ELF-VLA: Unleashing VLA Potentials in Autonomous Driving via Explicit Learning from Failures

Yuechen Luo, Fang Li, Qimao Chen, Shaoqing Xu, Jiaxin Liu, Ziying Song, Zhi-xin Yang, Fuxi Wen

Code is available at https://github.com/luo-yc17/ELF-VLA

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction

What are VLA models in Autonomous Driving?

VLA (Vision-Language-Action) models use cameras to look at the road and directly output driving commands. They also use CoT to reason before they act.

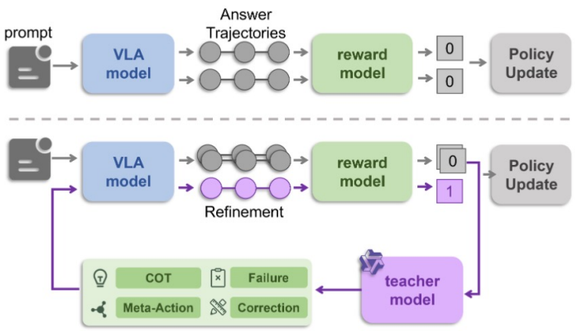

Figure 3. **Comparison between RL fine-tuning of general VLA and ELF-VLA.** Top: General VLA gets trapped in a performance plateau due to low-scoring rollouts. Bottom: ELF-VLA breaks this plateau using teacher model feedback for refinement.

Key Challenge: RL Training Gets Stuck

✓ **All-Zero Rewards**: In hard scenarios, the model fails every time and gets a driving score of zero.
✓ **Sparse Reward Problem**: A simple "zero" score provides no actionable feedback, leaving the model unaware of whether it failed due to a bad plan, wrong reasoning, or poor steering.

Stop Improving

Motivation and Method

Idea: Overcome the RL performance plateau by solving the "all-zero reward" problem in critical scenarios through explicit learning from failures.
✓ **Structured Diagnostic Feedback:** Replaces vague scalar rewards with teacher-guided reports pinpointing planning, reasoning, or execution errors.
✓ **Feedback-Guided Refinement:** Generates high-reward trajectories based on feedback and re-injects them into RL training for targeted gradient updates.

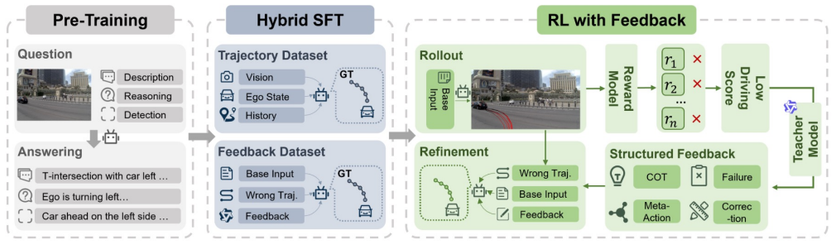

Figure 2. **Overview of ELF-VLA.** First, the model is pre-trained on an driving Q&A dataset for driving knowledge. Subsequently, it undergoes SFT on a mixed dataset of feedback inputs, learning trajectory prediction and feedback-based refinement. Finally, in the RL phase, a teacher model generates feedback, reducing zero-reward rollouts.

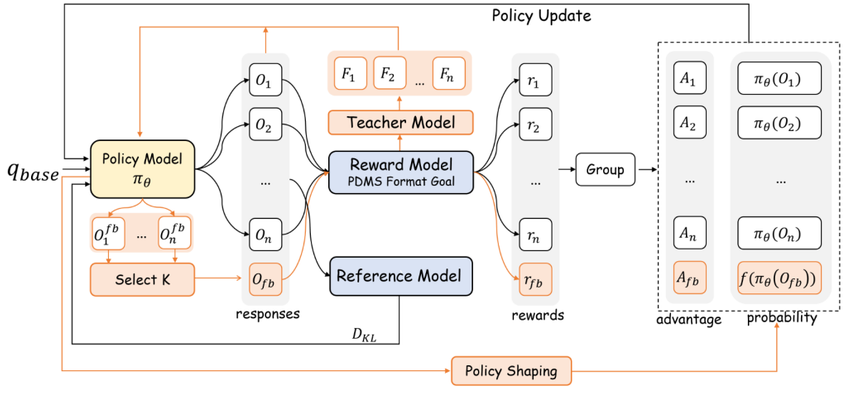

Figure 3. **Overview of GRPO with feedback.** The policy model generates initial responses. Based on rewards, the Qwen3VL-32B teacher provides feedback, guiding policy to sample improved responses. A high-quality response is combined with the initial set for joint optimization, followed by Policy Shaping.

