NoRD: A Data-Efficient Vision-Language-Action Model that Drives without Reasoning

Ishaan Rawal[1,2*] Shubh Gupta[1] Yihan Hu[1] Wei Zhan[1,3†]

[1]Applied Intuition [2]Texas A&M University [3]UC Berkeley

Abstract

Vision-Language-Action (VLA) models are advancing autonomous driving by replacing modular pipelines with unified end-to-end architectures. However, current VLAs face two expensive requirements: (1) massive dataset collection, and (2) dense reasoning annotations. In this work, we address both challenges with NoRD *(**No Reasoning** for **D**riving). Compared to existing VLAs,* NoRD *achieves competitive performance while being fine-tuned on <60% of the data and no reasoning annotations, resulting in 3× fewer tokens. We identify that standard Group Relative Policy Optimization (GRPO) fails to yield significant improvements when applied to policies trained on such small, reasoning-free datasets. We show that this limitation stems from difficulty bias, which disproportionately penalizes reward signals from scenarios that produce high-variance rollouts within GRPO.* NoRD *overcomes this by incorporating Dr. GRPO, a recent algorithm designed to mitigate difficulty bias in LLMs. As a result,* NoRD *achieves competitive performance on Waymo and NAVSIM with a fraction of the training data and no reasoning overhead, enabling more efficient autonomous systems.*

1. Introduction

The prevailing paradigm for end-to-end autonomous driving is increasingly shifting toward Vision-Language-Action (VLA) models. The dominant training methodology for these models is a two-stage training pipeline: (1) Supervised Fine-Tuning (SFT) on large-scale datasets with detailed, natural language Chain-of-Thought (CoT) reasoning annotations [41, 45, 51], followed by (2) a Reinforcement Learning (RL) stage to align outputs with driving metrics, for which Group Relative Policy Optimization (GRPO) [15] has been widely adopted [36, 51].

While this paradigm has achieved state-of-the-art performance on complex driving benchmarks [36], its reliance on both massive data and dense reasoning introduces three non-scalable costs:

1. *Data cost* of collecting and curating vast quantities of specialized driving scenarios
2. *Annotation cost* from generating high-quality reasoning traces for this data
3. *Training and inference cost* from resulting reasoning tokens, increasing training time and creating inference latency that is impractical for real-world deployment

This motivates a natural hypothesis: **Can we achieve competitive performance on driving benchmarks while being both reasoning-free and data-efficient?**

This investigation is supported by two distinct lines of work. First, recent studies provide a theoretical motivation by questioning the necessity of explicit reasoning, suggesting it may be a byproduct of planning rather than a causal determinant [38]. Second, existing work on end-to-end models like EMMA [18] and SimLingo [35], provides

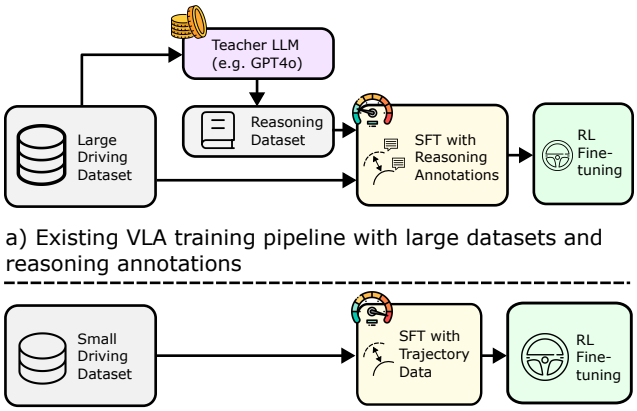


Figure 1. **Comparison of VLA training pipelines.** (a) Existing approaches depend on large-scale reasoning data generation, followed by extensive SFT and RL fine-tuning. (b) In contrast, NoRD directly utilizes a small-scale driving dataset for SFT, and performs RL fine-tuning tailored for weak SFT policy, enabling data-efficient learning without reasoning supervision.

*Work done during an internship at Applied Intuition.

†Corresponding author. Email: wei.zhan@applied.co

an empirical precedent by achieving strong performance on nominal benchmarks without reasoning. We therefore investigate if this data-efficient, reasoning-free approach can be extended to the more challenging benchmarks (e.g. NAVSIM [11], WaymoE2E [30]) that are currently dominated by their reasoning-centric counterparts.

We initially trained a reasoning-free NoRD-BASE VLA (based on Qwen-2.5VL-3B-Instruct [2]) using only SFT on 80,000 NAVSIM training samples; a greater than 60% reduction in data compared to state-of-the-art reasoning-based models [51]. This model was then post-trained with GRPO to optimize the PDM score [11].

However, this data-efficient, reasoning-free model achieves scores significantly lower than reasoning-based baselines (>12-point difference), and post-training with GRPO only results in a meager improvement (+0.67%). This initial failure creates the illusion that reasoning data is a necessary component for achieving high performance.

In this work, we argue that this conclusion is premature. We posit that the failure lies not in the reasoning-free SFT policy, but rather in the interaction between the policy optimization method (GRPO) and the reward landscape. We find that the complex, sparse reward signals from driving benchmarks (like PDM score from NAVSIM or the RFS from WaymoE2E) induce a highly polarized distribution of intra-group rewards. A significant mass of these mean rewards is clustered at the extremes (i.e., near 0 or 1), and correspond to rollouts with low variance. Conversely, the remaining scenarios, which yield intermediate mean rewards, are characterized by high-variance rollouts. When GRPO is applied in this landscape to a weaker, data-efficient SFT policy like NoRD-BASE, the resulting learning signal disproportionately penalizes the intermediate-mean (high-variance) scenarios, impeding effective optimization.

We are the first to identify that the failure to optimize weak SFT mode is caused by polarized intra-group reward landscape, and that it stems from *difficulty bias*, which has also been observed in LLM reasoning domain [22, 29]. Based on our analysis, we propose to mitigate this bias by using Dr. GRPO [29], an existing policy optimization algorithm specifically designed to address this flaw. We demonstrate that by applying Dr. GRPO as a drop-in replacement, our reasoning free VLA, NoRD, can be successfully trained.

Our key contributions are as follows:

1. We are the first to identify that the failure of reasoning-free and data-efficient VLA training for autonomous driving is an instance of difficulty bias, triggered by the combination of a weak SFT policy and complex driving metrics.
2. We empirically characterize this failure, showing that the data-efficient SFT policy induces a polarized reward distribution that deprives GRPO of a learning signal.
3. We propose using Dr. GRPO as a drop-in replacement to train NoRD, a data-efficient, reasoning-free VLA, and are the first to validate this policy optimization method in the autonomous driving domain (see Fig. 1).
4. We demonstrate performance competitive with the state-of-the-art on the NAVSIM and WaymoE2E benchmarks without using any reasoning annotations and at least 60% less data than reasoning VLAs, while improving on inference time, proving the viability of our approach.

