Devil is in Narrow Policy: Unleashing Exploration in Driving VLA Models

Supplementary Material

A. Training Implement Details

we provide more detailed configurations for both the Imitation Learning (SFT) and Reinforcement Learning (RL) stages.

A.1. Training Stages on SFT

Since the base Qwen2.5-VL model does not inherently support the specific `<think>` token, we introduce external tokens (`<thinking></thinking>` to wrap the *driving explanation* part, `<answer></answer>` to wrap the *trajectory prediction* part) to the tokenizer. To ensure the model effectively learns this structured reasoning format without compromising its visual encoding capabilities, we adopt a two-stage SFT strategy:

- **Thinking Alignment.** In this stage, we freeze the vision encoder and the projector while optimizing only the LLM backbone. It primarily focuses on aligning the model with the structured CoT format and external tokens, ensuring the LLM captures the syntactic structure of the thinking process without disturbing the pretrained visual representations.
- **End-to-End Fine-tuning.** In the second stage, we unfreeze all parameters to end-to-end fine-tuning. This step enables joint optimization of the vision encoder and LLM, effectively bridging the vision-language space while adapting the pretrained knowledge to the domain shift of autonomous driving.

A.2. Hyper-Parameters

We provide the detailed hyper-parameters for SFT in Table 7 and GRPO in Table 8.

Table 7. Training configurations for SFT.

Setting	SFT
Base Model	Qwen2.5-VL-3B
Max Pixels	262144
Global Batch Size	128
Epochs	3(align) / 3(fine-tune)
Trainset Samples	103k(navtrain) + 39k(DE)
Learning Rate	4e-5 / 5e-6
Weight Decay	0.05
Warmup Ratio	0.10
bfloat16	✓
Global Batch Size	128
GPUs	8

Table 8. Training configurations for GRPO.

Setting	GRPO
Max Pixels	262144
Rollout Batch Size	256
Actor Global Batch Size	256
N Rollout(Group Size)	8
Outer Loops	3
Total Steps	130
Active Samples	6k/3k/1k (for 3 outer-loop)
bfloat16	✓
GPUs	8

A.3. Details on Span Driving Reward

In Sec. 4.2, we redesign the reward function by adapting the EPDMS into an additive focal objective with strict safety constraints. Furthermore, we remove EC for training efficiency, and the sparse reward R_{sparse} defined by Eq. 10 uses $C' = \{\text{NC, DAC, DDC, TLC}\}$, and $M' = \{\text{EP, TTC, C, LK}\}$, the weights are $w'_m = \{5, 5, 2, 2\}$ and focal exponents are $\gamma_m = \{0.5, 0.5, 1.0, 1.0\}$.

B. Data Pipeline

B.1. Details on Exploratory Data Expansion

Semantic-Aware Challenging Sceneario Filtering. We first identify 12k challenging driving segments (e.g., multi-lane roads, intersections, occlusions) from the 103k `navtrain` split. Although visual grounding models like Grounding-DINO can detect road elements, and the Navsim dataset itself provides rich SemanticMapLayers annotations, these sources cannot directly filter for challenging scenes that semantically possess "multiple feasible trajectories." Therefore, we employ the following prompt with Qwen2.5-VL-72B to screen for these scenarios for data expansion. As shown in the prompt in Fig. 5, the VLM is instructed to focus on complex road elements and identify scenarios that allow for diverse maneuvers.

Generative Trajectory Expansion. To construct a robust exploratory dataset, we implement a rigorous expansion pipeline using the DDIM planner from ReCogDrive. To induce behavioral variance beyond deterministic outcomes, we modify the standard DDIM sampling by scaling the standard deviation of the Gaussian noise injected during the reverse denoising steps. The expansion process includes:

1. **Hybrid Sampling Strategy:** We apply distinct sampling methods based on scene complexity. For the *entire* `navtrain` dataset, we perform **intra-intent** ($k = 32$)

Prompt and Examples for Sceneario Filtering

Question:
<image> Give you a scene during driving. Observe the provided image to identify challenging scenarios, such as intersections, multi-lane roads, and occlusions. Pay strict attention to the ego vehicle's current lane, ground markings, and traffic signs. Provide your analysis strictly in the format below. Do not add any extra explanations. Always make conservative decisions. Unless the non-lane-keeping driving intention is very safe and necessary, only stay in the lane.

