SpaceDrive

Infusing Spatial Awareness into VLM-based Autonomous Driving

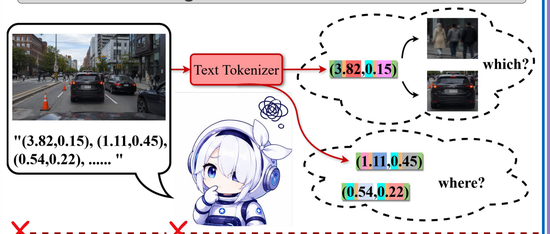
Peizheng Li*[1,2]; Zhenghao Zhang*[1,4]; David Holtz[1]; Hang Yu[1,5]; Yutong Yang[1,6]; Yuzhi Lai[2]; Rui Song[7]; Andreas Geiger[2,3]; Andreas Zell[2]

[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

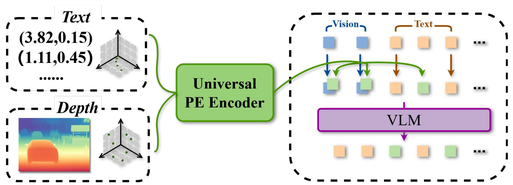
CVPR JUNE 3-7, 2026 DENVER COLORADO

TL ; DR

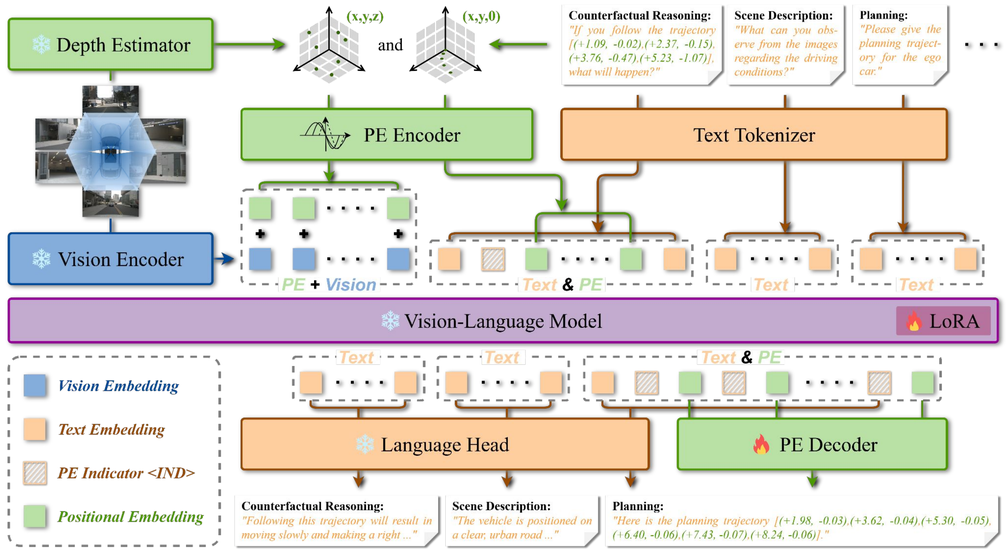
Existing VLM-based E2E Methods



SpaceDrive



Method



Ablations

Ablation 1: PE injection

PE Injection	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
baseline	2.51	4.53	6.77
+ PE on **Vision**	1.88	2.45	2.36
+ PE in **Text**	**1.80**	**1.88**	**4.21**
(+ Ego-state PE)	**0.32**	**0.23**	**1.27**

- Adding **PE on vision tokens** gives major gains. Unified PE for both **vision and text** improves further.

Ablation 3: PE normalization

α_{PE}	Learnable	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
1.0	✗	2.34	3.63	8.46
0.1	✗	2.43	3.79	9.42
0.02	✗	2.22	2.71	10.17
1.0	✓	1.82	2.04	4.62
0.1	✓	**1.80**	**1.88**	**4.21**
0.02	✓	1.86	2.03	5.42

- **Learnable PE** scaling is crucial for training stability.

