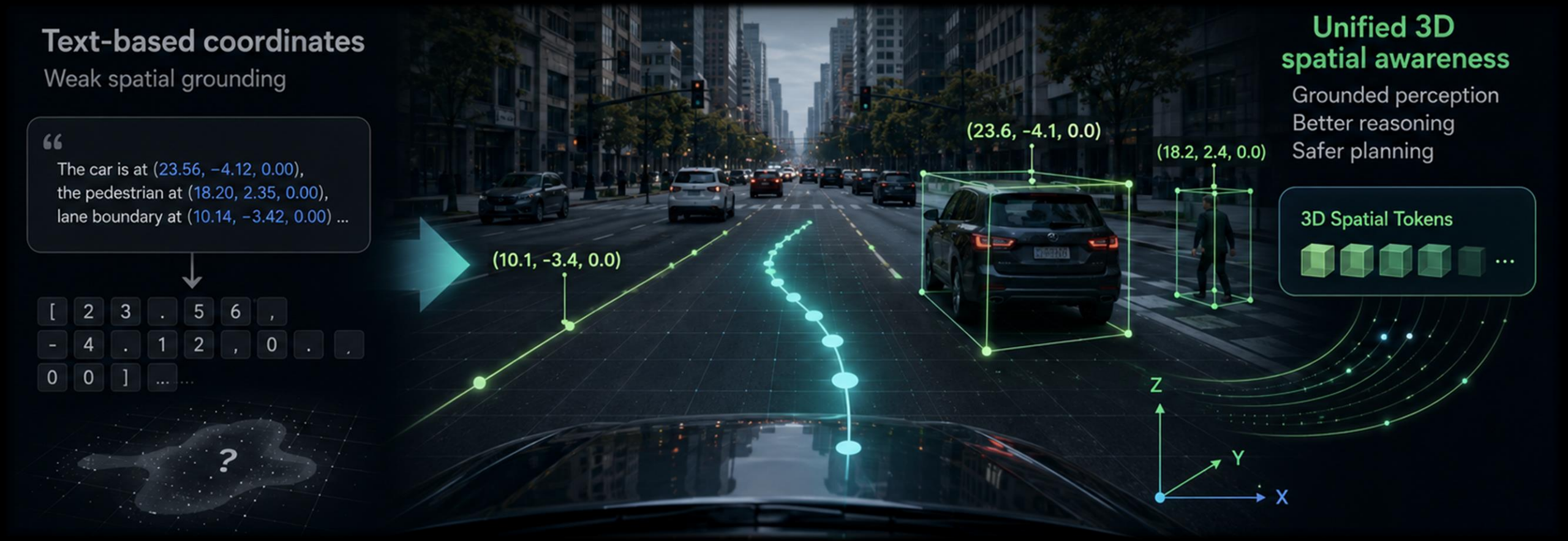


SpaceDrive: Infusing Spatial Awareness into VLM-based Autonomous Driving

Peizheng Li, Zhenghao Zhang*, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, Andreas Zell*