Analysis Guidelines

- An "intent" is an **immediately executable and traffic-compliant** high-level maneuver (e.g., turn left, change lane).
- Do **not** list intents that are clearly unreasonable or illegal (e.g., turning left from a straight-only lane, or changing lanes over a solid line).

Output Format

1. SCENE_SUMMARY: [Provide a brief (200-word max) description of the ego vehicle's situation.]
2. ALL_INTENTS: [List all plausible driving intents, separated by commas.]
3. NON_LANE_KEEPING_INTENT: [Based on the list in #2, is there any valid intent other than lane following? Answer "Yes" or "No"]

Answer Example 1:

1. SCENE_SUMMARY: The ego vehicle is on a two-lane road with a clear view ahead. The road is relatively empty, with no immediate traffic or pedestrians in sight. On the left side, there are trees, a sidewalk, and streetlights, while the right side has some vegetation and a N̈o Parkings̈ign. The road markings include a double yellow line in the center and a single yellow line on the right edge. The sky is partly cloudy, and a bridge is visible in the distance. The overall environment appears calm and safe for driving. There are no immediate obstacles or hazards on the road.
2. ALL_INTENTS: continue straight
3. NON_LANE_KEEPING_INTENT: No

Answer Example 2:

1. SCENE_SUMMARY: The ego vehicle is driving on a busy urban street, likely in a commercial area with multiple lanes and traffic signals. The vehicle is positioned in the right lane, following a red pickup truck. To the left, there is an orange van with advertisements, and to the right, there are various commercial buildings, including a Starbucks and a Planet 13 dispensary. The traffic signal ahead is red, indicating that vehicles must stop. The road markings are clear, with solid and dashed lines indicating lane boundaries. There are no immediate obstacles or pedestrians in the immediate vicinity of the ego vehicle. The overall environment suggests a typical city street with moderate traffic.
2. ALL_INTENTS: stop at the red light, change lane left, change lane right.
3. NON_LANE_KEEPING_INTENT: Yes

Figure 5. The prompt template and representative answer examples used to filter semantic-aware challenging scenarios with multiple feasible driving intents.

inference using the original prompts to explore execution variations within the same intent. For the 12k filtered challenging scenes, we additionally execute **inter-intent** inference by systematically altering the intention prompts (switching among *Go Straight*, *Turn Left*, *Turn Right*, and *Unknown*) to uncover plausible intents.

2. **Safety and Diversity Filtering:** All generated candidates undergo a dual-criteria filter. For safety, a trajectory is retained only if its PDMS score exceeds 95.0 and not less than the score of the human GT. For diversity, we employ a greedy selection based on geometric distance. We remove trajectories near the human GT and iteratively retain candidates, pruning those within a predefined distance margin.

This pipeline ultimately yields **39k** non-ground-truth yet physically feasible exploratory samples, significantly broadening the policy's behavioral coverage beyond the narrow human ground truth.

B.2. Details on Chain-of-Thought

Following Poutine [39], we adopt a structured reasoning approach to explicitly model the decision-making process. The model processes the input through input context and four thinking tasks:

Input Context. The input $\mathcal{X}$ comprises a single *CAMERA-FRONT* image, the ego-vehicle's kinematic history (past 1.5s trajectory, current velocity and acceleration), and the high-level driving intent.

Thinking Tasks. The CoT follows this sequence:

1. **Critical Object Perception:** Identify the presence of critical objects [39] that might influence the future path.

Table 9. Inference Latency Comparison. **Text**: Text waypoint; **Action**: Action token. AutoVLA uses a Fast-Slow Dual-System.

