Supplementary Material for
SpaceDrive: Infusing Spatial Awareness into VLM-based Autonomous Driving

Peizheng Li* [1,2], Zhenghao Zhang* [1,4], David Holtz[1], Hang Yu[1,5], Yutong Yang[1,6],
Yuzhi Lai[2], Rui Song[7], Andreas Geiger[2,3], Andreas Zell[2]
[1]Mercedes-Benz AG, [2]University of Tübingen, [3]Tübingen AI Center,
[4]TU Munich, [5]Karlsruhe Institute of Technology, [6]University of Stuttgart, [7]UCLA

https://zhenghao2519.github.io/SpaceDrive_Page/

Abstract

The main content of this supplementary material is organized as follows:

- *Section A: More implementation details of our method;*
- *Section B: Additional experiments and ablation studies;*
- *Section C: Additional visualization comparisons and qualitative analysis.*

A. Additional Implementation Details

A.1. SpaceDrive Framework

To ensure seamless adaptability, our method avoids any model-specific customization and fully preserves the original image preprocessing, patchification strategy, text tokenization, and chat template used by each base VLM model. Given the shape of the preprocessed visual patches, the depth map is resized accordingly using min-pooling (*e.g.* patch shapes of $6 \times 24 \times 24$ for LLaVA-1.5-7B [41] and $6 \times 23 \times 23$ for Qwen2.5-VL-7B [2] in our configuration). Newly introduced tokens, such as ⟨IND⟩, are set as learnable and appended to the frozen input embedding layer and output language-model head. The PE decoder for coordinates is implemented as a standard two-layer MLP with the same hidden dimensionality as the base VLM. In all experiments, we set the seed to 888. The LoRA configurations are listed in Tab. A.

A.2. Training Details

Open-loop Planning For open-loop planning, we follow prior works and use 6 future trajectory points sampled at 2 Hz over a 3-second horizon as ground truth supervision. As emphasized in previous studies [38, 74], strong open-loop planning performance can be achieved using only ego-status. To rigorously validate the effectiveness of our framework, the standard SpaceDrive variant intentionally excludes motion dynamics and high-level driving commands (*e.g.* “go straight”, “turn right”) from its inputs. In this configuration, the model performs trajectory planning exclusively from image observations, enabling a clean evaluation of the spatial reasoning capability brought by our design. The variant SpaceDrive+ includes the current commands and ego dynamics of past 2 frames that are widely used in other works [12, 61].

Table A. **LoRA configurations for VLM fine-tuning.**

Setting	Rank (r)	Alpha (α)	Dropout	Target Modules
Value	16	16	0.05	q_proj, k_proj, v_proj, o_proj

Table B. **Counterfactual reasoning comparison in the open-loop planning (without ego status).** P and R here stand for Precision and Recall. Results are highlighted in **bold** and underline for the best and the second-best performance.

Method	Safe		Red Light		Collision		Drivable Area	
	P	R	P	R	P	R	P	R
BEV-MLP	70.2	17.3	48.7	53.6	31.1	70.4	32.4	56.6
Omni-L [61]	**72.1**	58.0	59.2	63.3	34.3	71.3	49.1	**59.2**
Omni-Q [61]	70.7	49.0	57.6	58.3	32.3	**72.6**	48.5	58.6
SpaceDrive (**ours**)	65.7	**63.6**	**70.3**	**72.7**	**37.5**	66.4	**55.0**	37.0

For VQA training and evaluation, we adopt the dataset provided by OmniDrive [61], which includes scene description, attention, counterfactual reasoning, planning, as well as other general conversations. Consistent with the implementation of OmniDrive, the other VQA tasks are appended subsequent to the trajectory planning task to ensure semantic stability.

Closed-loop Planning Inspired by SimLingo [50], we augment the supervision of 6 trajectory points with 20 additional path waypoints, uniformly spaced at 1-meter intervals. In this setup, the trajectory points serve two purposes: estimating the target speed and identifying the appropriate waypoint for the target direction. This leads to generally more stable steering regardless of whether the ego vehicle is moving or not. Two PID contollers are applied to determine acceleration and steering, respectively. During training, we use a subset of SimLingo routes containing 3600 episodes with PDM-lite as the expert driver.

[*] Equal contribution, names are sorted alphabetically.

Table C. **Ablation of depth estimator.**

$f_{dep.}$	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
DepthAnythingV2 [68]	1.76	1.95	3.96
UniDepthV2 [49]	1.80	1.88	4.21

Table D. **Ablation of LoRA rank.** Learn. Par. is the abbreviation for the number of LoRA parameters when selecting Qwen2.5-VL-7B as the base VLM.