2. Related Works

Reasoning-based VLAs. Several works have incorporated high-level reasoning into the control loop, including hybrid architectures like ORION [14], unified transformers like AutoVLA [51], and a wide variety of reasoning strategies, such as retrieval-augmented CoT [9], spatio-temporal reasoning [47], multi-agent reasoning [34, 49], and models combining memory and tool use [23, 28]. While this approach has achieved state-of-the-art performance on complex driving benchmarks [36], its reliance on large-scale, specialized reasoning datasets [6, 41, 45] and the high inference latency of CoT generation [26, 31] motivate the exploration of alternatives.

Reasoning-Planning Disconnect. The high cost of reasoning-centric models has spurred an inquiry into their necessity, with recent work questioning whether the model's reasoning improves its planning output. One study [38], proposing a "Reasoning-Planning Decoupling Hypothesis" demonstrated that textual priors alone can match the performance of full multimodal reasoning models. This skepticism extends to RL post-training, as other works argue that RL does not instill new reasoning capacity but instead optimizes within the SFT model's existing latent distribution [46]. These findings motivate our reasoning-free approach and frame our central question: does the failure to align weak SFT models stem from an inherent limitation of these models, or from an optimization failure?

Reasoning-Free VLAs. A separate line of VLA models operate without explicit reasoning traces. This includes models that map raw sensor data directly to trajectories like EMMA [18], SimLingo [35], and S4-Driver [43], as well as generative approaches like ADriver-I [19], DrivingGPT [5], and DiffVLA [20]. While these methods have demonstrated strong performance on nominal driving benchmarks like nuScenes [3], they have not yet proven competitive on the complex, long-tail benchmarks where reasoning-centric models currently excel.

Data Efficient VLAs. Our work aims to close the performance gap between reasoning-free and reasoning-based methods while maintaining data efficiency. Many data-

Table 1. Comparison of RL fine-tuning (RLFT) on NoRD-BASE with GRPO and Dr. GRPO on NAVSIM test set. While GRPO fails to improve NoRD-BASE, we get significant gains with Dr. GRPO.

Model	PDMS ↑
NoRD-BASE	76.66
NoRD-BASE + GRPO	77.18 (+0.67%)
NoRD-BASE + Dr. GRPO	85.62 (+11.68%)

efficient VLAs in broader domains [12, 13, 42, 44] mitigate data scarcity by leveraging massive external out-of-domain datasets. We instead focus on the distinct and more challenging problem of training a competitive model using only small-sized specialized, in-domain driving data.

Mitigating Difficulty Bias. The literature to mitigate the difficulty bias in GRPO, largely from the LLM reasoning domain, is divided into two main strategies. Data-level interventions attempt to manage data before the optimization, using methods like online filtering of saturated or degenerate samples [10, 27], curriculum learning [32], or advanced sampling [1]. These approaches are designed for binary rewards and are generally computationally infeasible for expensive driving simulations, as they often require multiple rollouts to estimate sample difficulty. In contrast, algorithmic-level interventions modify the optimization algorithm itself. This includes reweighting schemes [48, 50], alternative objectives [8, 22], or difficulty-based priors [4]. Our work incorporates this second strategy. We select Dr. GRPO [29], a lightweight method that directly corrects the bias by identifying and adjusting the specific normalization term in the advantage estimation responsible for it. Dr. GRPO is thus a prime candidate for our setting, as it avoids the infeasible overhead of data-level methods.

3. Limitations of GRPO for Data-Efficient Training

VLAs for autonomous driving have achieved competitive performance through a two-stage training pipeline - first SFT followed by RL post-training. This paradigm relies on large-scale domain-specific datasets that are additionally annotated with reasoning data. During RL post-training, GRPO optimizes the SFT model (*i.e.*, the policy) for high-level objectives such as preference alignment or safety by maximizing the group-relative advantage.

However, this existing SFT-heavy approach is costly and inefficient. First, collecting and labeling thousands or even millions of driving scenarios is resource-intensive. Second, generating reasoning traces from a teacher model increases token load, training time, and compute requirements. Finally, reasoning tokens during inference add latency, limiting real-time deployment. These challenges raise our central question: *can VLAs fine-tuned on small-scale driving data without reasoning supervision achieve competitive performance, or is RL post-training inherently limited in optimizing weaker, data-efficient VLAs?*

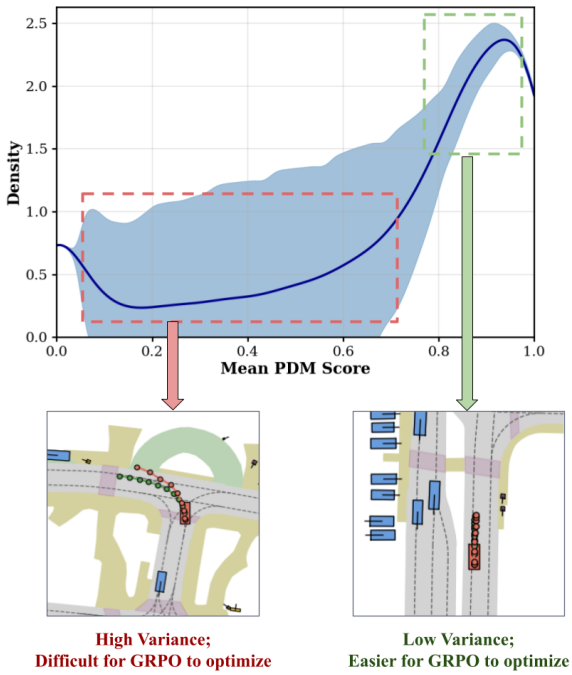
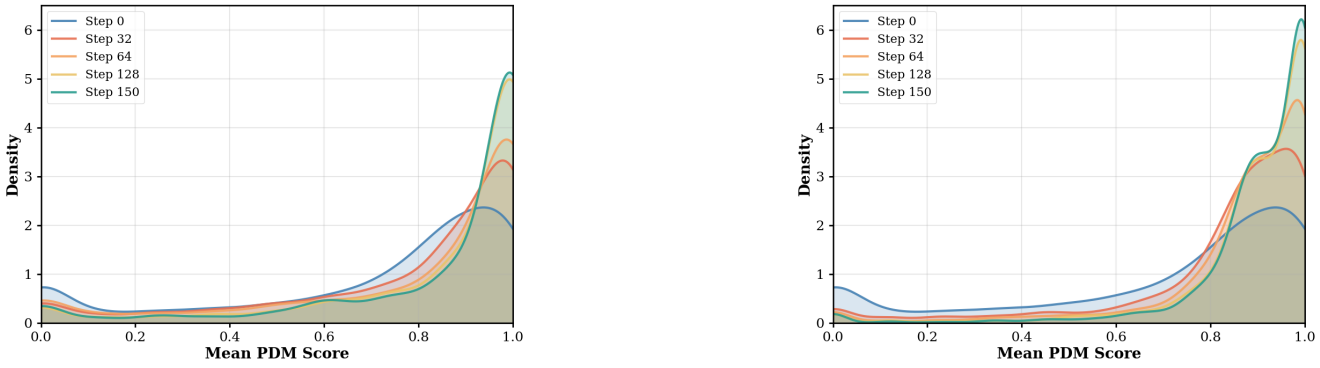


