf{I}_{t-K+1}, \dots, \mathbf{I}_t\}, \mathcal{C}), \tag{2}$$

where $\hat{\tau} = \{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2, \dots, \hat{\mathbf{p}}_T\}$ and each $\hat{\mathbf{p}}_j = (x_j, y_j) \in \mathbb{R}^2$ denotes the planned ego-vehicle position at future time j.

In our setting, training data are collected under a fixed camera configuration $\mathcal{C}_{\text{train}}$, whereas test-time deployment involves a different camera configuration $\mathcal{C}_{\text{test}} \neq \mathcal{C}_{\text{train}}$. A change in camera configuration shifts the input distribution from $p(\mathbf{I} \mid \mathcal{C}_{\text{train}})$ to $p(\mathbf{I} \mid \mathcal{C}_{\text{test}})$. As a result, the predictor F trained under $\mathcal{C}_{\text{train}}$ may produce degraded trajectories $\hat{\tau}$ under $\mathcal{C}_{\text{test}}$. Our goal is to improve robustness to such viewpoint shifts without requiring additional training data from the test platform.

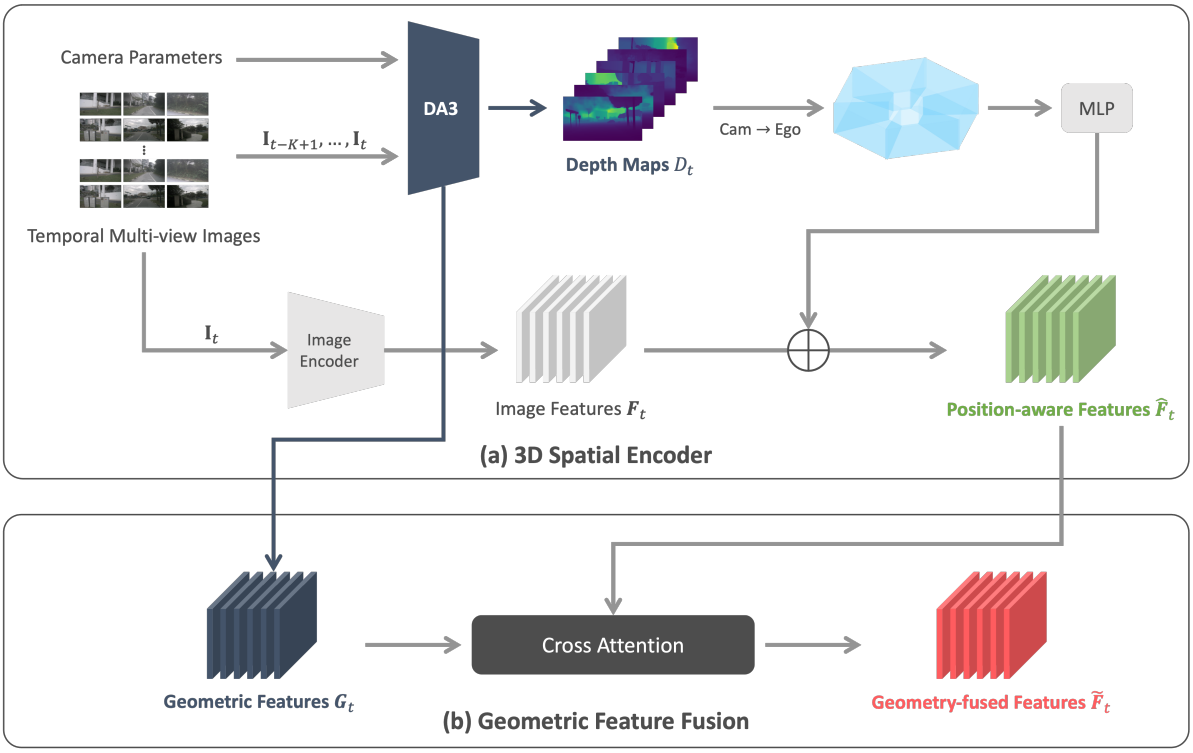

Figure 2. Architecture of the proposed method. (a) 3D Spatial Encoder: computes depth-derived 3D positions from DA3 and camera parameters, and injects them as positional embeddings into image features. (b) Geometric Feature Fusion: fuses DA3 intermediate features into image features via cross-attention.