Rank	Learn. Par.	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
16	10.09M	1.80	1.88	4.21
64	40.37M	1.88	2.13	4.08
128	80.74M	1.82	2.25	4.68

Table E. **Ablation of PE frequency.**

Frequency	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
$1000^{-2i/d_a}$	1.78	2.01	3.93
$10000^{-2i/d_a}$	1.83	1.83	3.18
$20000^{-2i/d_a}$	1.80	1.88	4.21

Table F. **Ablation of regression loss.**

$\mathcal{L}_{\text{reg.}}$	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
MAE	1.86	1.82	5.73
MSE	1.82	2.14	6.22
Huber Loss	1.80	1.88	4.21

B. Additional Experiments and Analyses

B.1. VQA for Counterfactual Reasoning

As aforementioned, to validate the spatial reasoning capabilities of SpaceDrive, we conduct counterfactual reasoning experiments following the setting in OmniDrive [61], as presented in Tab. B. In this evaluation, keywords such as “safety”, “collision”, “running a red light”, and “out of the drivable area” are extracted from the VQA outputs and compared against ground truth keywords to compute Precision and Recall. The results demonstrate that our framework achieves superior performance across the majority of metrics, *e.g.* a Recall of 63.6% in the safety task. It is particularly noteworthy that, without any specific prompt engineering for the dialogue, the mere incorporation of the unified 3D spatial representation enables significantly higher Precision in tasks demanding rigorous spatial understanding, such as Collision (37.5%) and Drivable Area (55.0%). This further confirms our SpaceDrive possesses strong spatial reasoning capabilities.

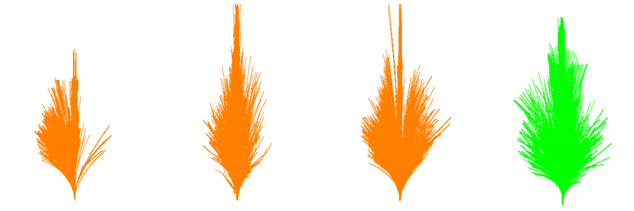
B.2. More Ablation Studies

Depth Estimator In Tab. C, we compare the influence of different pre-trained depth estimator on the planning performance. DepthAnythingV2 [68] and UniDepthV2 [49] are selected as representative examples of relative and metric depth estimation models, respectively. We observe that both variants perform similarly on the L2 error metric and Collision rate, which are the most reliable indicator of planning performance. This suggests that the effectiveness of our SpaceDrive is independent of a specific pre-trained depth model, implicitly demonstrating the adaptability of our framework. Notably, LiDAR-based depth ground truth (GT) is inherently sparse and lacks valid depth values in regions such as the sky, necessitating manual definition. Together with factors like camera distortion and projection error, GT-based comparisons are unreliable and thus excluded from the comparison.

LoRA Rank Table D presents a comparison of different LoRA [17] ranks in the VLM during fine-tuning. Benefiting from our universal spatial positional encoding, the coordinate regression process in the language model is simplified. Utilizing only low-rank fine-tuning (rank 16) achieves the optimal overall result (L2 error of 1.80, Collision rate of 1.88%, and Intersection rate of 4.21%). While increasing the rank to 128 substantially raises the number of learnable parameters from 10.09M to 80.74M, it fails to improve the planning accuracy and, instead, leads to a degradation in Collision and Intersection rates. We attribute this to the excessive degrees of training freedom in the high-rank adapter, which hinders the convergence. The above comparison further demonstrates that our method not only offers stronger planning reliability but also maintains parameter efficiency.

PE Frequency Table E investigates the base frequency of the Sin-Cos PE, which impacts both encoding resolution and smoothness. Utilizing a smaller frequency base (corresponding to a higher frequency) introduces a larger phase shift between adjacent positions but leads to positional aliasing at long distances, thereby inhibiting the representation of far-field positions. As shown in the comparison, setting the base to 1000 enhances local resolution and achieves the lowest L2 error of 1.78. However, distant coordinates exhibit near-random phase characteristics, which compromises overall safety (leading to worse Collision and Intersection rates). Conversely, an excessively large base (*e.g.* 20000) generates smoother, more stable encodings over long distances but diminishes local discriminative capability. Compared to the original base of 10000, the resulting L2 error reduction is less pronounced, at only -0.03, but the collision rate increase is negligible. Overall, comparing all variants reveals that the influence of different PE frequencies is relatively limited and non-decisive. We finally adopt 20000 as our PE frequency base.

Regression Loss In Tab. F, we compares different regression losses for trajectory prediction. MAE provides robustness to outliers but yields the worst L2 and intersection metrics, suggesting insufficient pressure on medium-scale errors. MSE reduces L2 compared to MAE, but its quadratic growth on



(a) OmniDrive-L (b) SpaceDrive (c) SpaceDrive-CoT (d) Ground Truth

Figure A. **Trajectory distribution of open-loop planning for different frameworks and ground truth.**

large residuals makes optimization more sensitive to outliers, leading to noticeably higher collision and intersection rates. Huber loss strikes a balance between them and achieves the best L2 error together with markedly improved safety metrics. So we adopt Huber loss as our final regression objective.

