LaViPlan : Language-Guided Visual Path Planning with RLVR

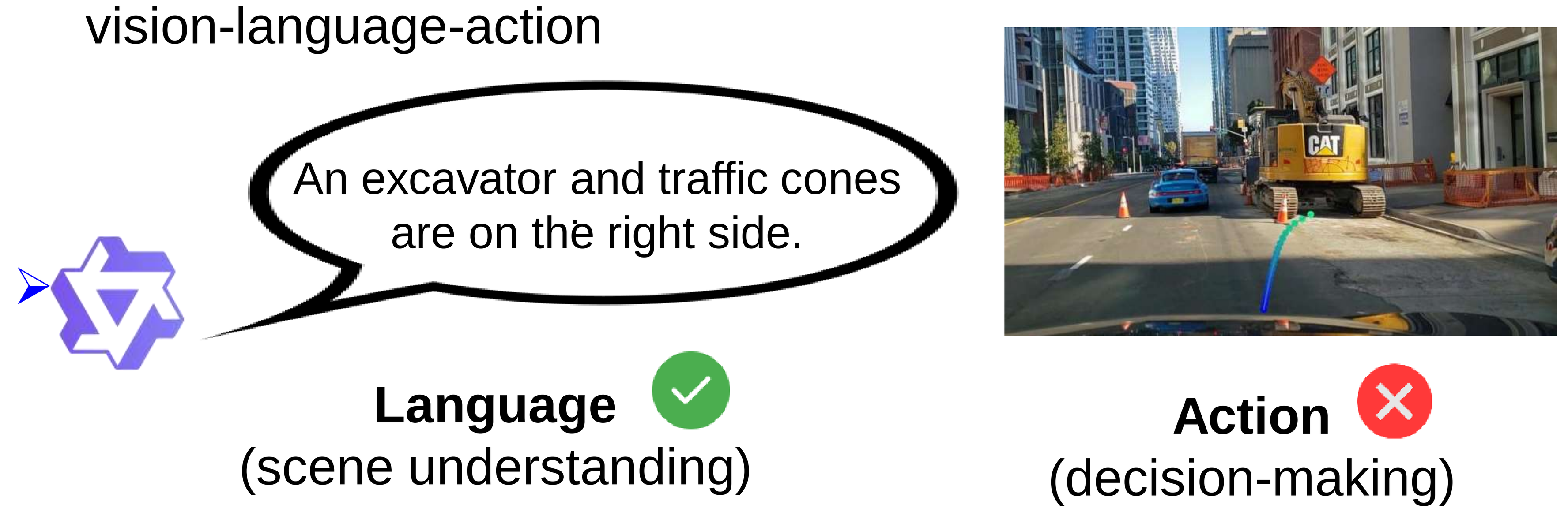
Hayeon Oh

Electronics and Telecommunications Research Institute

ICCV OCT 19-23, 2025 HONOLULU HAWAII

1. Motivation

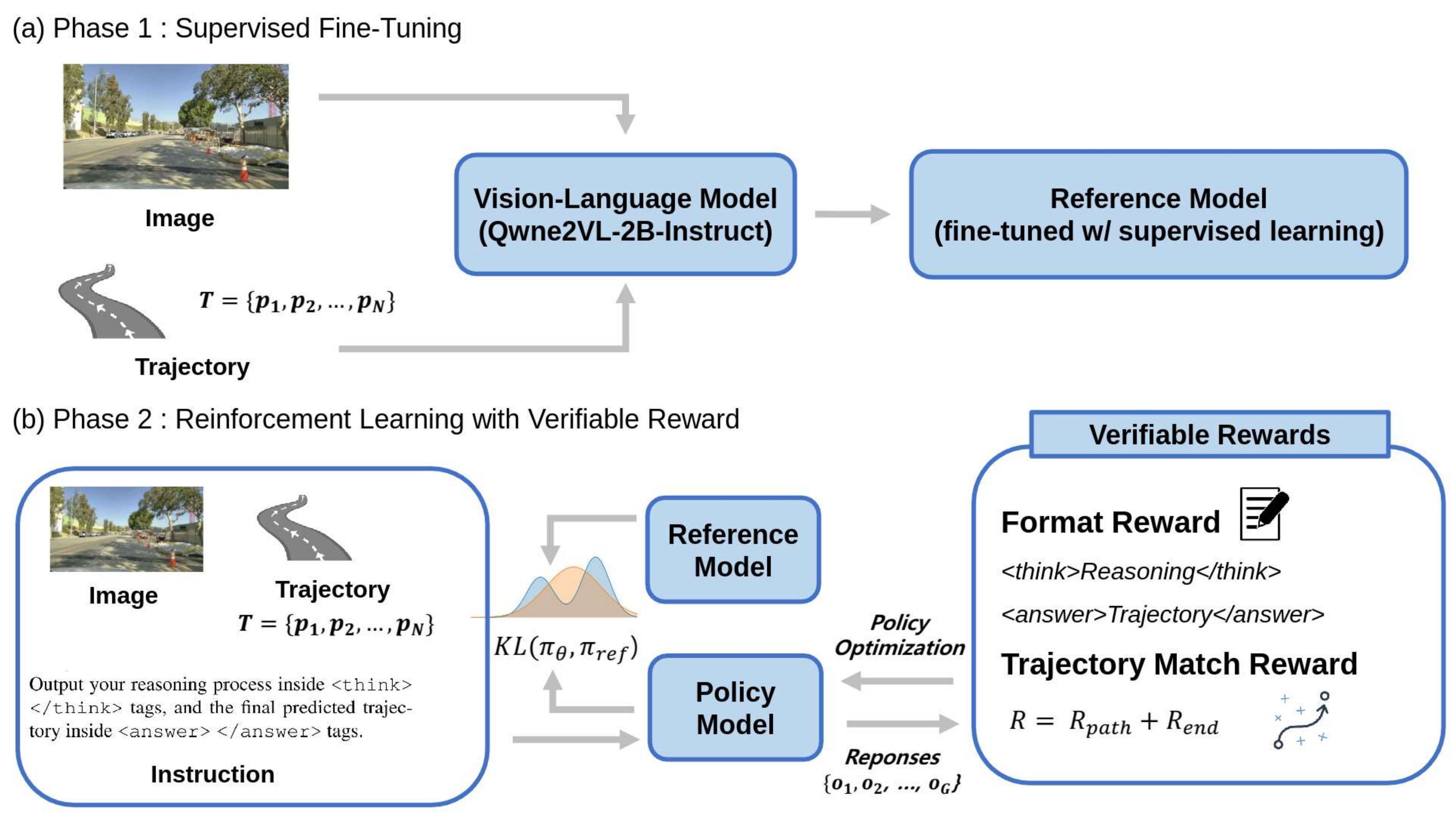
- LLM agent for autonomous driving has misalignment problem in vision-language-action



- Post-training with reinforcement learning (RL) has shown generalization, memory efficiency, and alignment (e.g., RLHF)

⇒ What if we leverage RL with LLM for autonomous driving?

2. Method



Phase 1 : supervised fine-tuning (SFT)

- Instruction tuning enables reasoning-guided waypoint prediction for path planning.

Phase 2 : reinforcement fine-tuning (RFT) with GRPO for vision-language-action alignment

- Post-training with the group relative policy optimization (GRPO) can align language and action by maximizing planning-oriented reward following the objective function below :

$$\max_{\pi_\theta} \mathbb{E}_{o \sim \pi_\theta(q)} \left[R_{\text{RLVR}}(q, o) \right]$$

$$= \left[R(q, o) - \beta \, \text{KL} \left(\pi_\theta(o \mid q) \,\|\, \pi_{\text{ref}}(o \mid q) \right) \right]$$

- The reward is based on image-plane displacement errors between predicted and ground-truth waypoints.