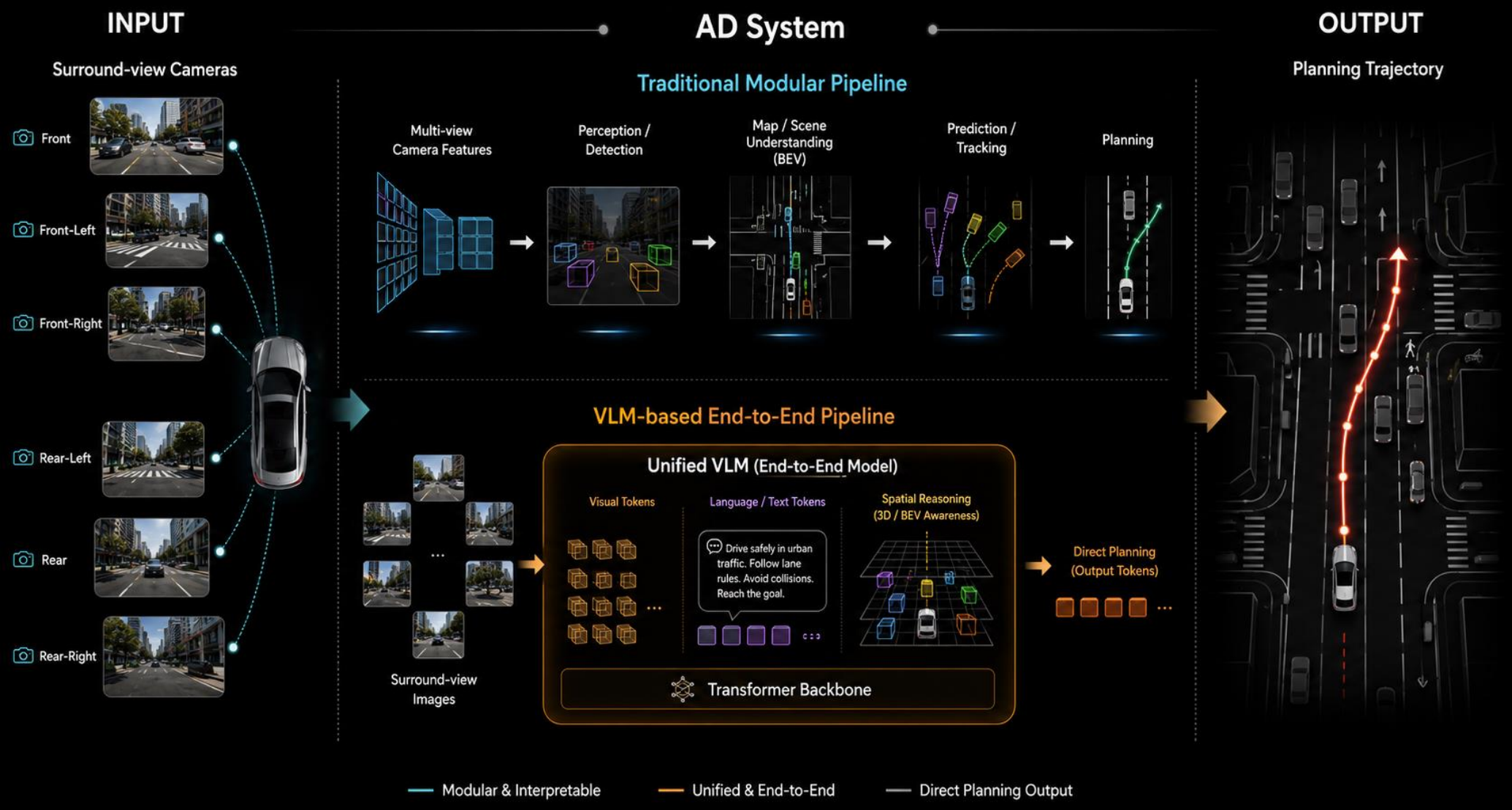
** Equal contribution*

Mercedes-Benz

Paper

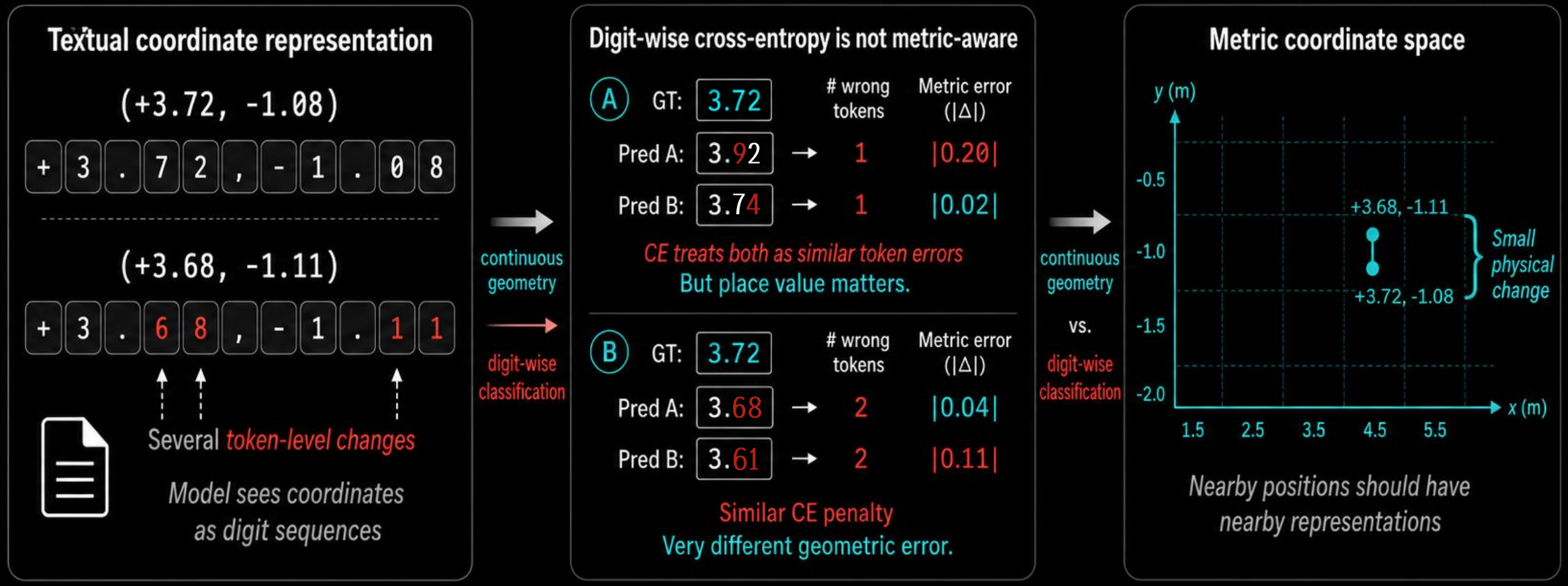
Code

End-to-End Planning



Bottleneck of VLM-based Pipeline

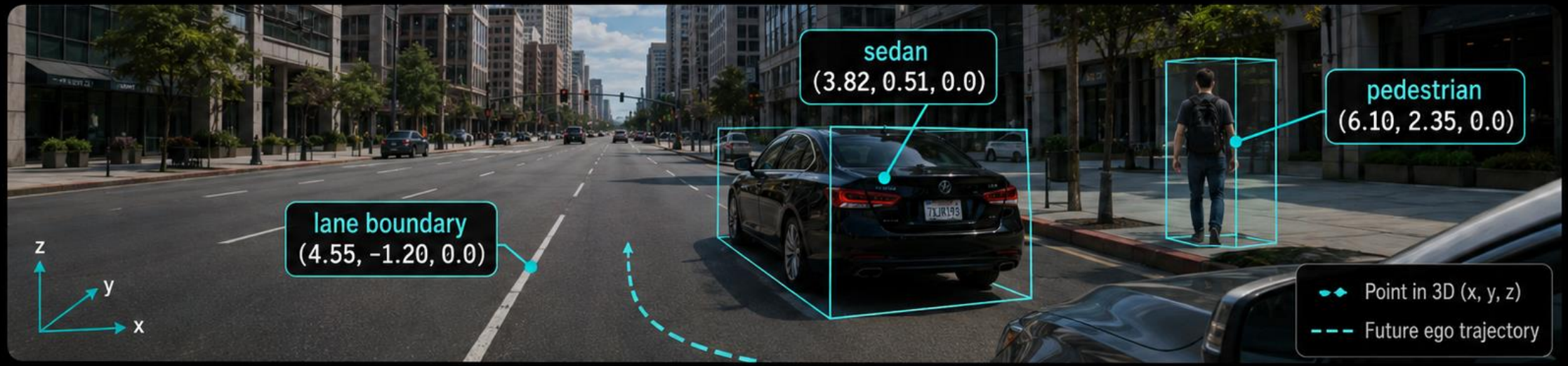
I: Textual Coordinates Break Metric Continuity



Digit-wise classification **dose not preserve** numerical metric continuity

Bottleneck of VLM-based Pipeline

II: Weak Association Between 2D Visual Semantics and 3D Coordinate



Q: What object is at (3.82 m, 0.51 m)?

Q: What happens if the ego follows this trajectory?