3.2. Proposed Method

We extend World4Drive [23] for robustness to camera viewpoint changes by integrating geometric priors from DA3 [13], without relying on viewpoint augmentation. As illustrated in Fig. 2, the extension consists of two modules.

The first module, **3D Spatial Encoder**, injects depth-derived per-pixel 3D positions as positional embeddings into image features. The second module, **Geometric Feature Fusion**, injects DA3 intermediate features into image features via cross-attention. The first module provides explicit 3D spatial cues, while the second transfers geometric features learned by the pretrained 3D foundation model. The resulting features are then passed to the planning head for trajectory prediction.

3.2.1. 3D Spatial Encoder

The 3D Spatial Encoder augments image features with depth-derived 3D positional information. The current surround-view images $\mathbf{I}_t$ are first processed by an image encoder (ResNet-50) to extract multi-view feature maps $\mathbf{F}_t$. To estimate scene geometry, DA3 takes the temporal multi-view sequence $\{\mathbf{I}_{t-K+1}, \ldots, \mathbf{I}_t\}$ together with the corresponding camera parameters transformed into the current ego-vehicle coordinate frame. DA3 then predicts a depth map D_t for the current frame.

For each pixel (u, v), its 3D position $\mathbf{p}$ in ego-vehicle coordinates is obtained from the depth value $D_t(u, v)$, camera intrinsics $\mathbf{K}$, and extrinsics $(\mathbf{R}, \mathbf{d})$:

$$\mathbf{p} = \mathbf{R}\left(\mathbf{K}^{-1}[u,\ v,\ 1]^\top \cdot D_t(u, v)\right) + \mathbf{d}. \tag{3}$$

Applying this operation to all pixels yields a 3D point cloud $\mathbf{P}_t$. The point cloud $\mathbf{P}_t$ is encoded with sinusoidal positional encoding (SPE), and the result is mapped to positional embeddings $\mathbf{E}_t$ via a learnable MLP:

$$\mathbf{E}_t = \mathrm{MLP}(\mathrm{SPE}(\mathbf{P}_t)). \tag{4}$$

The positional embeddings are then added to the image features:

$$\hat{\mathbf{F}}_t = \mathbf{F}_t + \mathbf{E}_t. \tag{5}$$

This operation provides the planner with explicit 3D spatial cues derived from the pretrained depth model.

3.2.2. Geometric Feature Fusion

The Geometric Feature Fusion (GFF) module injects DA3 intermediate features into the image features. From the same temporal multi-view input used in the 3D Spatial Encoder, DA3 intermediate features $\mathbf{G}_t$ are extracted for the

Figure 3. Camera images under different viewpoint perturbation conditions. Each column shows the Original, Pitch $-10°$, and Depth $+1.0$ m conditions for the front and front-left cameras.

current frame. $\mathbf{G}_t$ encodes geometrically consistent information across views because DA3 internally processes camera tokens.

Following WorldRFT [19], the module fuses DA3 features into the image features via cross-attention, using $\hat{\mathbf{F}}_t$ as the query and $\mathbf{G}_t$ as the key and value:

$$\tilde{\mathbf{F}}_t = \text{CrossAttention}(\hat{\mathbf{F}}_t,\, \mathbf{G}_t). \tag{6}$$

The DA3 parameters are frozen during training to preserve the geometric knowledge acquired through large-scale pretraining. The geometry-fused features $\tilde{\mathbf{F}}_t$ are then passed to the planning head for trajectory prediction.

4. Experiment

4.1. Setup

Dataset. We use the VR-Drive [3] evaluation framework based on the nuScenes [1] dataset to evaluate robustness to viewpoint perturbations. This framework defines five perturbation conditions: pitch rotation ($-10°$, $+5°$), height translation (-0.7 m, $+1.0$ m), and depth translation ($+1.0$ m). Evaluation images for each perturbation condition are generated via Novel View Synthesis (NVS) using OmniRe [2]. Generalization to unseen camera viewpoints is measured by evaluating models trained on the original nuScenes data using these perturbed images. Fig. 3 shows example images under representative perturbation conditions.