Experimental Results

Table 1. Comparison with state-of-the-art methods on the NAVSIMv1 with PDMS.

Method	Image	Lidar	NC↑	DAC↑	TTC↑	CF↑	EP↑	PDMS↑
Constant Velocity			68.0	57.8	50.0	100	19.4	20.6
Ego Status MLP			93.0	77.3	83.6	100	62.8	65.6
UniAD [11]	✓		97.8	91.9	92.9	100	78.8	83.4
TransFuser [6]	✓	✓	97.7	92.8	92.8	100	84.0	84.0
DiffusionDrive [22]	✓	✓	98.2	96.2	94.7	100	82.2	88.1
WoTE [18]	✓	✓	98.5	96.8	94.9	99.9	81.9	88.3
Hydra-NeXt [21]	✓		98.1	97.7	94.6	100	81.8	88.6
AutoVLA-3B [41]	✓		98.4	95.6	98.0	100	81.9	89.1
DriveVLA-W0-3B [36]	✓		98.7	**99.1**	95.3	99.3	83.3	90.3
GoalFlow [36]	✓	✓	98.4	98.3	94.6	100	85.0	90.3
InternVL3-8B-SFT	✓		98.5	95.5	95.3	100	81.2	87.4
InternVL3-8B-RL	✓		98.5	96.7	95.4	100	83.2	89.0
ELF-VLA-8B(Ours)	✓		**98.9**	98.1	**96.0**	**100**	**85.3**	**91.0**

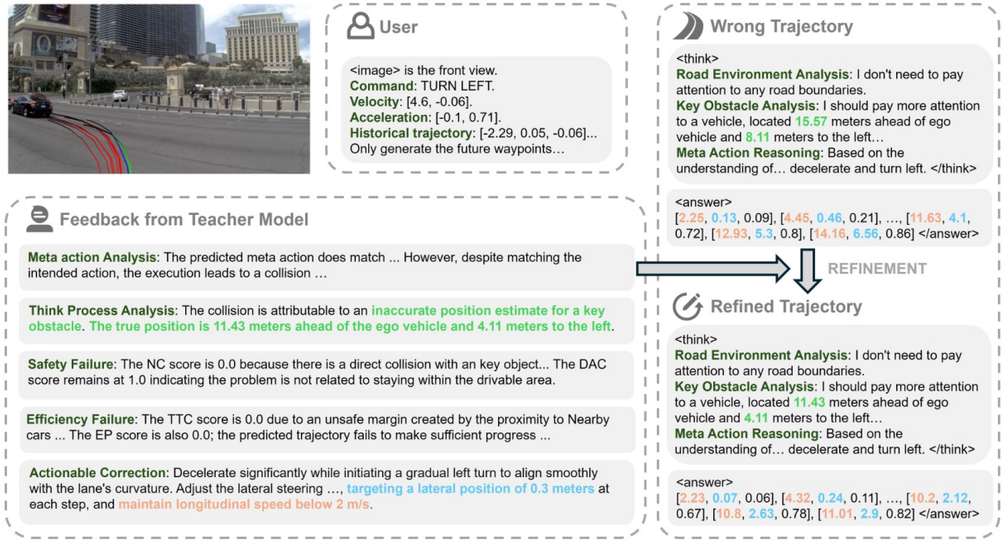

Figure 4. Visualization of trajectory refinement process by ELF-VLA on the NAVSIM dataset. Visualization of the initial Wrong Trajectories (red), the Ground Truth (green), and the final Refined Trajectory (blue).