Figure 2. **Reward distribution in the weak SFT model.** The group-mean PDM score is shown with band representing the mean of the corresponding group standard deviation for NoRD-BASE. GRPO struggles to optimize high-variance regions (the majority) and is effective only in low-variance regions (the trajectories in green and red are for ground truth and NoRD-BASE prediction).

To investigate this, we train NoRD-BASE, a VLA built on Qwen-2.5VL-3B-Instruct [2], using supervised fine-tuning on only 80,000 NAVSIM training samples and without reasoning annotations. NoRD-BASE predicts physical trajectory tokens from current images and historical vehicle states, followed by GRPO optimization on the PDM score (details in Sec. 4.2). The PDM score evaluates predicted trajectories in simulation across metrics such as safety, comfort, collision avoidance, and adherence to driving areas, with higher scores indicating better performance.

As shown in Tab. 1, GRPO post-training leads to only a 0.67% improvement, resulting in negligible overall gains. This outcome is inconsistent with our goal of shifting the primary learning burden from SFT to RL post-training. This minimal improvement is in stark contrast to prior works like AutoVLA [51], which have demonstrated a 9% performance boost by post-training an SFT model, but one that was trained on 212,000+ samples and with reasoning data. This discrepancy between GRPO's effectiveness on strong versus weak SFT policies motivates a deeper investigation



(a) GRPO (b) Dr. GRPO

Figure 3. **Evolution of group-mean PDM score during RL fine-tuning.** (a) GRPO struggles to optimize samples with high group variance during training, particularly in the range $[0.2–0.65]$. (b) Dr. GRPO effectively optimizes high-variance samples during training, resulting in significant overall performance gains.

(a) Left Turn

(b) Lane Change

Figure 4. **Qualitative comparison of RL fine-tuning (RLFT) on the weak SFT model using GRPO and Dr. GRPO.** With Dr. GRPO, NoRD successfully learns complex maneuvers such as sharp turns and lane changes without collisions, whereas GRPO fails to optimize the weak SFT model (NoRD-BASE) and collides (in red).

into the underlying cause.

To understand this discrepancy, we analyze the reward characteristics of the training set for NoRD-BASE, shown in Fig. 2. For each training example, we perform 8 rollouts and plot the distribution of group-mean PDM scores, with bands indicating the corresponding group standard deviation, averaged across groups. Our key observations are:

1. **Low variance occurs in samples with high or very low group mean (≥ 0.8 or ≤ 0.15).** The weak SFT model performs reliably on simple behaviors, such as maintaining a straight trajectory at constant speed, resulting in high mean reward and low intra-group variance. Conversely, in extremely difficult scenarios, *e.g.*, out-of-distribution driving, the model predicts trajectories with trivially low PDM scores, yielding low mean and low variance within the rollout group. Notably, the proportion of such samples is very small, suggesting that NoRD is sufficiently expressive.
2. **High variance occurs in samples with intermediate group-mean values in the range $[0.2, 0.65]$.** The PDM score penalizes collisions, off-road behavior, and other safety violations. Given the weakness of the SFT model, complex maneuvers such as sharp turns fail more often than they succeed, producing low mean rewards and high variance within the rollout group.

With these observations, we analyze the evolution of group-mean reward distributions across GRPO training steps (see Fig. 3a). We find that the density in the high-variance region ($[0.2, 0.65]$) remains largely unchanged throughout training, whereas the density in the lowest-variance region (close to 1) steadily increases. This pattern explains the marginal improvement in the final PDM score after GRPO post-training: GRPO primarily optimizes the small subset of samples with low intra-group reward variance while failing to improve the majority of samples with high intra-group variance.

Our findings suggest that GRPO is fundamentally ineffective at learning from samples with high intra-group variance, which dominate the training dataset for our weak SFT model, NoRD-BASE, and therefore provides limited benefit for RL post-training. We interpret this failure as a form of *difficulty bias* in GRPO. Originally, difficulty bias was proposed for binary reward settings, measured as the mean group reward [22], and used to post-train LLMs for mathematical reasoning [29]. Building on this, our analysis shows

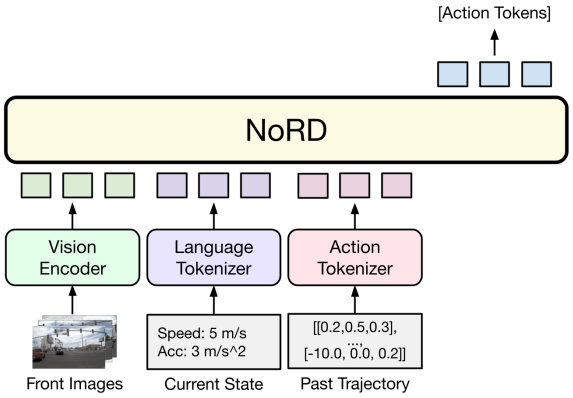


Figure 5. **Model architecture of NoRD.** NoRD directly predicts action tokens without requiring reasoning traces, enabling a significantly more efficient training and inference pipeline.

that the limitation of optimizing weak SFT policies with GRPO for data-efficient VLA training stems from this inherent difficulty bias. To address this, we post-train NoRD-BASE using Dr. GRPO, a GRPO variant originally designed to mitigate difficulty bias in LLM reasoning. Dr. GRPO enables training with significantly less data and without any reasoning annotations, as explained in the next section.