B.3. Comprehensive Benchmark

Constrained by the limited space in the main paper, we list only the primary relevant works in the benchmark comparisons. Therefore, we provide more comprehensive benchmark comparisons for open-loop and closed-loop planning in Tab. G and Tab. H, respectively. It is worth noting that existing nuScenes [4] open-loop evaluations utilize differing sets of metrics in different studies. While the main paper employs the OmniDrive [61] version commonly used by VLM-based frameworks, Table G provides results derived using the evaluation metrics from ST-P3 [18] and UniAD [19].

C. Additional Visualization

C.1. More Adaptability Analysis

Figure A illustrates the distribution of planned trajectories across all scenarios under the open-loop setting of nuScenes [4]. We first analyze the output trajectory distribution of OmniDrive-L [61], a typical scheme utilizing textual digit tokens for waypoint coordinates, shown in Fig. A.a. Due to the VLM's limitations in numerical processing, as discussed in Section 1, OmniDrive-L exhibits clear mode collapse for right-turn cases. In sharp contrast, our SpaceDrive, which is based on the universal 3D PE representation, significantly mitigates this issue, as shown in Fig. A.b. Furthermore, when adopting inference techniques such as Chain-of-Thought during inference, the output trajectory planning demonstrates enhanced robustness (Fig. A.c) and closer alignment with the ground truth distribution (Fig. A.d). This result further supports the strong adaptability of our method to language model inference techniques.

C.2. Failure Analysis of Textual Coordinate Output

A further quantitative analysis is conducted to assess the driving capability of conventional VLM-based models that output trajectory coordinates as textual digit tokens in closed-loop simulation, as shown in Fig. B. We the exact same scenario as in Fig. 3 and employ OmniDrive-L [61], a framework structurally analogous to SpaceDrive, utilizing the same closed-loop training configuration as in Sec. A.2. This figure clearly illustrates that in the closed-loop setting, the planned trajectories generated by OmniDrive-L collapse into an approximately straight line, and the directional control exhibits random oscillation. This phenomenon aligns with the mode collapse previously observed during open-loop evaluation (See Sec. C.1). Critically, this oscillation is amplified over time, leading to vehicle instability and ultimately making the vehicle veer off the road and collide with the guardrail. This result provides strong empirical support for our analysis in Sec. 4.2: purely text-based trajectory coordinate output from VLMs is inadequate for reliable closed-loop driving.

C.3. More Qualitative Closed-Loop Results

We present additional closed-loop simulation visualizations for SpaceDrive in Fig. C, covering 3 representative safety-critical scenarios: (a) navigating around a construction zone requiring a brief excursion into the oncoming lane; (b) decelerating and yielding due to a sudden pedestrian crossing during normal driving; and (c) performing an emergency stop and yielding to an ambulance rapidly approaching from the rear. All these scenarios demand the model to quickly establish a deep understanding of the 3D spatial context and generate a sound trajectory in a minimal timeframe. The visualizations clearly indicate that our proposed framework, by leveraging its unified 3D representation, effectively manages these critical, unforeseen situations. This further substantiates the efficacy of our proposed SpaceDrive framework.

Table G. **Open-loop planning results on nuScenes [4].** SpaceDrive+ denotes the adoption of the ego planner input. ‡: The model is trained using only the trajectory prediction task for open-loop planning, without utilizing our generated OmniDrive Q&A data. Methods marked as Hybrid Paradigm here stack traditional and VLM-based approaches, and are thus incomparable. Results are highlighted in **bold** and underline for the best and the second-best performance among VLM-based methods.

