The 39th Conference on Neural Information Processing Systems

SURDS: Benchmarking Spatial Understanding and Reasoning in Driving Scenarios with Vision Language Models

Xianda Guo[1,*] , Ruijun Zhang[2,3,*], Yiqun Duan[4,*], Yuhang He[5], Dujun Nie[2,3],
Wenke Huang[1], Chenming Zhang[6,3], Shuai Liu[7] ,Hao Zhao[8], Long Chen[2,3,6,†]

(Wechat) [1] School of Computer Science, Wuhan University [2]CASIA [3]Waytous [4] University of Technology Sydney
[5] Microsoft Research [6] IAIR, Xi'an Jiaotong University [7] ByteDance [8] AIR, Tsinghua University

Introduction

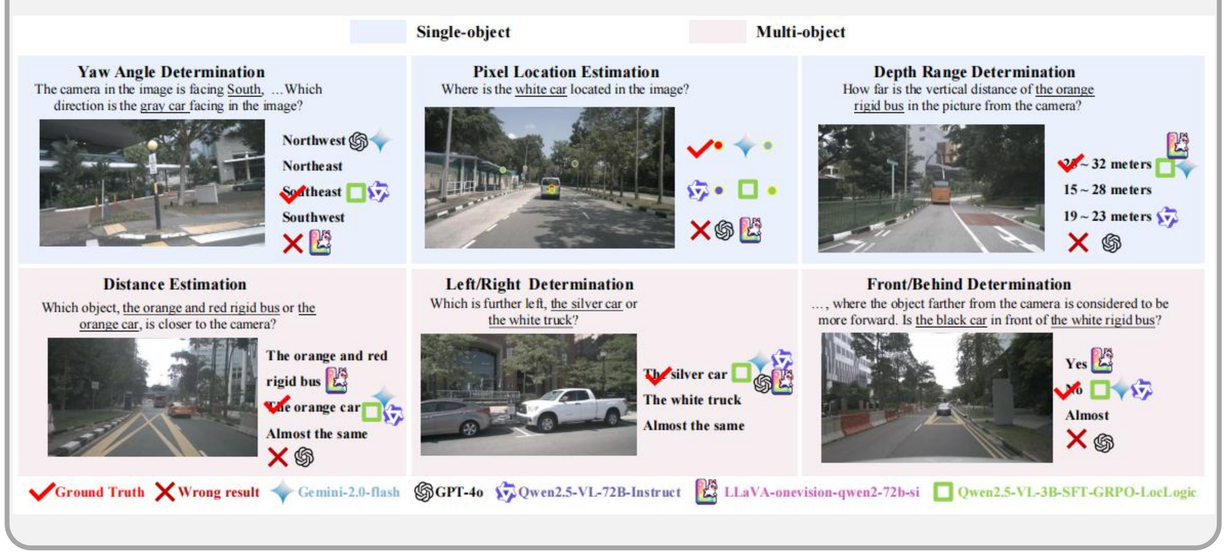
Paper	Scale	Annotation Types	Data Source	Reasoning	Method	w/o Depth	w/o Visual Mark
BLINK (ECCV2024) [27]	3,807	Image QA pairs	Web	✗	✗	✓	w/ Marked point
SpatialRGPT (NeurIPS2024) [18]	1,406	Image QA pairs	Web	✓	✓	✗	w/ Mask
SpatialBot (ICRA2025) [11]	174	Image QA pairs	Web	✓	✓	✗	w/ Marked point & Bbox
VSI bench (CVPR2025) [81]	~5,000	Video QA pairs	Indoor	✓	✗	✓	✓
SURDS (ours)	9,250	Image QA pairs	Driving	✓	✓	✓	✓

Table : Comparison between our work and other spatial understanding benchmarks. Reasoning indicates whether the framework focuses on reasoning. Method denotes whether it proposes a specific approach to enhance spatial understanding. w/o Depth means the framework does not use depth information during training or evaluation. w/o Visual Mark indicates no visual annotationsare added to the image.