4. NoRD: No Reasoning for Driving

NoRD (**No** **R**easoning for **D**riving) is our Vision-Language-Action (VLA) model for autonomous driving, built upon Qwen-2.5VL-3B-Instruct. NoRD achieves both token and data efficiency by omitting reasoning annotations entirely from the training and inference stages, by emphasizing learning during the RL post-training phase rather than during supervised fine-tuning. However, as discussed in Sec. 3, naively reducing the amount of SFT data significantly degrades performance because GRPO is ineffective in learning from samples with high intra-group variance. This section presents the design of NoRD and approach for training it effectively with limited data.

The inputs to NoRD are the past ego-trajectory, current speed, acceleration, and RGB images from the front, front-left, and front-right cameras, as shown in Fig. 5. The model predicts the future ego-trajectory at 10 Hz. To improve token efficiency, we represent trajectories using k-disc tokenization [33] with a vocabulary size of 2048. Specifically, all future trajectories in the training set are first interpolated to 10 Hz and segmented into 0.5 second intervals. These segments are then clustered into 2048 clusters based on the contour distance between trajectory segments. The resulting cluster centers form a discrete trajectory codebook that can reconstruct any trajectory using vocabulary tokens. These trajectory tokens are appended to the original vocabulary of the base Qwen model initialized by sample from a multivariate normal distribution parameterized by the mean and covariance of the existing token embeddings [16]. The model is trained in two stages: (1) *Supervised Fine-Tuning* with limited data, followed by (2) *RL Post-Training* using Dr. GRPO for effective policy optimization starting from a weak SFT model, as illustrated in Fig. 1.

4.1. Supervised Fine-Tuning with Limited Data

NoRD-BASE is intentionally trained on a limited dataset during supervised fine-tuning to offload the majority of learning to the subsequent RL post-training phase. We model trajectory prediction as a next-token prediction problem, where the model outputs trajectory tokens conditioned on the inputs. As expected, the reduced training data results in lower initial performance for NoRD-BASE (see Tab. 1). In the following section, we describe how to effectively optimize this weak SFT policy using Dr. GRPO by explicitly accounting for intra-group reward variance.

4.2. RL Post-Training for Weak SFT Policy

As discussed in Sec. 3, weak SFT policies cannot be effectively optimized using standard GRPO, which can be understood as an instance of the difficulty bias problem. To address this, we employ Dr. GRPO, a recently proposed RL fine-tuning algorithm, for post-training our weak SFT model, NoRD-BASE. In the original GRPO formulation, the group relative advantage is computed as

$$\hat{A}_{i,t}^{\text{GRPO}} := \frac{r(o_i \mid x) - \frac{1}{G}\sum_{j=1}^{G} r(o_j \mid x)}{\text{std}_{j=1,\ldots,G}(r(o_j \mid x))}.$$

Here, $r(o_i \mid x)$ is the reward for sample i given input x, G is the group size, and std denotes the standard deviation across the group. Recent studies have shown that this formulation unintentionally favors groups with low reward variance [29]. When the standard deviation of the reward within the group is small (*i.e.*, $<< 1$), the group-relative advantage is disproportionately large, whereas it is heavily attenuated for groups with high reward variance. This poses a major problem for us since NoRD-BASE, being a weak SFT model, produces groups with high intra-group variance during the GRPO rollout for the majority of samples (Fig. 2). Dr. GRPO mitigates difficulty bias by removing the standard deviation term from the group relative advantage, enabling more effective optimization of weak policies. Notably, while other variants like VD-GRPO [39] preserve absolute reward magnitudes to balance objective priorities (e.g., safety vs. comfort), Dr. GRPO ensures that 'hard' scenarios contribute a sufficient gradient signal. Additionally, we employ DAPO-style asymmetric clipping to prevent entropy collapse during RL training and follow Liu et al. [29] by not using KL-divergence regularization. The resulting Dr. GRPO post-training objective is given by:

$$\hat{A}_{i,t}^{\text{DrGRPO}} = r(o_i \mid x) - \frac{1}{G}\sum_{j=1}^{G} r(o_i \mid x), \tag{1}$$

$$L_{\text{DrGRPO}} = \sum_{t=1}^{|o_i|} \min\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})} \hat{A}_{i,t}^{\text{DrGRPO}}, \right.$$
$$\left. \text{clip}\Big(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}, 1-\epsilon_{\text{l}}, 1+\epsilon_{\text{h}} \Big) \hat{A}_{i,t}^{\text{DrGRPO}} \right) \tag{2}$$

This formulation enables NoRD, *i.e.*, NoRD-BASE trained with Dr. GRPO, to achieve improved performance during RL post-training by mitigating difficulty bias and stabilizing policy optimization for data- and token-efficient VLAs. We find that with Dr. GRPO finetuning, NoRD-BASE learns from mid-variance samples, leading to an overall improvement of 11.68% from the base model (as compared to 0.67% with GRPO). We notice that Dr. GRPO is able to optimize even on the samples with high intra-group variance, as shown in Fig. 3b. This enables NoRD even learn complex maneuvers, as compared to the GRPO counterpart (see Fig. 4).

5. Experiments

5.1. Datasets

NAVSIM [11]: NAVSIM is a curated redistribution of the OpenScenes dataset, comprising real-world urban driving scenarios. The dataset comprises 120 hours of driving data from OpenScene. The dataset features diverse and challenging traffic situations, providing synchronized 360° camera imagery, LiDAR scans, HD map data, and bounding-box annotations of dynamic agents, along with historical control signals. The task is to predict the ego-vehicle trajectory for the next 4 seconds at 2 Hz, with performance evaluated by executing the predicted trajectory within a simulation environment, scored using PDM Score in terms of driving safety, progress and comfort metrics.

Waymo Vision-Based End-to-End Dataset (WaymoE2E) [30]: WaymoE2E is a challenging long-tail dataset providing 360° camera views and ego-vehicle trajectories. Validation and test scenarios include three alternative trajectories labeled with human per scene, representing varying driving preferences and scored in the range $[4, 10]$. Predicted trajectories are evaluated using the Rated Feedback Score (RFS), which measures weighted similarity with the reference preference trajectories. WaymoE2E contains 4,021 challenging driving segments, partitioned into 2,037 training, 479 validation, and 1,505 test segments.

Table 2. Test results on the Waymo Vision-based End-to-End Driving Benchmark. NoRD achieves competitive performance, without reasoning or ensembling.