Method	Ego Status		ST-P3 Metrics								UniAD Metrics							
			L2 (m) ↓				Collision (%) ↓				L2 (m) ↓				Collision (%) ↓			
	BEV	Planner	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Traditional Modular Paradigm																		
ST-P3 [18]	-	-	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	1.72	3.26	4.86	3.28	0.44	1.08	3.01	1.51
UniAD [19]	-	-	0.44	0.67	0.96	0.69	0.04	0.08	0.23	0.12	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
VAD-Base [27]	-	-	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22	0.54	1.15	1.98	1.22	0.10	0.24	0.96	0.43
UAD [14]	-	-	0.28	0.41	0.65	0.45	0.01	0.03	0.14	0.06	0.39	0.80	1.50	0.90	0.01	0.12	0.43	0.19
MomAD [55]	-	-	0.28	0.49	0.78	0.52	0.08	0.14	0.34	0.19	0.36	0.83	1.56	0.91	0.06	0.23	1.00	0.43
GenAD [80]	-	✓	0.31	0.57	0.91	0.60	0.01	0.05	0.22	0.09	0.43	0.88	1.62	0.98	0.06	0.16	0.68	0.30
Drive-WM [63]	✓	✓	0.43	0.77	1.20	0.80	0.10	0.21	0.48	0.26	-	-	-	-	-	-	-	-
SparseDrive [57]	-	✓	0.29	0.55	0.91	0.58	0.01	0.02	0.13	0.06	0.44	0.92	1.69	1.01	0.07	0.19	0.71	0.32
DiffusionDrive [40]	-	✓	0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08	-	-	-	-	-	-	-	-
VLM-based Paradigm																		
EMMA [22]	-	-	**0.14**	**0.29**	0.54	**0.32**	-	-	-	-	-	-	-	-	-	-	-	-
RDA-Driver [21]	✓	✓	0.17	0.37	0.69	0.40	**0.01**	**0.05**	0.26	**0.10**	0.23	0.73	1.54	0.80	**0.00**	**0.13**	0.83	**0.32**
DriveVLM [59]	-	✓	0.18	0.34	0.68	0.40	-	-	-	-	-	-	-	-	-	-	-	-
ORION [12]	✓	-	0.17	0.31	0.55	0.34	-	-	-	-	-	-	-	-	-	-	-	-
OmniDrive-Q [61]	-	-	1.15	1.96	2.84	1.98	-	-	-	-	-	-	-	-	-	-	-	-
OmniDrive-Q++ [61]	✓	✓	**0.14**	**0.29**	0.55	0.33	-	-	-	-	-	-	-	-	-	-	-	-
OmniDrive-L‡ [61]	-	-	1.47	2.43	3.38	2.43	-	-	-	-	-	-	-	-	-	-	-	-
OmniDrive-L++‡ [61]	-	✓	0.31	0.62	1.06	0.66	-	-	-	-	-	-	-	-	-	-	-	-
OmniDrive-L [61]	-	-	1.43	2.34	3.24	2.34	-	-	-	-	-	-	-	-	-	-	-	-
OmniDrive-L++ [61]	-	✓	0.15	0.36	0.70	0.40	-	-	-	-	-	-	-	-	-	-	-	-
SpaceDrive (**ours**)	-	-	1.06	1.79	2.55	1.80	0.35	0.61	1.31	0.76	1.41	2.88	4.51	2.93	0.59	1.72	4.53	2.28
SpaceDrive+ (**ours**)	-	✓	0.15	**0.29**	**0.51**	**0.32**	0.05	0.08	**0.16**	**0.10**	**0.20**	**0.53**	**1.13**	**0.62**	0.10	0.31	**0.80**	0.40
Hybrid Paradigm																		
VLP [48]	✓	-	0.30	0.53	0.84	0.55	0.01	0.07	0.38	0.15	0.36	0.68	1.19	0.74	0.03	0.12	0.32	0.16
ReAL-AD [46]	✓	-	0.30	0.48	0.67	0.48	0.07	0.10	0.28	0.15	0.40	0.71	1.14	0.77	0.02	0.12	0.37	0.17
DriveVLM-Dual [59]	✓	-	0.15	0.29	0.48	0.31	0.05	0.08	0.17	0.10	-	-	-	-	-	-	-	-
SOLVE-VLM [7]	✓	-	0.13	0.25	0.47	0.28	-	-	-	-	-	-	-	-	-	-	-	-
Senna [28]	✓	-	0.11	0.21	0.35	0.22	0.04	0.08	0.13	0.08	-	-	-	-	-	-	-	-