$$R_{\text{planning}} = -\log \left(1 + \frac{1}{N} \sum_{i=1}^{N} \|\hat{p}_i - p_i\|_2 \right) - \log \left(1 + \|\hat{p}_N - p_N\|_2 \right)$$

3. Experiment

Results in ROADWork (in-domain dataset)

- ADE : average displacement error, FDE : final displacement error

	ADE ↓		FDE ↓	
	Easy	Hard	Easy	Hard
Baseline				
Vision-Language Models				
Qwen2VL-2B	52.44	52.77	102.39	105.05
Qwen2VL-7B	60.73	60.71	66.61	67.57
Qwen2.5-VL-3B	16.37	16.40	20.60	20.77
LLaMA3.2-11B	59.27	58.88	74.16	71.44
Domain-Specific Models				
Senna	N/A	N/A	N/A	N/A
DriveLM (w/ LLaMA-Adapter)	37.10	38.40	56.99	56.90
Supervised Fine-tuning				
Vision-Language Models				
Qwen2VL-2B	4.52	5.66	4.46	6.46
Qwen2VL-7B	4.80	6.04	5.08	7.35
Qwen2.5-VL-3B	4.97	6.22	5.07	7.34
LLaMA3.2-Vision-11B	4.52	5.46	5.20	7.10
Domain-Specific Models				
Senna	5.71	5.73	6.58	7.46
DriveLM (w/ LLaMA-Adapter)	6.73	7.79	6.87	8.43
Reinforcement Fine-tuning				
LaViPlan (ours)	**3.62**	**4.83**	**3.85**	**6.09**

- All models used 5K samples: 4K for LaViPlan's SFT and 1K for its RFT

	ADE ↓		FDE ↓	
	Easy	Hard	Easy	Hard
Baseline	52.44	52.77	102.39	105.05
SFT (4k)	4.12	5.31	4.44	6.51
LaViPlan	**3.62**	**4.83**	**3.85**	**6.09**
Δ	-12.1%	-9.1%	-13.3%	-6.5%

- RFT after SFT yields performance gains across all scenarios

Ratio	ADE ↓		FDE ↓	
	Easy	Hard	Easy	Hard
9:1	3.84 (-6.8%)	5.05 (-4.9%)	4.09 (-7.9%)	6.31 (-3.1%)
7:3	5.55 (+34.7%)	6.70 (+26.2%)	4.05 (-8.8%)	6.16 (-5.4%)
6:4	3.62 (**-12.1%**)	4.83 (**-9.1%**)	3.85 (**-13.3%**)	6.09 (**-6.5%**)

- Effect of Easy-to-Hard Data Ratio (Fixed Total Samples)

Results in CODA-LM (out-domain dataset)