Model	w/o Reason	w/o Ensemble	RFS↑	ADE@3↓
Poutine [36]	✗	✓	7.986	1.2055
HMVLM [40]	✗	✓	7.736	1.3269
DiffusionLTF	✓	✗	7.717	1.3561
UniPlan	✓	✗	7.692	1.3083
AutoVLA [51]	✗	✓	7.556	1.3507
NoRD	✓	✓	7.709	1.2504

5.2. Implementation Details

We use Qwen-2.5VL-3B-Instruct as our base model, as it offers a good trade-off between model capacity and computational efficiency. The model is fine-tuned on the NAVSIM and WaymoE2E datasets separately using 16 A100 GPUs with a batch size of 128. We employ the AdamW optimizer with a learning rate of 5×10^{-5} and a cosine decay schedule, fine-tuning all layers of Qwen-2.5VL-3B. This stage yields the NoRD-BASE model. Subsequently, we apply Dr. GRPO for RL post-training. For NAVSIM, we use 30 A100 GPUs to optimize the base model for 160 steps with a constant learning rate of 5×10^{-6}. For WaymoE2E, we post-train the model for 150 steps on 32 GPUs with a learning rate of 1×10^{-6}.

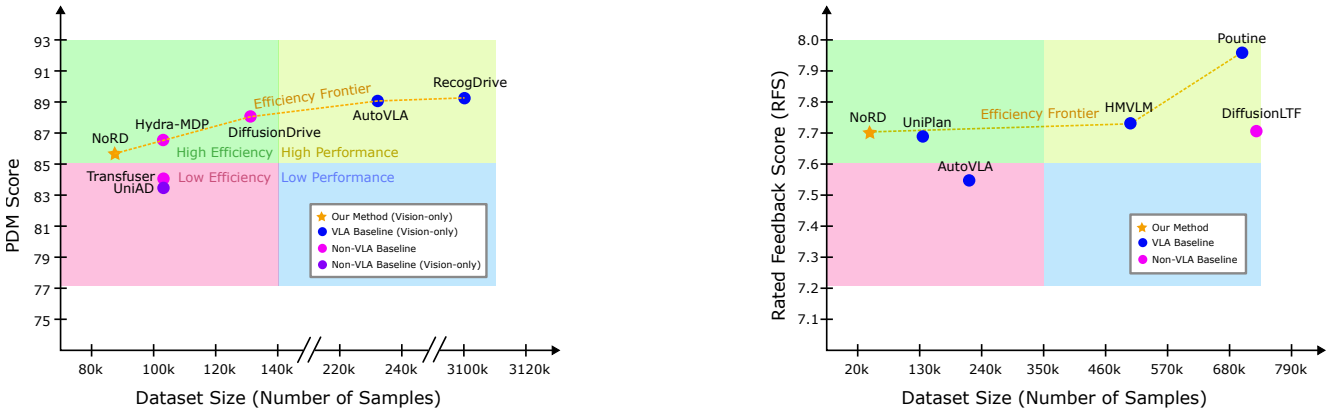
The RL post-training pipeline is implemented in `verl` [37] with Fully Sharded Data Parallel (FSDP) for memory-efficient training and `vLLM` [21] for rollout generation. During rollouts, we set the group size to 8 and fix the sampling temperature to 1.0. For validation, we employ deterministic sampling with a temperature of 0.01. For NAVSIM, we use the PDM score as the primary reward function, while for WaymoE2E, we employ the normalized RFS score. In both cases, we include additional rewards for trajectory length and output format, each weighted by 0.25, and then normalize to $[0, 1]$.

5.3. Results

WaymoE2E performance: The WaymoE2E dataset poses a challenging evaluation setting that emphasizes robustness under out-of-distribution driving scenarios. Consequently, most existing approaches rely on large-scale training datasets and explicit reasoning annotations to achieve competitive performance (Tab. 2). In contrast, NoRD attains a RFS of 7.709, ranking as the third best-performing VLA on the benchmark, while being the only top model trained without reasoning traces or ensembling. Remarkably, NoRD achieves this with merely 12,000 samples for supervised training and 450 samples for RLFT,

Table 3. Test results on NAVSIM benchmark (navtest subset). NoRD achieves competitive performance (*w/o R*: Without reasoning data, *w/o L*: without LiDAR data, and *C*: Number of RGB frames; * *BoN* refers to the average over best score per sample out of 6 outputs with different random seeds).

Method	w/o R	w/o L	C	PDMS↑	Collision↑	DAC↑	Direction↑	Progress↑	TTC↑	Comfort↑
BEV-based Methods										
UniAD [17]	✓	✓	32	83.4	97.7	91.9	-	78.8	92.9	**100**
Transfuser [7]	✓	✗	3	84.0	97.7	92.8	**97.9**	79.2	92.8	**100**
Hydra-MDP [24]	✓	✗	3	86.5	98.2	96.2	95.8	78.7	94.6	**100**
DiffusionDrive [25]	✓	✗	3	88.1	98.2	96.2	-	82.2	94.7	88.1
VLA-based Methods										
AutoVLA [51]	✗	✓	12	89.1	98.4	95.6	95.4	81.9	**98.0**	99.9
AutoVLA-BoN* [51]	✗	✓	12	92.1	99.1	97.1	95.5	**87.6**	97.1	**100**
RecogDrive [23]	✗	✓	12	89.6	98.2	97.9	-	83.5	95.2	99.8
NoRD	✓	✓	3	85.6	97.6	94.9	95.9	79.3	93.5	**100**
NoRD-BoN*	✓	✓	3	**92.4**	**99.2**	**98.3**	95.9	86.4	97.8	99.9



(a) NAVSIM (b) WaymoE2E

Figure 6. **Pareto-optimal curves on two driving benchmarks.** (a) NoRD is the only VLA in NAVSIM operating in the high-performance, high–data-efficiency region using only RGB inputs. (b) NoRD achieves competitive RFS on WaymoE2E with a fraction of the training data, without ensembling or reasoning supervision. Shaded regions provide a qualitative categorization of model efficiency and performance for ease of visualization.

whereas Poutine and HMVLM require 17× and 12× larger datasets for only marginal RFS gains. Furthermore, NoRD surpasses all other competitive models on the ADE metric, despite a $\geq 6\times$ reduction in training data, underscoring its strong generalization ability, as shown in Fig. 8.