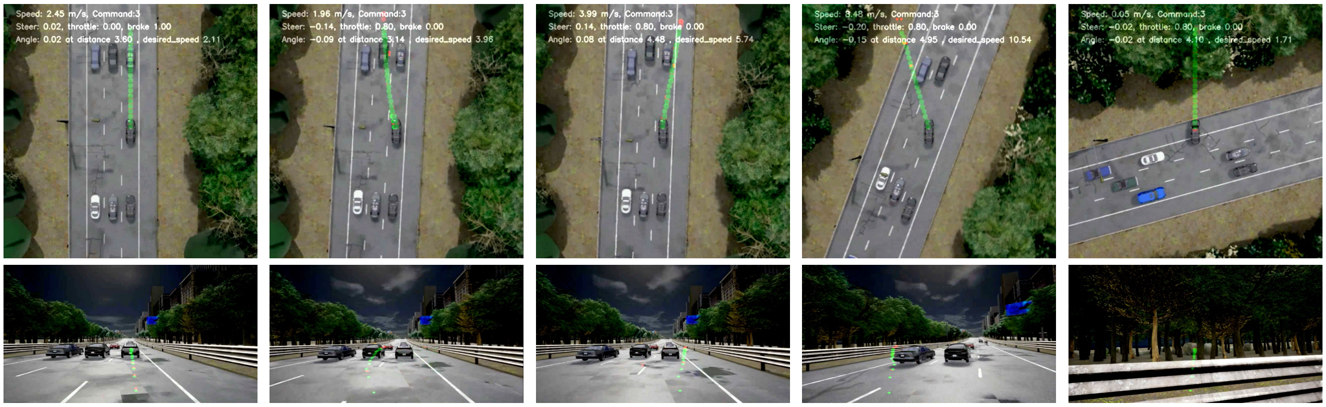
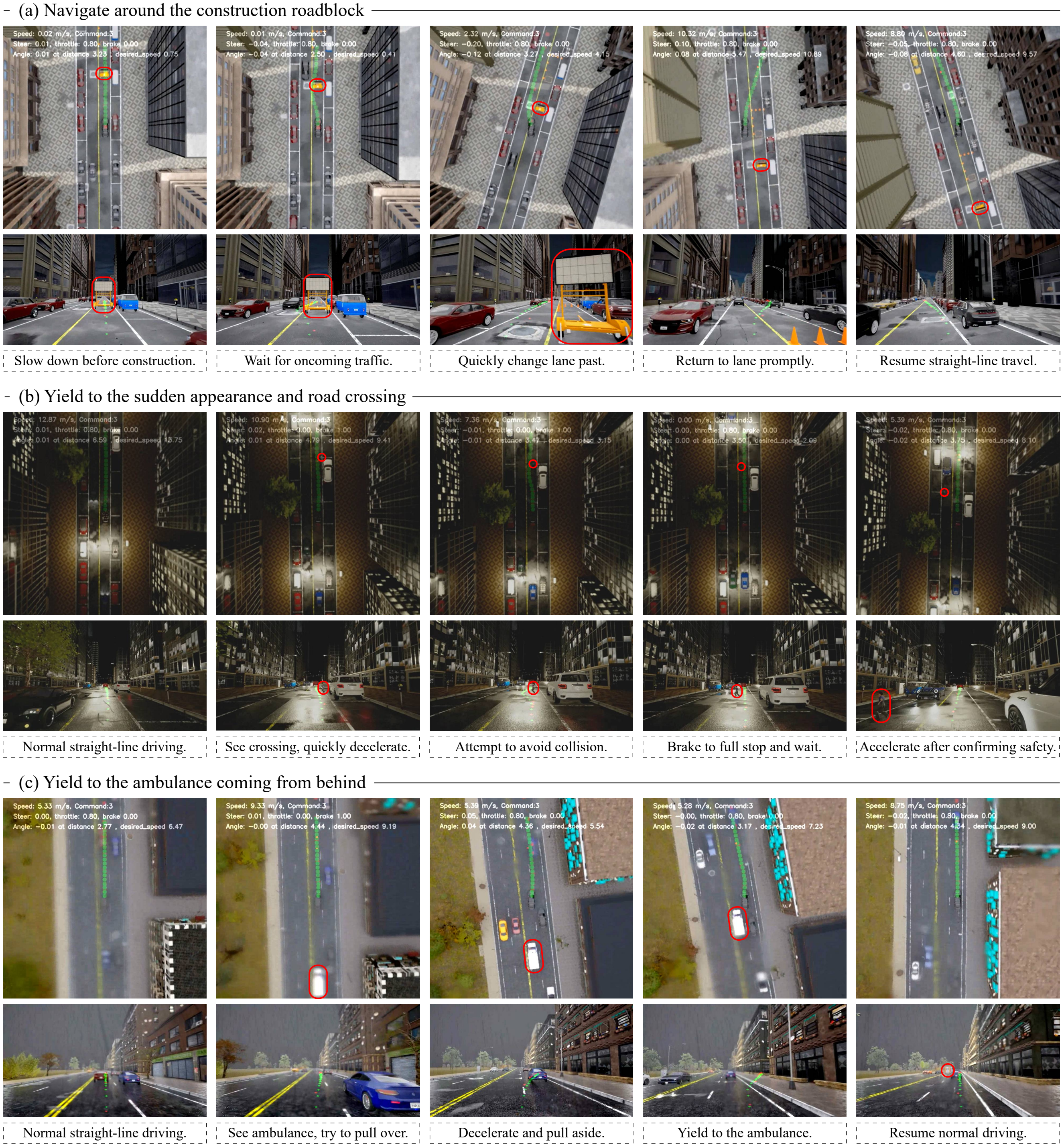


Figure B. **Qualitative results of OmniDrive-L [61] in closed-loop setting on Bench2Drive [25].** Green and pink dots represent path and speed waypoints, respectively. Parameters such as speed and steering wheel angle can be found in the figures.