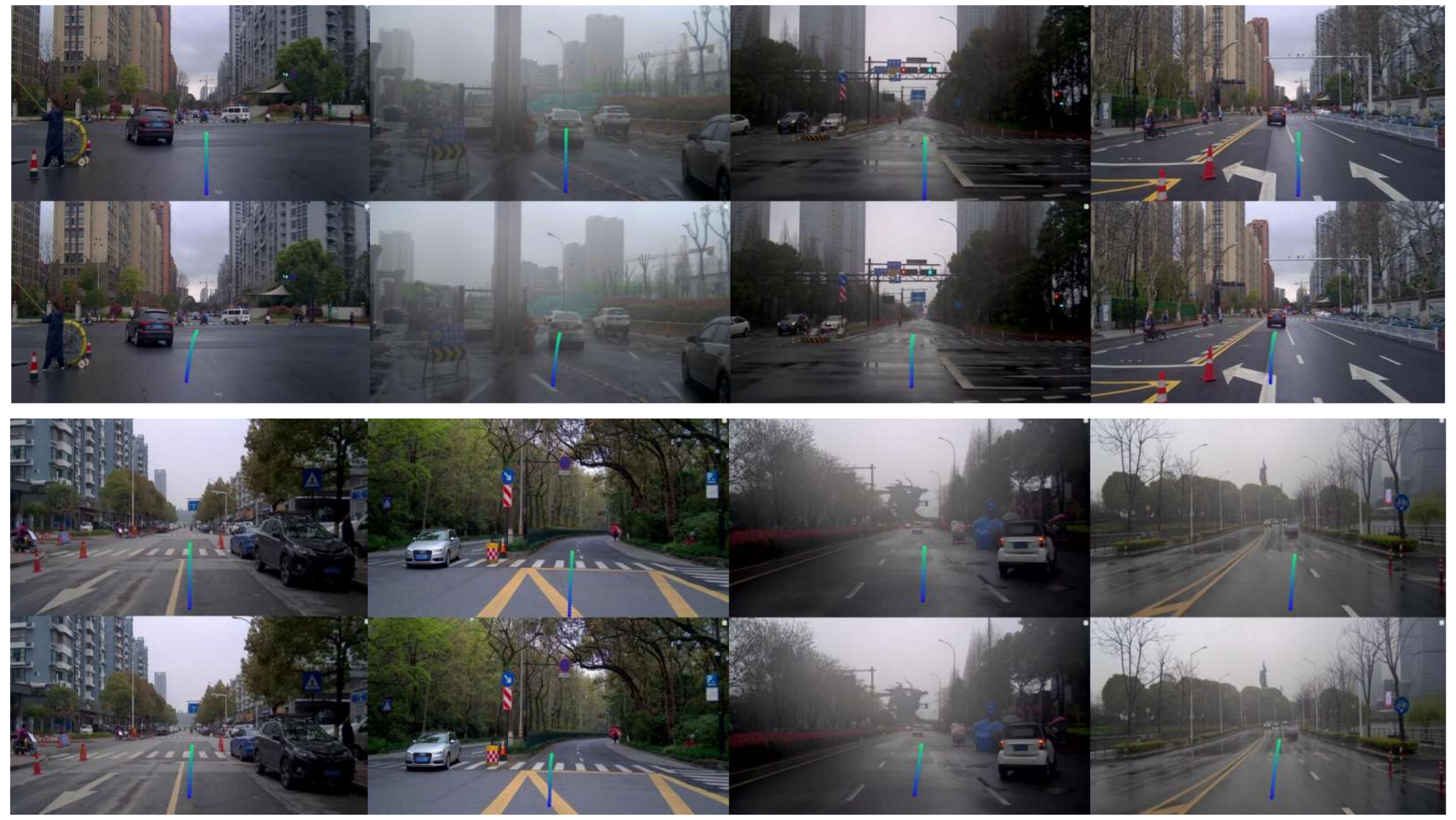
Model	Balanced↑	Safety-Focused↑	Performance-Focused↑	Equal↑
Baseline	0.40	0.30	0.50	0.33
SFT (5k)	0.60	0.59	**0.56**	0.63
LaViPlan	**0.64**	**0.73**	**0.56**	**0.70**

- Evaluation under varying penalty weights in zero-shot scenarios

Ratio / Model	Balanced ↑	Safety-Focused ↑	Performance-Focused ↑	Equal ↑
SFT (5K)	0.60 (+0.20)	0.59 (+0.29)	0.56 (**+0.06**)	0.63 (+0.30)
LaViPlan (9:1)	0.58 (+0.18)	0.62 (+0.32)	0.51 (+0.01)	0.63 (+0.30)
LaViPlan (7:3)	0.64 (**+0.24**)	0.73 (**+0.43**)	0.56 (**+0.06**)	0.70 (**+0.37**)
LaViPlan (6:4)	0.45 (+0.05)	0.49 (+0.19)	0.39 (-0.11)	0.51 (+0.18)

- Effect of Easy-to-Hard Data Ratio in out-domain dataset

Qualitative Analysis



- Trajectories before **(up)** and after RFT **(down)**, showing alignment

4. Conclusion

- **Summary** : leveraging GRPO to align vision, language, and action
- **Limitation** : spare reward depending on the entire rollout
- **Future work** :LLM agent for autonomous driving capable of causal and counterfactual reasoning for safe, interpretable decision