NAVSIM performance: The NAVSIM benchmark rigorously evaluates trajectory prediction by executing models in a simulator and scoring them using the PDM metric, a weighted measure of high-level driving factors such as comfort, time-to-collision, and ego-progress. As shown in Tab. 3 (with qualitative results in Fig. 7), NoRD is the only model that requires no reasoning traces, relies solely on 3 camera frames, and uses no additional features like, LiDAR and HD Map. While other VLA models, such as AutoVLA and RecogDrive, require 1.6× and 34× more training data, NoRD achieves competitive performance with fewer than 90,000 samples. We also evaluate the best-of-N performance, where the oracle selects the best trajectory out of 6 predictions based on the PDM score. In this configuration, NoRD-BoN surpasses reasoning-based AutoVLA-BoN, achieving a PDM score of 92.4, highlighting its capabilities and data-efficiency.

Efficiency and Scalability: A central contribution of our work is the remarkable data efficiency of NoRD,

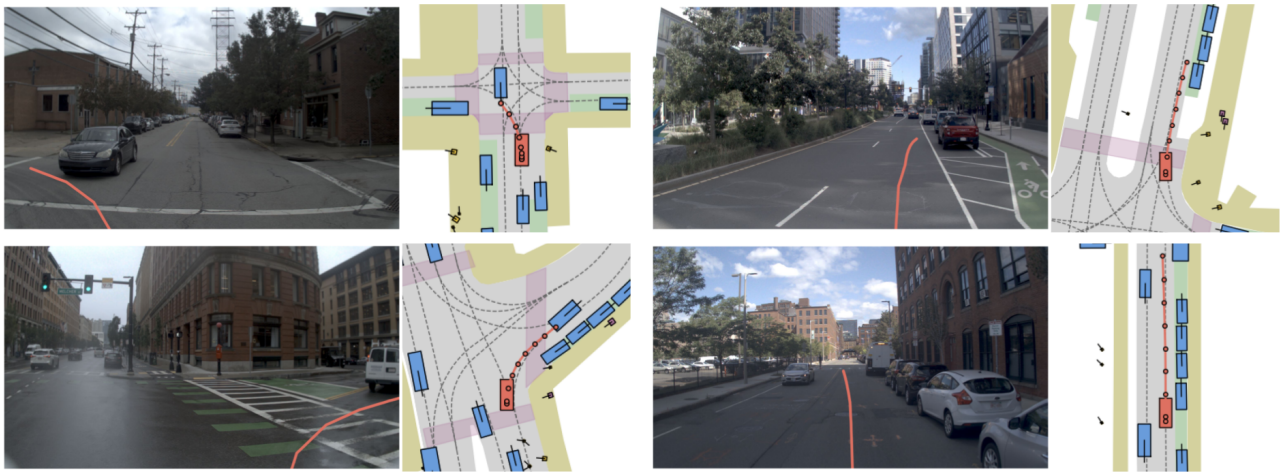


Figure 7. **Qualitative Results on NAVSIM (navtest subset).** NoRD safely executes sharp turns, respects traffic lights, and avoids collisions, demonstrating robust driving behavior. The predicted trajectory is shown in red.

Figure 8. **Qualitative Results on WaymoE2E test set.** NoRD drives safely in challenging out-of-distribution scenarios like unsafe pedestrian crossing and construction site. The predicted trajectory is shown in red (stitched, center-cropped for visualization).

as highlighted in the Pareto-front analyses (Fig. 6a and Fig. 6b). On both benchmarks, NoRD establishes a competitive performance baseline while operating in the high-efficiency (*i.e.*, low data) regime. While some VLA-based methods eventually achieve marginally higher absolute scores, they do so at a prohibitive data cost of at least 3× more data. NoRD, in contrast, firmly establishes itself on the efficiency frontier, presenting an optimal and practical trade-off. While maintaining data efficiency, since it directly predicts the trajectory tokens, it is extremely lightweight and this enables it to achieve significantly lower inference time and token count, as compared to other

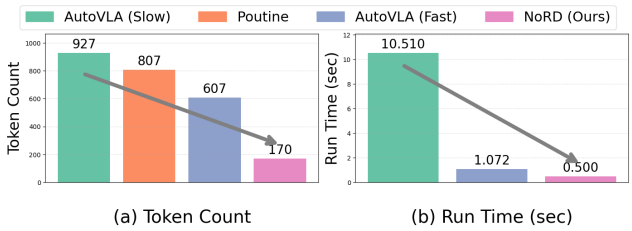


Figure 9. **Comparison of token and runtime efficiency.** NoRD is the most (a) token and (b) runtime efficient VLA.

VLAs as shown in Fig. 9. Our findings strongly suggest that high-performance autonomous driving VLAs do not necessarily require large datasets, paving the way for more accessible and scalable yet efficient models.

6. Conclusion

We proposed NoRD, a reasoning-free, data-efficient VLA for autonomous driving. NoRD achieves strong performance while eliminating language reasoning and significantly reducing training data requirements. By analyzing rewards and modifying training pipelines, we demonstrate that VLAs can be trained with substantially fewer samples while improving token efficiency and inference speed. While Dr. GRPO mitigates difficulty bias better than GRPO, it remains imperfect [22], leaving room for future work. Importantly, NoRD does not suggest that VLAs cannot benefit from language-based reasoning; rather, it shows that efficient, high-performing VLAs can be trained without reasoning and large-scale datasets, pushing the boundaries of data and inference efficiency.

References

[1] Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. Polaris: A post-training recipe for scaling reinforcement learning on advanced reasoning models, 2025. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2

[4] Mingrui Chen, Haogeng Liu, Hao Liang, Huaibo Huang, Wentao Zhang, and Ran He. Unlocking the potential of difficulty prior in rl-based multimodal reasoning, 2025. 3

[5] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 26890–26900, 2025. 2

[6] Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. *arXiv preprint arXiv:2505.23757*, 2025. 2

[7] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022. 7

[8] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025. 3

[9] Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. Retrieval-based interleaved visual chain-of-thought in real-world driving scenarios, 2025. 2

[10] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025. 3

[11] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 6

[12] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025. 3

[13] Shichao Fan, Quantao Yang, Yajie Liu, Kun Wu, Zhengping Che, Qingjie Liu, and Min Wan. Diffusion trajectory-guided policy for long-horizon robot manipulation. *arXiv preprint arXiv:2502.10040*, 2025. 3

[14] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 2

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. 1

[16] John Hewitt. Initializing new word embeddings for pretrained language models, 2021. 5

[17] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. 7

[18] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: End-to-end multimodal model for autonomous driving. *Transactions on Machine Learning Research*, 2025. 1, 2

[19] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving, 2023. 2

[20] Anqing Jiang, Yu Gao, Zhigang Sun, Yiru Wang, Jijun Wang, Jinghao Chai, Qian Cao, Yuweng Heng, Hao Jiang, Yunda Dong, et al. Diffvla: Vision-language

guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*, 2025. 2

[21] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 6

[22] Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. DisCO: Reinforcing large reasoning models with discriminative constrained optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2, 3, 4, 8