Ablation 5: Depth estimators

$f_{dep.}$	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
DepthAnythingV2	1.76	1.95	3.96
UniDepthV2	1.80	1.88	4.21

- Effectiveness of SpaceDrive is **independent** from specific pre-trained depth models.

Ablation 2: PE encoder & decoder

Encoder (ϕ)	Decoder (ψ)	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
Sine-Cosine	Coordinate-wise	**1.80**	**1.88**	**4.21**
MLP	Coordinate-wise	1.96	3.17	6.76
RoPE	Coordinate-wise	1.93	3.71	11.40
Sine-Cosine	Sine-Cosine	1.87	2.62	9.20
Sine-Cosine	Task-specific	1.93	2.41	5.58

- Sine-Cosine inherently incorporates **relative position modeling**.
- Inverting Sine-Cosine encoding is **worse than** learnable decoder.

Ablation 4: Adaptability across VLMs

Method	Backbone	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
SpaceDrive	LLaVA	1.82	2.44	4.08
SpaceDrive	Qwen-VL	1.80	1.88	4.21
SpaceDrive+	LLaVA	0.31	0.23	1.42
SpaceDrive+	Qwen-VL	0.32	0.23	1.27

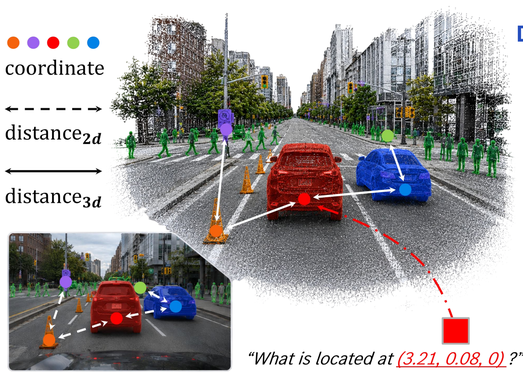
- The unified spatial representation works on **many** VLM backbones.

Ablation 6: Robustness to Depth Errors

Setting	Depth Noise	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
Training & Inference	w. -5% Global Shift	1.86	1.80	5.22
Training & Inference	w. 5% Global Shift	1.86	1.93	4.41
Training & Inference	w. 5% Random Noise	1.83	2.16	4.44
Inference only	w. -2.5% Global Shift	1.80	1.89	4.22
Inference only	w. 2.5% Global Shift	1.80	1.86	4.24
Inference only	w. 2.5% Random Noise	1.80	1.89	4.17

- PE serves as hints for **spatial association in training**, not an extra modality input.

Insights



Desired Positional Representation

Explicit 3D spatial correlation

Universal PE representation

Benchmarks

A. Open-loop planning on nuScenes

Method	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
UniAD	0.46	0.37	1.59
UAD	0.30	0.27	1.37
Drive-WM	0.80	0.26	-
EMMA	0.32	-	-
RDA-Driver	0.80	0.32	-
DriveVLM	0.40	0.27	-
ORION	0.34	0.37	-
OmniDrive-Q	1.98	3.79	4.59
OmniDrive-Q++	0.33	0.30	3.00
SpaceDrive	**1.80**	**1.88**	**4.21**
SpaceDrive+	**0.32**	**0.23**	**1.27**

B. Closed-loop planning on Bench2Drive

Method	Driving Score ↑	Success Rate ↑
ReAL-AD	41.17	11.36
Dual-AEB	45.23	10.00
VDRive	66.25	50.51
StuckSolver	70.89	50.01
DriveMoE	74.22	48.64
ETA	74.33	48.33
VLR-Drive	75.01	50.00
ORION	77.74	54.62
SimLingo	85.07	67.27
SpaceDrive+	**78.02**	**55.11**

***SpaceDrive** achieves the **best** open-loop performance among VLM-based methods on unScenes.*

***SpaceDrive** achieves **2nd-best** in VLM-based planners on Bench2Drive w/o specific closed-loop optimizations.*

Visualization