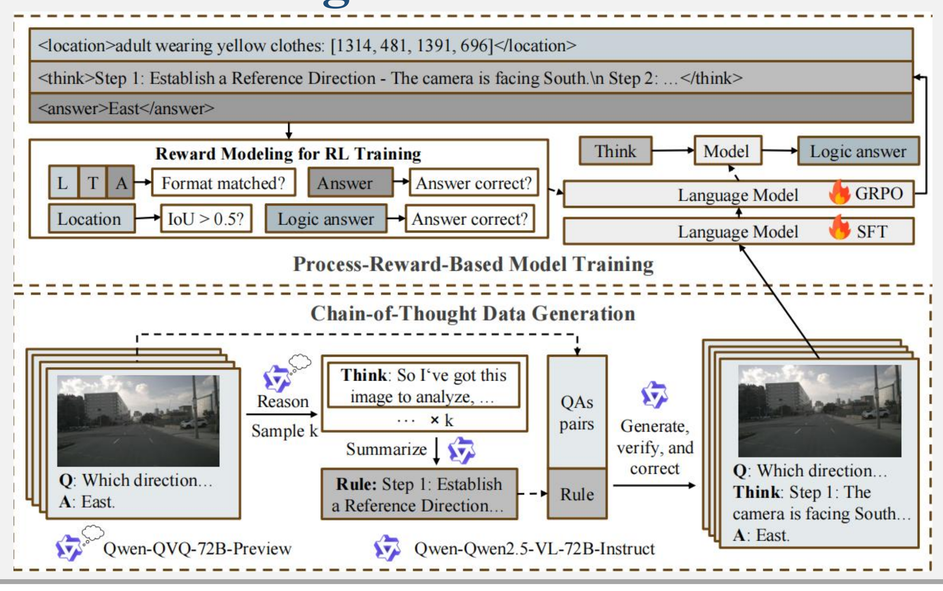
Contributions

- We propose SURDS, the first large-scale benchmark for evaluating fine-grained spatial understanding of VLMs in realistic driving scenarios, respectively, contains 41,080 training pairs and 9,250 test pairs.
- Our evaluations on SURDS reveal fundamental spatial reasoning limitations in existing models and demonstrate that model scale alone does not ensure spatial competence.
- Comprehensive experiments investigated different training strategies from supervised finetuning to reinforcement learning post-train alignment, providing valuable reference for follow-up researchers.

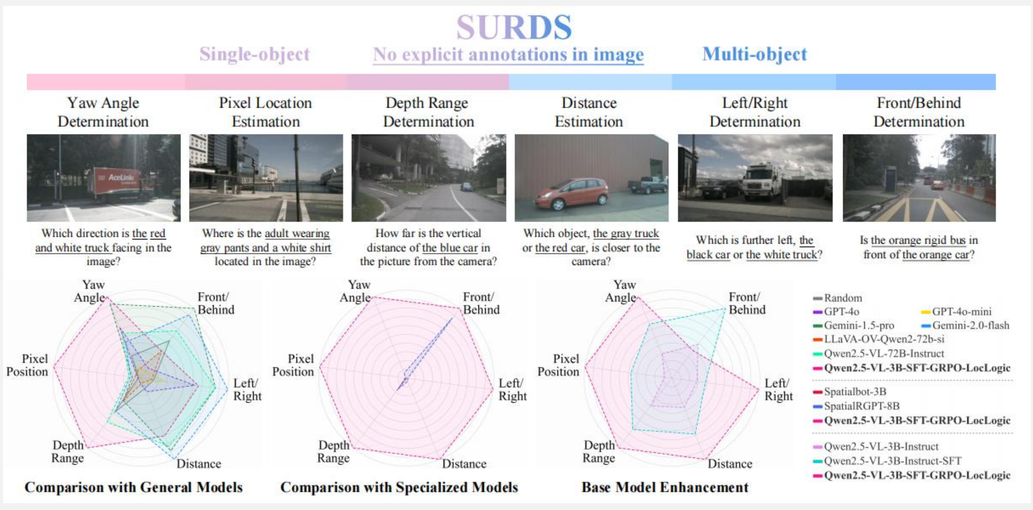
Illustrative examples of the benchmark QA pairs on both single-object and multi-object



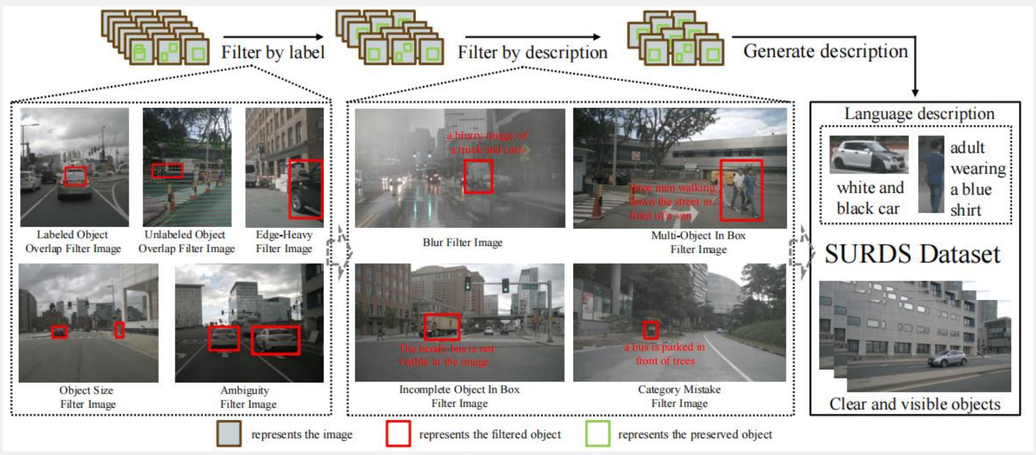
Overview of data generation and model training