[23] Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*, 2025. 2, 7

[24] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 7

[25] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12037–12047, 2025. 7

[26] Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Wang Ye, Zhengbing He, Chengzhong Xu, and Zhenning Li. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*, pages 1–15, 2025. 2

[27] Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, and Yang Liu. DAPO : Improving multi-step reasoning abilities of large language models with direct advantage-based policy optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[28] Xueyi Liu, Zuodong Zhong, Qichao Zhang, Yuxin Guo, Yupeng Zheng, Junli Wang, Dongbin Zhao, Yun-Fu Liu, Zhiguo Su, Yinfeng Gao, Qiao Lin, and Chen Huiyong. Reasonplan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. In *9th Annual Conference on Robot Learning*, 2025. 2

[29] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *COLM*, 2025. 2, 3, 4, 5

[30] Waymo LLC. Vision-based end-to-end driving - 2025: Waymo open dataset. https://waymo.com/open/challenges/2025/e2e-driving/, 2025. 2, 6

[31] Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, et al. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*, 2025. 2

[32] Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*, 2025. 3

[33] Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *The Twelfth International Conference on Learning Representations*. 5

[34] Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, et al. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *arXiv preprint arXiv:2505.15298*, 2025. 2

[35] Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11993–12003, 2025. 1, 2

[36] Luke Rowe, Rodrigue de Schaetzen, Roger Girgis, Christopher Pal, and Liam Paull. Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving. *arXiv preprint arXiv:2506.11234*, 2025. 1, 2, 6

[37] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024. 6

[38] Xurui Song, Shuo Huai, JingJing Jiang, Jiayi Kong, and Jun Luo. More than meets the eye? uncovering the reasoning-planning disconnect in training vision-language driving models. *arXiv preprint arXiv:2510.04532*, 2025. 1, 2

[39] Xiaolong Tang, Meina Kan, Shiguang Shan, and Xilin Chen. Plan-R1: Safe and feasible trajectory planning as language modeling. In *International Conference on Learning Representations (ICLR)*, 2026. 5

[40] Daming Wang, Yuhao Song, Zijian He, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. Hmvlm: Multistage reasoning-enhanced vision-language model for long-tailed driving scenarios. *arXiv preprint arXiv:2506.05883*, 2025. 6

[41] Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, et al. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025. 1, 2

[42] Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. DiffusionVLA: Scaling robot foundation models via unified diffusion and autoregression. In *Forty-second International Conference on Machine Learning*, 2025. 3

[43] Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Z Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, Dragomir Anguelov, and Mingxing Tan. S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[44] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025. 3

[45] Zhenlong Yuan, Jing Tang, Jinguo Luo, Rui Chen, Chengxuan Qian, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025. 1, 2

[46] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[47] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[48] Jixiao Zhang and Chunsheng Zuo. GRPO-LEAD: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5642–5665, Suzhou, China, 2025. Association for Computational Linguistics. 3

[49] Weicheng Zheng, Xiaofei Mao, Nanfei Ye, Pengxiang Li, Kun Zhan, Xianpeng Lang, and Hang Zhao. Driveagent-r1: Advancing vlm-based autonomous driving with active perception and hybrid thinking. *arXiv preprint arXiv:2507.20879*, 2025. 2

[50] Jingyu Zhou, Lu Ma, Hao Liang, Chengyu Shen, Bin Cui, and Wentao Zhang. Daro: Difficulty-aware reweighting policy optimization. *arXiv preprint arXiv:2510.09001*, 2025. 3

[51] Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1, 2, 3, 6, 7

NoRD: A Data-Efficient Vision-Language-Action Model that Drives without Reasoning

Supplementary Material

7. Comparison between GRPO and Dr. GRPO

We present a component-wise breakdown of Tab. 1 in Tab. 4. Except for Ego Progress, Dr. GRPO significantly outperforms GRPO. As shown in the training and validation curves in Fig. 11, both GRPO and Dr. GRPO improve over time; however, GRPO consistently lags behind Dr. GRPO. To further illustrate this, we visualize the change in mean PDM scores of the group, relative to the SFT model (step 0), across different variance groups in Fig. 10. The variance groups are defined based on intra-group tertiles. Our analysis reveals that:

1. **Low-variance samples** (Fig. 10 (a)): GRPO exhibits higher density above the $y = x$ line, particularly for initial scores in $[0.8, 1.0]$.
2. **Medium- and high-variance samples** (Fig. 10 (b,c)): Dr. GRPO outperforms GRPO, with a denser concentration above the $y = x$ line. The performance gap widens for high-variance samples, consistent with our observation that GRPO attenuates policy updates for such samples.

8. Detailed Results

8.1. Prompt Example

We show an illustrative example in Fig. 12. NoRD maintains token and inference efficiency by directly predicting the trajectory tokens.

8.2. Waymo E2E Scores

We present the detailed results of the performance of NoRD on WaymoE2E test set in Tab. 6. As is evident, NoRD is capable of performing complex multi-lane switching maneuvers, while also performing well in less-represented scenes, such as intersections and construction sites.

8.3. Effect of Vocabulary Size

We experimented with a smaller k-disc vocabulary, consisting of 512 trajectory tokens (as compared to 2048 trajectory tokens in NoRD) and found that the performance on NAVSIM degrades (Tab. 5). This is perhaps because the smaller vocabulary size cannot represent complex maneuvers like sharp turn faithfully.

9. Reward Functions

In this section, we elaborate on the reward functions used for RL post-training. The reward consists of length reward, format reward and dataset-specific reward (PDM score for NAVSIM and Normalized RFS for WaymoE2E). The output of the model is a string of action tokens like `TRAJ_0242 TRAJ_150 TRAJ_172` that are decoded to a list of waypoints of tuples `[x,y,yaw]` at 10 Hz.

Format Reward (r_f)**:** A binary reward taking values in $\{0,\ 0.25\}$. A reward of 0.25 is assigned if the prediction consists of valid space-separated trajectory tokens of the form `TRAJ_i`, where i is a zero-padded 4-digit integer in $[0, 2047]$; otherwise the reward is 0.

Length Reward (r_l)**:** A binary reward taking values in $\{0,\ 0.25\}$. The model receives a reward of 0.25 if the prediction contains the correct number of trajectory tokens (8 for NAVSIM and 10 for WaymoE2E); otherwise, the reward is 0.