Speed: 0.02 m/s, Command:3
Steer: 0.01, throttle: 0.80, brake 0.00
Angle: 0.01 at distance 3.23 , desired_speed 0.75
Speed: 0.01 m/s, Command:3
Steer: −0.04, throttle: 0.80, brake 0.00
Angle: −0.04 at distance 2.50 , desired_speed 0.41
Speed: 2.32 m/s, Command:3
Steer: −0.20, throttle: 0.80, brake 0.00
Angle: −0.12 at distance 3.27 , desired_speed 4.15
Speed: 10.32 m/s, Command:3
Steer: 0.10, throttle: 0.80, brake 0.00
Angle: 0.08 at distance 5.47 , desired_speed 10.89
Speed: 8.80 m/s, Command:3
Steer: −0.05, throttle: 0.80, brake 0.00
Angle: −0.08 at distance 4.60 , desired_speed 9.57
Speed: 12.87 m/s, Command:3
Steer: 0.01, throttle: 0.80, brake 0.00
Angle: 0.01 at distance 6.59 , desired_speed 13.75
Speed: 10.90 m/s, Command:3
Steer: 0.02, throttle: 0.00, brake 1.00
Angle: 0.01 at distance 4.79 , desired_speed 9.41
Speed: 7.36 m/s, Command:3
Steer: −0.01, throttle: 0.00, brake 1.00
Angle: −0.01 at distance 3.47 , desired_speed 3.15
Speed: 0.00 m/s, Command:3
Steer: 0.00, throttle: 0.80, brake 0.00
Angle: 0.00 at distance 3.50 , desired_speed 2.09
Speed: 5.39 m/s, Command:3
Steer: −0.02, throttle: 0.80, brake 0.00
Angle: −0.02 at distance 3.75 , desired_speed 8.10
Speed: 5.33 m/s, Command:3
Steer: 0.00, throttle: 0.80, brake 0.00
Angle: −0.01 at distance 2.77 , desired_speed 6.47
Speed: 9.33 m/s, Command:3
Steer: 0.01, throttle: 0.00, brake 1.00
Angle: −0.00 at distance 4.44 , desired_speed 9.19
Speed: 5.39 m/s, Command:3
Steer: 0.05, throttle: 0.80, brake 0.00
Angle: 0.04 at distance 4.36 , desired_speed 5.54
Speed: 5.28 m/s, Command:3
Steer: −0.00, throttle: 0.80, brake 0.00
Angle: −0.02 at distance 3.17 , desired_speed 7.23
Speed: 8.75 m/s, Command:3
Steer: −0.02, throttle: 0.80, brake 0.00
Angle: −0.01 at distance 4.34 , desired_speed 9.00

(a) Navigate around the construction roadblock

Slow down before construction. | Wait for oncoming traffic. | Quickly change lane past. | Return to lane promptly. | Resume straight-line travel.

(b) Yield to the sudden appearance and road crossing

Normal straight-line driving. | See crossing, quickly decelerate. | Attempt to avoid collision. | Brake to full stop and wait. | Accelerate after confirming safety.

(c) Yield to the ambulance coming from behind

Normal straight-line driving. | See ambulance, try to pull over. | Decelerate and pull aside. | Yield to the ambulance. | Resume normal driving.

Figure C. **More qualitative results of closed-loop evaluation on Bench2Drive [25].** We include 3 scenarios here to demonstrate the closed-loop planning capability of SpaceDrive: (a) urban road construction; (b) a sudden pedestrian crossing; (c) yielding to an ambulance. Green and pink dots represent path and speed waypoints, respectively. Red circles indicate objects requiring attention in the scenario. Parameters such as speed and steering wheel angle can be found in the figures.

Table H. **Closed-loop planning results on Bench2Drive [25].** Results are highlighted in **bold** and underline for the best and the second-best performance among VLM-based methods.

Method	Closed-loop Metric	
	Driving Score ↑	Success Rate(%) ↑
Traditional Modular Paradigm		
AD-MLP [74]	18.05	0.00
UniAD-Base [19]	45.81	16.36
VAD-Base [27]	42.35	15.00
MomAD [55]	44.54	16.71
GenAD [80]	44.81	15.90
SparseDrive [57]	47.38	17.72
UAD [14]	49.22	20.45
SeerDrive [75]	58.32	30.17
WoTE [35]	61.71	31.36
DriveDPO [51]	62.02	30.62
ThinkTwice [24]	62.44	37.17
DriveTransformer-L [26]	63.46	38.60
DriveAdapter [23]	64.22	42.08
Raw2Drive [70]	71.36	50.24
Hydra-NeXt [39]	73.86	53.22
DiffusionDrive [40]	77.68	52.72
PGS [20]	78.08	48.64
GaussianFusion [43]	79.10	54.40
TF++ [84]	84.21	64.39
R2SE [42]	86.28	67.76
HiP-AD [58]	86.77	69.09
VLM-based Paradigm		
ReAL-AD [46]	41.17	11.36
Dual-AEB [77]	45.23	10.00
X-Driver [44]	51.70	18.10
GEMINUS [60]	65.39	37.73
VDRive [15]	66.25	50.51
StuckSolver [3]	70.89	50.01
DriveMoE [69]	74.22	48.64
ETA [16]	74.33	48.33
VLR-Drive [30]	75.01	50.00
ORION [12]	77.74	54.62
SimLingo [50]	**85.07**	**67.27**
SpaceDrive+ (**ours**)	78.02	55.11

References

[1] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 5, 1

[3] Zhipeng Bao and Qianwen Li. Large language model-assisted autonomous vehicle recovery from immobilization. *arXiv preprint arXiv:2510.26023*, 2025. 6