Metrics. Following ST-P3 [5], we adopt L2 displacement error (m) and collision rate (%) between predicted and ground-truth trajectories.

Baselines. We use the reported results of AD-MLP [21], BEV-Planner [11], VAD [7], SparseDrive [17], DiffusionDrive [12], and VR-Drive [3] used in VR-Drive's evaluation. World4Drive [23] and the proposed method are trained and evaluated under the same setting.

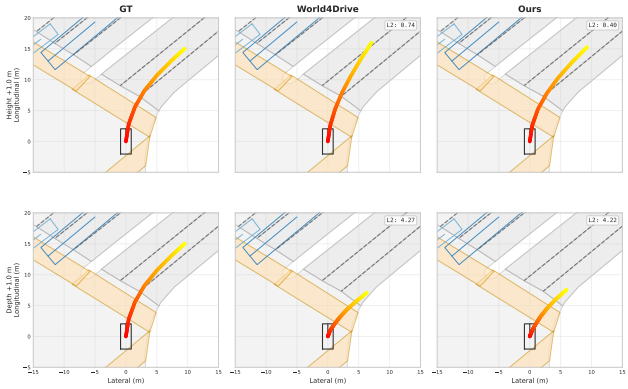

Figure 4. BEV trajectory comparison under viewpoint perturbations. Rows correspond to Height $+1.0$ m and Depth $+1.0$ m conditions. Under Height perturbation, the proposed method remains close to the ground truth, whereas World4Drive shows a larger deviation. Under Depth perturbation, both methods fail.

Training Setup. Following World4Drive [23], we adopt ResNet-50 as the image encoder with an input image resolution of 360×640. For DA3 [13], we use the Giant-size pretrained model and freeze its parameters during training. We use AdamW as the optimizer with an initial learning rate of 5×10^{-5} and weight decay of 0.01, with cosine annealing for learning rate scheduling. We train the proposed model for 12 epochs with a batch size of 8, using 8 NVIDIA RTX 6000 Ada GPUs for all experiments.

4.2. Main Results

Tab. 1 shows that the proposed method improves robustness to pitch and height perturbations relative to World4Drive [23], while maintaining comparable performance under the original condition. In contrast, gains under Depth $+1.0$ m are smaller, indicating that longitudinal translation remains challenging. VR-Drive, which augments training viewpoints through novel view synthesis, achieves the best robustness across all perturbation conditions. These results suggest that augmentation-free geometric priors improve viewpoint robustness, but do not yet match augmentation-based robustness under severe viewpoint shifts.

Compared with perception-based end-to-end methods (VAD, SparseDrive, DiffusionDrive), the proposed method does not show a consistent advantage across all perturbation conditions. Fig. 4 shows that the proposed method maintains accurate trajectory prediction under Height perturbation, whereas both World4Drive and our method degrade under Depth perturbation.

4.3. Further Analysis

The main results show that the proposed method is effective under pitch and height perturbations, whereas gains under

Table 1. Comparison of trajectory planning performance under each viewpoint perturbation condition in the VR-Drive evaluation framework. $*$ denotes methods that use only ego-status without camera images as input. The Type column indicates P for Perception-based and S for Self-Supervised.