Text-only VLM

- ✕ Ambiguous object grounding
- ✕ Unstable trajectory reasoning

(3.82, 0.51, 0.0) query coordinate

trajectory prediction

versus

Desired spatial grounding

- ✓ Object ↔ 3D coordinate
- ✓ 3D coordinate ↔ trajectory constraint

(3.82, 0.51, 0.0) query coordinate

future ego trajectory

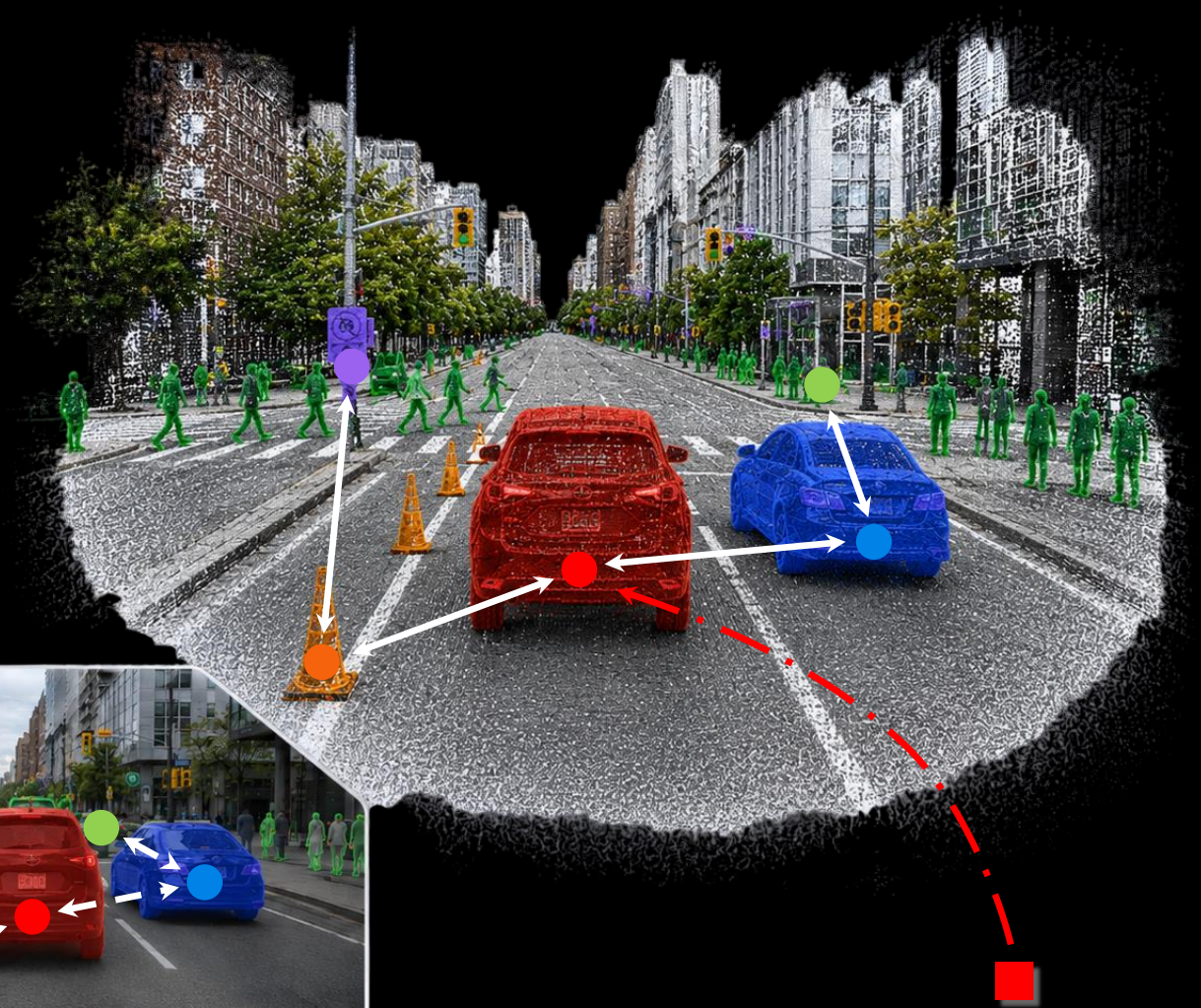
Recognition in 2D is not enough → planning requires <u>stable metric grounding</u>

Desired Positional Representation

coordinate

distance_{2d}

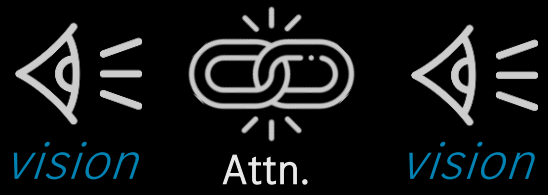
distance_{3d}



"What is located at (3.21, 0.08, 0) ?"