Method	Setting	Latency (s)
AutoVLA	Dual-Sys (Text)	9.31
AutoVLA	Dual-Sys (Action)	3.95
AutoVLA	Dual-Sys (Action + RFT)	**1.31**
Curious-VLA	Slow Think only (Text)	<u>1.57</u>

2. **Driving Explanation:** Generate a concise, natural-language rationale (approximate 100 words).
3. **Meta-Behavior Description:** Classify the intended behavior into discrete categories [39], specifically selecting the appropriate Speed and Command.
4. **Trajectory Prediction:** Predict the optimal 4-second trajectory waypoints (normalized) conditioned on the previous steps.

.

C. Inference Efficiency

We evaluate real-time efficiency under serial inputs. As shown in Tab. 9, Curious-VLA achieves 1.57s per sample, which is 7.74s faster than AutoVLA's text waypoint mode and competitive with its optimized Action+RFT mode (1.31s). The efficiency gain primarily benefits from our concise output template and single-view (1x C) input.

D. RL Training Stability

We provide the RL training curves to demonstrate the stability of our training pipeline. As shown in Fig. 6, we report Val Reward and Test PDMS over the entire 130 training steps (ADAS 3x, 3 outer-loops as described in Alg. 1). Outer-loop transitions are determined by Val Reward trends. In contrast, the *Random Sample* baseline (without ADAS) shows RL collapse, confirming the necessity of diversity-aware sampling.

We further analyze training stability across multiple runs in Fig. 7. We perform $k = 4$ independent training runs (ADAS 1x). The Critic (on trainset) and Val Reward curves demonstrate consistent improvement with low variance across all trials.

E. External Analytical Experiments

We further extend the analysis to DiffusionDrive [31] in Tab. 10. Despite its diverse 20-candidate pool(@all), the final Top-1 confidence-selected trajectory(@1) collapses to a single mode with the lowest diversity (0.037 / 0.076 mean-pADE/FDE), confirming that the narrow policy problem persists even in diffusion-based planners. In contrast, Curious-VLA achieves a superior balance across Quality, Diversity, and Performance.

Table 10. Extended exploration analysis. DiffusionDrive is evaluated with its 20 denoised candidates(@all) and the confidence Top-1 selection(@1). All metrics @k=8.

Method	Output	Quality	Diversity	Perf.
DiffusionDrive	Diff.@all	0.218 / 0.430	0.571 / 1.175	87.60
DiffusionDrive	Diff.@1	0.350 / 0.720	0.037 / 0.076	88.10
Curious-VLA	AR@1	**0.269 / 0.547**	**0.641 / 1.415**	**91.55**

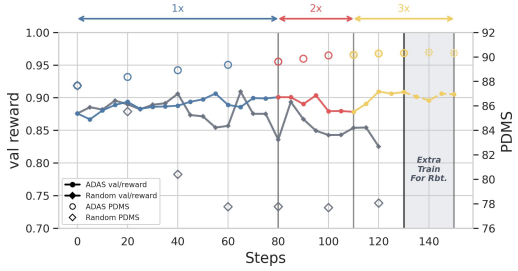

Figure 6. **RL Training Curves.** Val Reward and Test PDMS over 130 steps (ADAS 3x). The *Random Sample* baseline shows RL collapse.

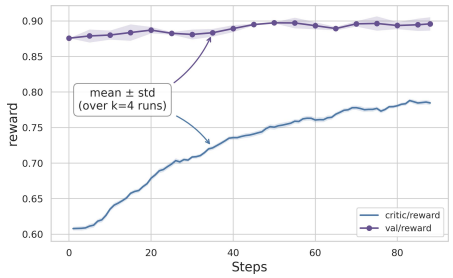

Figure 7. **Stability Analysis.** $k = 4$ training runs (ADAS 1x). Critic and Val Reward curves show consistent improvement with low variance.

F. More Visualization of Curious-VLA

As shown in Fig. 8, **Curious-VLA** successfully alleviates the narrow policy bottleneck, ensuring diverse feasible behaviors.

Figure 8. More visualization between Curious-VLA(top) and Qwen2.5-VL(bottom).