Method	Type	Original		Pitch $+5°$		Pitch $-10°$		Height $+1.0\,m$		Height $-0.7\,m$		Depth $+1.0\,m$	
		L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓
AD-MLP$*$	-	0.29	0.19	0.29	0.19	0.29	0.19	0.29	0.19	0.29	0.19	0.29	0.19
BEV-Planner$*$	-	0.55	0.22	0.59	0.37	0.54	0.76	0.57	0.29	0.58	0.64	0.58	0.28
VAD	P	0.72	0.22	0.68	0.28	1.02	0.88	0.73	0.47	0.74	0.22	0.71	0.26
SparseDrive	P	0.61	0.08	0.66	0.15	0.96	0.23	0.87	0.54	1.01	0.30	1.27	0.31
DiffusionDrive	P	0.57	0.08	0.67	0.11	0.96	0.24	1.46	0.81	1.21	0.20	1.57	0.41
VR-Drive	P	0.60	0.06	0.60	0.06	0.70	0.11	0.69	0.11	0.69	0.14	0.72	0.13
World4Drive	S	0.48	0.26	0.83	0.48	0.64	0.36	1.06	1.29	5.01	5.44	7.10	6.40
Ours	S	0.49	0.32	0.54	0.30	0.69	0.21	0.63	0.42	0.90	0.46	6.61	5.21

Table 2. Ablation of module contributions. The Depth column indicates the depth estimator used in the 3D Spatial Encoder, and the GFF column indicates whether Geometric Feature Fusion is enabled.

Depth	GFF	Original		Pitch $+5°$		Pitch $-10°$		Height $+1.0\,m$		Height $-0.7\,m$		Depth $+1.0\,m$	
		L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓	L2 [m]↓	Col. [%]↓
Metric3D		0.48	0.26	0.83	0.48	0.64	0.36	1.06	1.29	5.01	5.44	7.10	6.40
Metric3D	✓	0.58	0.35	0.65	0.38	0.73	0.33	0.83	0.46	0.78	0.34	4.94	5.92
DA3		0.52	0.28	0.70	0.37	1.06	0.71	4.58	4.54	5.27	5.80	7.91	6.73
DA3	✓	0.49	0.32	0.54	0.30	0.69	0.21	0.63	0.42	0.90	0.46	6.61	5.21

Table 3. Counterfactual diagnosis under Depth $+1.0\,m$. Camera extrinsic parameters used for 3D positional embeddings are replaced with their training-time values at test time.

Replaced cameras	L2 [m]↓	Col. [%]↓
None (Depth $+1.0\,m$)	6.61	5.21
Front/rear only	5.62	4.81
Four side cameras	1.30	0.30
All cameras	0.51	0.29

depth perturbation are smaller. This section investigates the source of this discrepancy through an ablation of the two proposed modules and a counterfactual analysis of the 3D positional embeddings.

4.3.1. Ablation of Module Contributions

The proposed method introduces two changes to World4Drive: replacing the depth estimator in the 3D Spatial Encoder with DA3 [13], and adding Geometric Feature Fusion (GFF). To isolate the contribution of each change, we compare the full model with a configuration that retains Metric3D [20] as the depth estimator and introduces only GFF (Tab. 2).

The GFF-only configuration is worse than the baseline under the original condition, but reduces performance degradation under pitch and height perturbations. This result indicates that GFF is the primary source of the improvement in viewpoint generalization. By contrast, replacing the depth estimator with DA3 alone degrades performance, while adding GFF mitigates this degradation and recovers gains under several perturbation conditions. These results show that changing the depth estimator alone does not improve viewpoint generalization.

4.3.2. Analysis of Distribution Shift in 3D Positional Embedding

We analyze the limited gains under depth perturbation through a counterfactual diagnosis of the 3D positional embeddings. The 3D positional embeddings directly depend on camera extrinsic parameters (Eq. 3). When camera viewpoints change at test time, the 3D Spatial Encoder produces 3D position distributions that differ from those observed during training. Training under a single camera configuration can therefore make the model sensitive to the 3D position patterns seen during training.

To verify this hypothesis, we evaluate Depth $+1.0\,m$ perturbed images while replacing only the camera extrinsic parameters used for 3D positional embedding computation with their training-time values (Tab. 3). Replacing the four side cameras (front-left, front-right, back-left, back-right) improves L2, and replacing all cameras yields performance close to the original condition. This result demonstrates that the distributional shift in 3D positional embeddings caused by camera extrinsic changes is the main source of the performance degradation.