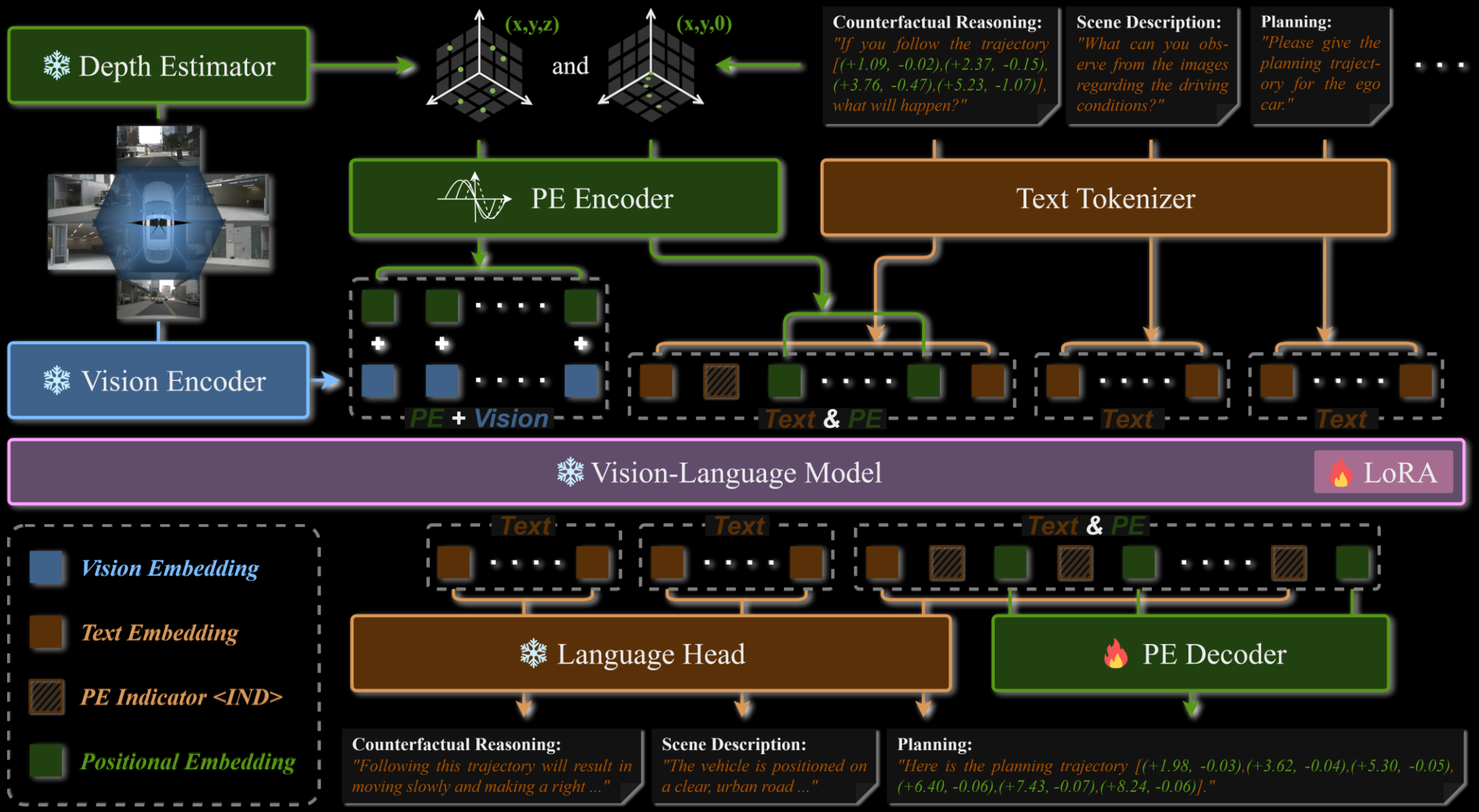
Desired Positional Representation

Explicit 3D spatial correlation



Universal PE representation

SpaceDrive Framework



Open-Loop Evaluation on nuScenes

Method	Ego Status		L2 (m) ↓				Collision (%) ↓				Intersection (%) ↓			
	BEV	Planner	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	2.53	8.17	14.40	8.37
UniAD [19]	✓	✓	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37	0.20	1.33	3.24	1.59
VAD-Base [27]	✓	✓	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33	0.21	2.13	5.06	2.47
AD-MLP [77]	-	✓	0.15	0.32	0.59	0.35	0.00	0.27	0.85	0.37	0.27	2.52	6.60	2.93
BEV-Planner++ [36]	✓	✓	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34	0.35	2.62	6.51	3.16
UAD [14]	✓	✓	0.13	0.28	0.48	0.30	0.00	0.19	0.61	0.27	0.13	1.08	2.89	1.37
Drive-WM [64]	✓	✓	0.43	0.77	1.20	0.80	0.10	0.21	0.48	0.26	-	-	-	-
VLM-based Paradigm														
EMMA [22]	-	-	**0.14**	**0.29**	0.54	**0.32**	-	-	-	-	-	-	-	-
RDA-Driver [21]	✓	✓	0.23	0.73	1.54	0.80	**0.00**	**0.13**	0.83	0.32	-	-	-	-
DriveVLM [60]	-	✓	0.18	0.34	0.68	0.40	0.10	0.22	**0.45**	0.27	-	-	-	-
ORION [12]	✓	-	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37	-	-	-	-
OmniDrive-Q [63]	-	-	1.15	1.96	2.84	1.98	0.80	3.12	7.46	3.79	1.66	3.86	8.26	4.59
OmniDrive-Q++ [63]	✓	✓	**0.14**	**0.29**	0.55	0.33	**0.00**	**0.13**	0.78	0.30	0.56	2.48	5.96	3.00