[4] 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*, 2020. 2, 5, 6, 3, 4

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[6] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 2

[7] Xuesong Chen, Linjiang Huang, Tao Ma, Rongyao Fang, Shaoshuai Shi, and Hongsheng Li. Solve: Synergy of language-vision and end-to-end networks for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 6, 4

[8] An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, et al. 3d aware region prompted vision language model. *arXiv preprint arXiv:2509.13317*, 2025. 3

[9] Shuxiao Ding, Yutong Yang, Julian Wiederer, Markus Braun, Peizheng Li, Juergen Gall, and Bin Yang. Tqd-track: Temporal query denoising for 3d multi-object tracking. *arXiv preprint arXiv:2504.03258*, 2025. 2

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, 2017. 5

[11] Xiang Fei, Jinghui Lu, Qi Sun, Hao Feng, Yanjie Wang, Wei Shi, An-Lan Wang, Jingqun Tang, and Can Huang. Advancing sequential numerical prediction in autoregressive models. *arXiv preprint arXiv:2505.13077*, 2025. 2, 4

[12] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 1, 2, 3, 4, 6, 8

[13] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 3

[14] Mingzhe Guo, Zhipeng Zhang, Yuan He, Ke Wang, Liping Jing, and Haibin Ling. End-to-end autonomous driving without costly modularization and 3d manual annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 6, 4

[15] Ziang Guo and Zufeng Zhang. Vdrive: Leveraging reinforced vla and diffusion policy for end-to-end autonomous driving. *arXiv preprint arXiv:2510.15446*, 2025. 6

[16] Shadi Hamdan, Chonghao Sima, Zetong Yang, Hongyang Li, and Fatma Guney. Eta: Efficiency through thinking ahead, a dual approach to self-driving with large models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 6

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 5, 2

[18] 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*, 2022. 2, 6, 3, 4

[19] 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*, 2023. 2, 6, 3, 4

[20] Yi Huang, Lihui Jiang, Bingbing Liu, Hongbo Zhang, et al. Prioritizing perception-guided self-supervision: A new paradigm for causal modeling in end-to-end autonomous driving. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 6

[21] Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*, 2024. 6, 4

[22] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 6, 4

[23] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 6

[24] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6

[25] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmark-

ing of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 2024. 2, 5, 6, 8, 4

[26] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In *International Conference on Learning Representations (ICLR)*, 2025. 6

[27] 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*, 2023. 2, 6, 4

[28] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024. 1, 6, 4

[29] Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2021. 2

[30] Fanjie Kong, Yitong Li, Weihuang Chen, Chen Min, Yizhe Li, Zhiqiang Gao, Haoyang Li, Zhongyu Guo, and Hongbin Sun. Vlr-driver: Large vision-language-reasoning models for embodied autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 6

[31] Yuzhi Lai, Shenghai Yuan, Peizheng Li, Jun Lou, and Andreas Zell. Seer-var: Semantic egocentric environment reasoner for vehicle augmented reality. *arXiv preprint arXiv:2508.17255*, 2025. 3

[32] Yuzhi Lai, Shenghai Yuan, Boya Zhang, Benjamin Kiefer, Peizheng Li, Tianchen Deng, and Andreas Zell. Fam-hri: Foundation-model assisted multi-modal human-robot interaction combining gaze and speech. *arXiv preprint arXiv:2503.16492*, 2025. 3

[33] Peizheng Li, Shuxiao Ding, Xieyuanli Chen, Niklas Hanselmann, Marius Cordts, and Juergen Gall. Powerbev: A powerful yet lightweight framework for instance prediction in bird's-eye view. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 2023. 2

[34] Peizheng Li, Shuxiao Ding, You Zhou, Qingwen Zhang, Onat Inak, Larissa Triess, Niklas Hanselmann, Marius Cordts, and Andreas Zell. Ago: Adaptive grounding for open world 3d occupancy prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2025. 2

[35] Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*, 2025. 2, 6

[36] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*, 2025. 3

[37] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[38] 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*, 2024. 2, 5, 6, 1

[39] Zhenxin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Zuxuan Wu, and Jose M. Alvarez. Hydra-next: Robust closed-loop driving with open-loop training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 6

[40] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2, 4, 6

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2023. 4, 1

[42] Haochen Liu, Tianyu Li, Haohan Yang, Li Chen, Caojun Wang, Ke Guo, Haochen Tian, Hongchen Li, Hongyang Li, and Chen Lv. Reinforced refinement with self-aware expansion for end-to-end autonomous driving. *arXiv preprint arXiv:2506.09800*, 2025. 6

[43] Shuai Liu, Quanmin Liang, Zefeng Li, Boyang Li, and Kai Huang. Gaussianfusion: Gaussian-based multi-sensor fusion for end-to-end autonomous driving. *arXiv preprint arXiv:2506.00034*, 2025. 2, 6