Overview of the SURDS benchmark and the proposed method's performance.



Overview of the pipeline for constructing the SURDS dataset



Comparison of our proposed method with other open-source and proprietary VLMs, as well as specialized spatial understanding models

Model	Single-object			Multi-object			Score
	Yaw	Pixel	Depth	Dis	L/R	F/B	
Random	5.73	1.12	34.27	8.76	11.57	11.89	12.22
GPT-4o	13.08	1.62	2.49	11.57	47.89	3.14	13.30
GPT-4o-mini	3.24	0.28	0.22	4.22	21.51	2.05	5.25
Gemini-1.5-pro	19.14	4.41	22.70	61.95	66.38	**22.05**	32.77
Gemini-2.0-flash	9.30	5.41	32.97	**69.30**	**77.30**	20.00	35.71
LLaVA-OV-Qwen2-72b-si	1.95	3.03	23.57	3.78	9.73	8.65	8.45
Qwen2.5-VL-72B-Instruct	11.57	6.13	44.00	58.05	66.16	14.92	33.47
Qwen2.5-VL-7B-Instruct	7.57	3.46	25.95	11.46	17.95	9.30	12.61
Qwen2.5-VL-3B-Instruct	6.27	3.81	27.68	17.84	14.81	10.49	13.48
SpatialBot [11]	0.00	0.00	12.00	0.00	0.00	0.00	2.00
SpatialRGPT [18]	1.30	0.55	10.59	1.95	0.86	7.35	3.77
Qwen2.5-VL-3B-SFT-GRPO-LocLogic	**20.97**	**44.81**	**69.84**	49.30	51.35	8.54	**40.80**

Ablation study on different training setups of our method, analyzing the impact of SFT, GRPO, and the addition of location and logic rewards (LocLogic)

Model	Single-object			Multi-object			Score
	Yaw	Pixel	Depth	Dis	L/R	F/B	
Qwen2.5-VL-3B	6.27	3.81	27.68	17.84	14.81	10.49	13.48
Qwen2.5-VL-3B-GRPO	14.59	3.75	29.19	35.68	39.89	22.70	24.30
Qwen2.5-VL-3B-GRPO-LocLogic	8.11	57.82	27.24	22.05	19.35	12.43	24.50
Qwen2.5-VL-3B-SFT	13.95	21.11	51.35	33.95	19.68	21.62	26.94
Qwen2.5-VL-3B-SFT-GRPO	19.24	15.02	62.59	39.14	32.65	9.30	29.66
Qwen2.5-VL-3B-SFT-GRPO-LocLogic	20.97	44.81	69.84	49.30	51.35	8.54	**40.80**

Ablation study on model performance under different reward settings after SFT

Base model	Training reward				Single-object			Multi-object			Score
	Format	Loc	Acc	Logic	Yaw	Pixel	Depth	Dis	L/R	F/B	
Qwen2.5-VL-3B-SFT	✗	✗	✗	✗	13.95	21.11	51.35	33.95	19.68	21.62	26.94
Qwen2.5-VL-3B-SFT	✓	✗	✓	✗	19.24	15.02	62.59	39.14	32.65	9.30	29.66
Qwen2.5-VL-3B-SFT	✓	✓	✓	✗	17.84	22.72	64.65	41.41	30.92	11.68	31.53
Qwen2.5-VL-3B-SFT	✓	✗	✓	✓	20.54	14.81	62.49	36.54	32.65	11.35	29.73
Qwen2.5-VL-3B-SFT	✓	✓	✓	✓	20.97	44.81	69.84	49.30	51.35	8.54	**40.80**
Qwen2.5-VL-3B-SFT†	✓	✓	✓	✓	18.16	22.61	62.05	39.14	32.22	15.14	31.55
Qwen2.5-VL-3B-SFT	✓	✓	✓	✓	20.97	44.81	69.84	49.30	51.35	8.54	**40.80**