SpaceDrive (**ours**)	-	-	1.06	1.79	2.55	1.80	0.35	1.33	3.97	1.88	0.96	3.38	8.28	4.21
SpaceDrive+ (**ours**)	-	✓	0.15	**0.29**	**0.51**	**0.32**	0.04	0.18	0.49	**0.23**	**0.22**	**0.80**	**2.79**	**1.27**
Hybrid Paradigm														
VLP [46]	✓	-	0.30	0.53	0.84	0.55	0.01	0.07	0.38	0.15	-	-	-	-
ReAL-AD [44]	✓	-	0.30	0.48	0.67	0.48	0.07	0.10	0.28	0.15	-	-	-	-
DriveVLM-Dual [60]	✓	-	0.15	0.29	0.48	0.31	0.05	0.08	0.17	0.10	-	-	-	-
SOLVE-VLM [8]	✓	-	0.13	0.25	0.47	0.28	0.00	0.16	0.43	0.20	-	-	-	-
Senna [28]	✓	-	0.11	0.21	0.35	0.22	0.04	0.08	0.13	0.08	-	-	-	-

Compared to all traditional modular approaches or VLM-based methods, SpaceDrive+ achieves the best performance across all metrics.

Closed-Loop Evaluation on Bench2Drive

Method	Closed-loop Metric	
	Driving Score ↑	Success Rate(%) ↑
VLM-based Paradigm		
ReAL-AD [44]	41.17	11.36
Dual-AEB [80]	45.23	10.00
VDRive [15]	66.25	50.51
StuckSolver [3]	70.89	50.01
DriveMoE [72]	74.22	48.64
ETA [16]	74.33	48.33
VLR-Drive [30]	75.01	50.00
ORION [12]	77.74	54.62
SimLingo [49]	**85.07**	**67.27**
SpaceDrive+ (**ours**)	<u>78.02</u>	<u>55.11</u>

Without extensive optimizations designed for closed-loop planning, SpaceDrive+ ranks as the <u>2nd-best</u> VLM-based method simply by the infusion of spatial awareness, surpassing most traditional modular E2E approaches.