[44] Wei Liu, Jiyuan Zhang, Binxiong Zheng, Yufeng Hu, Yingzhan Lin, and Zengfeng Zeng. X-driver: Explainable autonomous driving with vision-language models. *arXiv preprint arXiv:2505.05098*, 2025. 6

[45] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European conference on computer vision*. Springer, 2022. 2

[46] Yuhang Lu, Jiadong Tu, Yuexin Ma, and Xinge Zhu. Real-ad: Towards human-like reasoning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 6, 4

[47] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*. Springer, 2024. 3

[48] 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*, 2024. 1, 3, 6, 4

[49] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 5, 2

[50] Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the Computer*

Vision and Pattern Recognition Conference, 2025. 2, 4, 6, 7, 1

[51] Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and Zhaoxiang Zhang. Drivedpo: Policy learning via safety dpo for end-to-end autonomous driving. *arXiv preprint arXiv:2509.17940*, 2025. 2, 6

[52] 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*, 2024. 3

[53] Yinzhe Shen, Omer Sahin Tas, Kaiwen Wang, Royden Wagner, and Christoph Stiller. Divide and merge: Motion and semantic learning in end-to-end autonomous driving. *arXiv preprint arXiv:2502.07631*, 2025. 2

[54] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, 2024. 1, 2, 3

[55] Ziying Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 6, 4

[56] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 4

[57] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 6, 4

[58] Yingqi Tang, Zhuoran Xu, Zhaotie Meng, and Erkang Cheng. Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder. *arXiv preprint arXiv:2503.08612*, 2025. 6

[59] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conference on Robot Learning*, 2025. 1, 3, 6, 4

[60] Chi Wan, Yixin Cui, Jiatong Du, Shuo Yang, Yulong Bai, Peng Yi, Nan Li, and Yanjun Huang. Geminus: Dual-aware global and scene-adaptive mixture-of-experts for end-to-end autonomous driving. *arXiv preprint arXiv:2507.14456*, 2025. 6

[61] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 1, 2, 3, 4, 5, 6

[62] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on robot learning*, 2022. 2

[63] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[64] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[65] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 2, 3

[66] 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

[67] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2, 3

[68] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 2024. 2

[69] Zhenjie Yang, Yilin Chai, Xiaosong Jia, Qifeng Li, Yuqian Shao, Xuekai Zhu, Haisheng Su, and Junchi Yan. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. *arXiv preprint arXiv:2505.16278*, 2025. 6

[70] Zhenjie Yang, Xiaosong Jia, Qifeng Li, Xue Yang, Maoqing Yao, and Junchi Yan. Raw2drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in carla v2). *arXiv preprint arXiv:2505.16394*, 2025. 6

[71] Hang Yu, Julian Jordan, Julian Schmidt, Silvan Lindner, Alessandro Canevaro, and Wilhelm Stork. Hype: Hybrid planning with ego proposal-conditioned predictions. *arXiv preprint arXiv:2510.12733*, 2025. 2

[72] Haibao Yu, Wenxian Yang, Ruiyang Hao, Chuanye Wang, Jiaru Zhong, Ping Luo, and Zaiqing Nie. Drivee2e: Closed-loop benchmark for end-to-end autonomous driving through real-to-simulation. *arXiv preprint arXiv:2509.23922*, 2025. 2

[73] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context multi-modal large language model learning. In *Robotics: Science and Systems*, 2024. 3

[74] 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, 5, 6, 1

[75] Bozhou Zhang, Nan Song, Xiatian Zhu, Jiankang Deng, Li Zhang, et al. Future-aware end-to-end driving: Bidirectional

modeling of trajectory planning and scene evolution. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 6

[76] Qingwen Zhang, Yi Yang, Peizheng Li, Olov Andersson, and Patric Jensfelt. Seflow: A self-supervised scene flow method in autonomous driving. In *European Conference on Computer Vision*. Springer, 2024. 2

[77] Wei Zhang, Pengfei Li, Junli Wang, Bingchuan Sun, Qihao Jin, Guangjun Bao, Shibo Rui, Yang Yu, Wenchao Ding, Peng Li, et al. Dual-aeb: Synergizing rule-based and multimodal large language models for effective emergency braking. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025. 6

[78] Zhiyuan Zhang, Xiaofan Li, Zhihao Xu, Wenjie Peng, Zijian Zhou, Miaojing Shi, and Shuangping Huang. Mpdrive: Improving spatial understanding with marker-based prompt learning for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 2

[79] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. 3

[80] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, 2024. 2, 6, 4

[81] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025. 3

[82] Zewei Zhou, Tianhui Cai, Seth Z Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025. 3

[83] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 3, 4

[84] Julian Zimmerlin, Jens Beißwenger, Bernhard Jaeger, Andreas Geiger, and Kashyap Chitta. Hidden biases of end-to-end driving datasets. *arXiv preprint arXiv:2412.09602*, 2024. 2, 6