5. Conclusion

We propose an augmentation-free method for improving robustness to camera viewpoint changes in end-to-end autonomous driving by integrating geometric priors from a 3D foundation model. The proposed method uses two modules: the 3D Spatial Encoder for depth-derived 3D positional embeddings and Geometric Feature Fusion for injecting DA3 intermediate features into image features. Experiments on the VR-Drive viewpoint perturbation benchmark show clear gains under pitch and height perturbations, while gains under longitudinal translation remain limited.

The analysis shows that the main bottleneck comes from 3D positional embeddings that depend explicitly on camera extrinsic parameters. This result suggests that geometric priors from 3D foundation models are useful for viewpoint robustness, but their effectiveness depends critically on how they are integrated into the driving model.

A promising direction for future work is therefore to construct explicit 3D intermediate representations, such as BEV, voxel, or Gaussian-based scene representations, from the geometric knowledge of 3D foundation models. Such representations inherently decouple scene understanding from camera extrinsic parameters and can provide a more viewpoint-agnostic integration to exploit 3D foundation priors.

References

[1] 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 *CVPR*, pages 11618–11628, 2020. 2, 4

[2] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. OmniRe: Omni Urban Scene Reconstruction. In *ICLR*, 2025. 4

[3] Hoonhee Cho, Jae-Young Kang, Giwon Lee, Hyemin Yang, Heejun Park, Seokwoo Jung, and Kuk-Jin Yoon. VR-Drive: Viewpoint-Robust End-to-End Driving with Feed-Forward 3D Gaussian Splatting. In *NeurIPS*, 2025. 1, 2, 4

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2

[5] 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 *ECCV*, pages 533–549. Springer, 2022. 1, 2, 4

[6] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented Autonomous Driving. In *CVPR*, pages 17853–17862, 2023.

[7] 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 *ICCV*, pages 8340–8350, 2023. 1, 2, 4

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM TOG*, 42(4):139:1–139:14, 2023. 1

[9] 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 *ICCV*, pages 4015–4026, 2023. 2

[10] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing End-to-End Autonomous Driving with Latent World Model. In *ICLR*, 2025. 1, 2

[11] 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 *CVPR*, pages 14864–14873, 2024. 4

[12] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, pages 12037–12047, 2025. 1, 2, 4

[13] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 1, 2, 3, 4, 5

[14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[15] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc van Gool, and Fisher Yu. UniDepth: Universal Monocular Metric Depth Estimation. In *CVPR*, pages 10106–10116, 2024. 1, 2

[16] Malte Stelzer, Jan Pirklbauer, Jan Bickerdt, Volker P Schomerus, Jan Piewek, Thorsten Bagdonat, and Tim Fingscheidt. On Camera and LiDAR Positions in End-to-End Autonomous Driving. In *European Conference on Computer Vision*, pages 224–240. Springer, 2024. 1, 2

[17] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. SparseDrive: End-to-End Autonomous Driving via Sparse Scene Representation. In *ICRA*, pages 8795–8801, 2024. 1, 2, 4

[18] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual Geometry Grounded Transformer. In *CVPR*, pages 5294–5306, 2025. 1, 2

[19] Pengxuan Yang, Ben Lu, Zhongpu Xia, Chao Han, Yinfeng Gao, Teng Zhang, Kun Zhan, Xianpeng Lang, Yupeng

Zheng, and Qichao Zhang. WorldRFT: Latent World Model Planning with Reinforcement Fine-Tuning for Autonomous Driving. In *AAAI*, 2026. 1, 2, 4

[20] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, pages 9043–9053, 2023. 1, 2, 5

[21] 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. 4

[22] Hanxue Zhang, Haoran Jiang, Qingsong Yao, Yanan Sun, Renrui Zhang, Hao Zhao, Hongyang Li, Hongzi Zhu, and Zetong Yang. Detect anything 3d in the wild. In *CVPR*, pages 5048–5059, 2025. 1, 2

[23] Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, and Dongbin Zhao. World4Drive: End-to-End Autonomous Driving via Intention-aware Physical Latent World Model. In *ICCV*, pages 28632–28642, 2025. 1, 2, 3, 4