Ablations

The usage of positional encodings across different components

Exp.	$\phi(\mathbf{c}_p)$	$\phi(\mathbf{c}_r)$	$\phi(\mathbf{c}_\tau^{ego})$	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
SpaceDrive						
1				2.51	4.53	6.77
2	✓			1.88	2.45	2.36
3		✓		2.42	5.06	8.94
4	✓	✓		1.80	1.88	4.21
SpaceDrive+						
5				0.41	0.60	4.40
6	✓	✓		0.33	0.23	1.32
7	✓	✓	✓	0.32	0.23	1.27

$\phi(c_p)$, $\phi(c_r)$ and $\phi(c_\tau^{ego})$ denote 3D coordinates for vision tokens, spatial positional encodings for textual coordinates and past ego locations, respectively.

- **(Exp. 1 vs. 2 & Exp. 3 vs. 4)** Explicitly injecting 3D spatial information into visual tokens enhances the model's spatial understanding.
- **(Exp. 2 vs. 4 & Exp. 6 vs. 7)** With unified positional encoding (e.g., in text prompts and ego states), joint spatial reasoning further improves planning performance.

Ablations

Different PE encoder and decoder

Exp.	Encoder $\phi(\cdot)$	Decoder $\psi(\cdot)$	Avg. L2 ↓	Avg. Col. ↓	Avg. Int. ↓
4	Sine-Cosine	Coordinate-wise	1.80	1.88	4.21
8	MLP	Coordinate-wise	1.96	3.17	6.76
9	RoPE	Coordinate-wise	1.93	3.71	11.40
10	Sine-Cosine	Sine-Cosine	1.87	2.62	9.20
11	Sine-Cosine	Task-specific	1.93	2.41	5.58

Gray row indicates that only 4929 out of 5119 output samples are semantically reasonable, while the rest fail.

- **(Exp. 4 vs. 8)** As Sine-Cosine encoding inherently incorporates relative position modeling, it serves as more efficient spatial inputs compared to learnable encoders.
- **(Exp. 4 vs. 9)** RoPE leads to a large performance degradation and instability due to the confusion with existing RoPE used in the base VLM
- **(Exp. 4 vs. 10,11)** Coordinate-wise MLP decoders demonstrate greater advantages compared to task-specific decoders and inverting Sine–Cosine encodings with coarse interpolation.

Ablations

Different PE normalization

Init. α_{PE}	Learnable	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
1		2.34	3.63	8.46
0.1		2.43	3.79	9.42
0.02		2.22	2.71	10.17
1	✓	1.82	2.04	4.62
0.1	✓	1.80	1.88	4.21
0.02	✓	1.86	2.03	5.42

Gray row indicates that only 4929 out of 2421 output samples are semantically reasonable, while the rest fail.

- Since the optimal PE scale α_{PE} differs between setups and may shift over training.
- A learnable normalization factor with extra freedom can solve this issue, leading to better planning performance.

Ablations

Different depth estimator and depth pooling strategy

$f_{dep.}$	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
DepthAnythingV2 [71]	1.76	1.95	3.96
UniDepthV2 [48]	1.80	1.88	4.21

- Both DepthAnythingV2 and UniDepthV2 perform similarly on the L2 error metric and Collision rate. showing the effectiveness of our SpaceDrive is independent of a specific pre-trained depth module

Pooling Strategy	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
Min	1.80	1.88	4.21
Average	1.83	1.88	4.08
Median	1.83	2.17	4.64

- Min pooling yields the lowest L2 error and collision rate, as is relatively reliable and safety-critical as it preserves the closest obstacles

Robustness

Training and Inference with Depth Estimation Errors

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

- Strong inference-time robustness: depth shift / random noise causes negligible degradation (Avg. L2 remains 1.80).
- Depth mainly helps learn spatially grounded visual–text associations, rather than serving as an extra planning modality at inference. It shows SpaceDrive treats depth as **geometric guidance**, not a precise test-time dependency.

Adaptability

Different base VLM

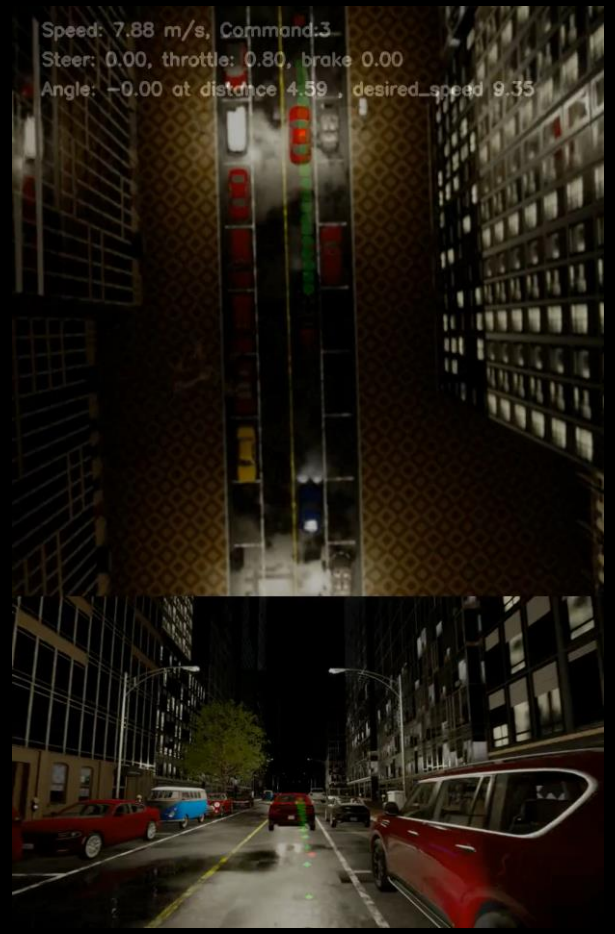
Method	Base VLM	Avg. L2 ↓	Avg. Collision ↓	Avg. Intersection ↓
SpaceDrive	LLaVA	1.82	2.44	4.08
	Qwen-VL	1.80	1.88	4.21
SpaceDrive+	LLaVA	0.31	0.23	1.42
	Qwen-VL	0.32	0.23	1.27

- For both SpaceDrive and SpaceDrive+, our approach **does not rely on** the specific foundation VLM.
- It demonstrates stable and superior performance with both LLaVA and Qwen-VL.

Visualization on Bench2Drive

Way Points Planning ▪▪▪▪▪▪ Trajectory Planning ▪▪▪▪▪▪

Visualization on Bench2Drive



Way Points Planning ▪▪▪▪▪▪

Trajectory Planning ▪▪▪▪▪▪

Takeaway

- Unified 3D Spatial Representation:

  Introduces a universal 3D Positional Encoding (PE) interface that unifies perception, reasoning, and planning, moving beyond the limitations of treating coordinates as discrete, non-continuous text tokens.
- Explicit Semantic-Spatial Association:

  Establishes a dense, explicit association between 2D perspective semantics and 3D physical coordinates by additively injecting depth-derived 3D PEs directly into visual tokens.
- Continuous Coordinate Regression:

  Replaces traditional digit-by-digit language generation with direct coordinate-level regression using a learnable PE decoder, ensuring numerical continuity and significantly enhancing the geometric precision and safety of planned trajectories.
- Generalizable & BEV-Free E2E Planning:

  Achieves SOTA open-loop results and 2nd best closed-loop performance across multiple VLM backbones (e.g., Qwen, LLaVA) without relying on dense Bird's-Eye-View (BEV) representations

References

Wang, Shihao, et al. "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning." *CoRR* (2024).

Wang, Shihao, et al. "Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

Zhang, Zhiyuan, et al. "Mpdrive: Improving spatial understanding with marker-based prompt learning for autonomous driving." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

Fei, Xiang, et al. "Advancing sequential numerical prediction in autoregressive models." *arXiv preprint arXiv:2505.13077* (2025).

Fu, Haoyu, et al. "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation." *arXiv preprint arXiv:2503.19755* (2025).

Tian, Xiaoyu, et al. "Drivevlm: The convergence of autonomous driving and large vision-language models." *arXiv preprint arXiv:2402.12289* (2024).