Dataset Specific Reward (r_d):

1. **PDM Score for NAVSIM**: The PDM score (range: $[0, 1]$) comprehensively measures the driving quality and safety. Is it given by:

$$\text{PDM Score} = \text{NC} \times \text{DAC} \times \frac{5 \cdot \text{TTC} + 2 \cdot \text{C} + 5 \cdot \text{EP}}{12}$$

   where, No at-fault Collision (NC), Drivable Area Compliance (DAC), Ego Progress (EP), Comfort (C), and Time-to-Collision (TTC) are all within $[0, 1]$.
2. **Normalized RFS for WaymoE2E**: The RFS quantifies the alignment of the model's predicted trajectory $\hat{T}$ with a set of three pre-rated human trajectories T_r. A score $s_r \in [3, 10]$ is assigned to each rater trajectory based on whether $\hat{T}$ falls within a *trust region* defined by dynamic longitudinal $\bar{\tau}_{\text{lng}}$ and lateral $\bar{\tau}_{\text{lat}}$ thresholds (scaled by current velocity). The final score is $\max_r(s_r)$, averaged over $t \in \{3, 5\}$ seconds, and clipped to $\min(\cdot, 4)$. The Normlized RFS, with range $[0, 1]$ is then given by:

$$\text{Normalized RFS} = \frac{\max(\max_r(s_r), 4) - 4}{6}$$

The overall reward r for the predicted trajectory is therefore given as:

$$r = \frac{r_f + r_l + r_d}{1.5}$$

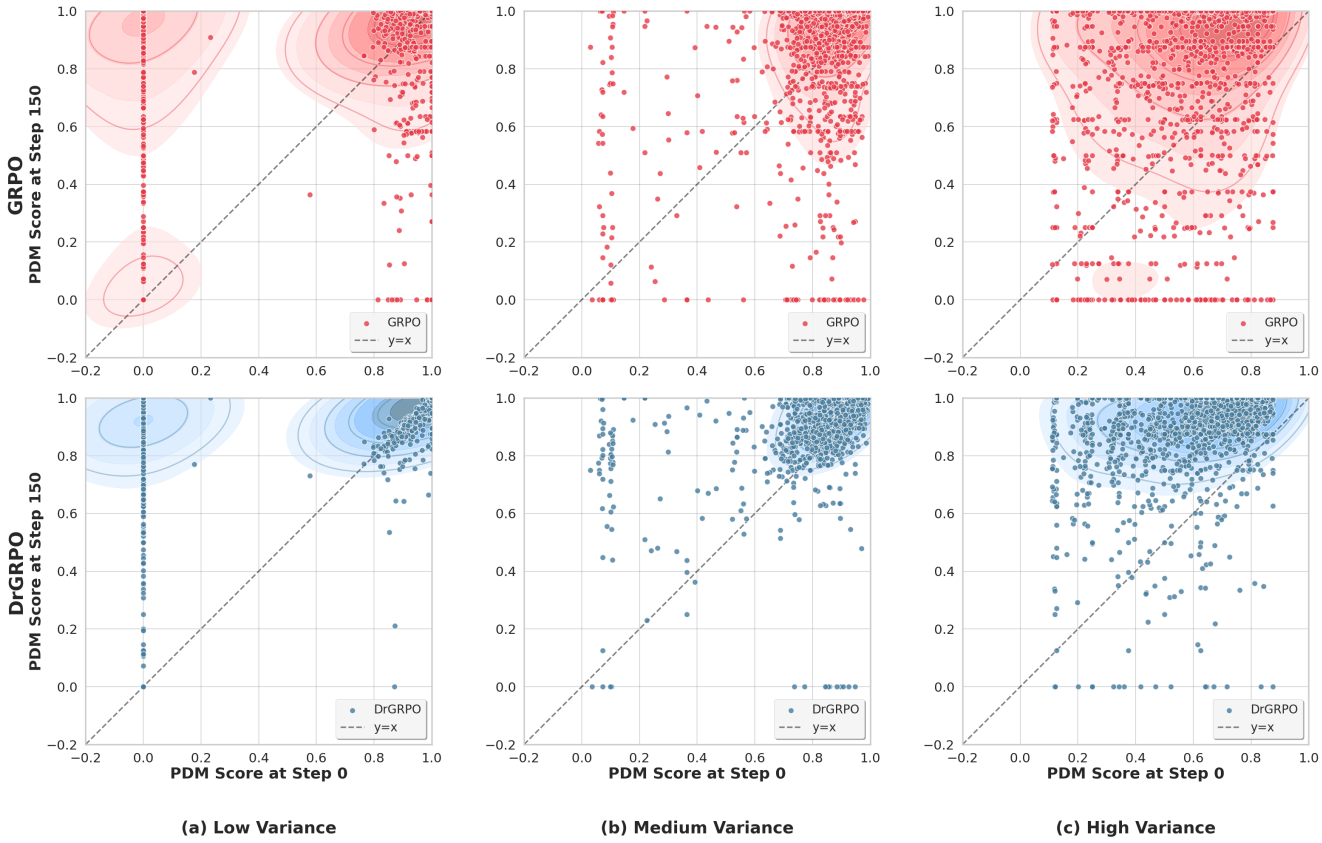


Figure 10. **Training improvement patterns for GRPO (top, red) and Dr. GRPO (bottom, blue) across intra-group variance levels.** The $y = x$ line indicates no change in PDM score. GRPO shows strong improvements for low-variance samples with initial scores in $[0.8, 1.0]$ (panel (a)), while Dr. GRPO outperforms GRPO for medium- and high-variance samples (panels (b) and (c)), with denser concentration above $y = x$.

Table 4. Detailed comparison of RL-fine-tuning of NORD-BASE with GRPO and Dr. GRPO. Dr. GRPO based RL fine-tuning is almost always better than GRPO.

Method	**PDMS↑**	**Collision↑**	**DAC↑**	**Direction↑**	**Progress↑**	**TTC↑**	**Comfort↑**
NORD-BASE	76.66	96.45	86.37	94.62	71.58	90.37	99.97
NORD-BASE+GRPO	77.18	91.89	90.12	91.84	**80.06**	80.13	99.96
NORD-BASE+Dr. GRPO	**85.62**	**97.56**	**94.92**	**95.94**	79.30	**93.53**	**100**

Table 5. Effect of k-disc vocabulary size on the performance of NORD on navtest.

Vocabulary Size	**PDMS ↑**
512	83.07
2048	85.62

10. Dataset Details

10.1. WaymoE2E

Supervised Finetuning: We curated the SFT dataset from the official WaymoE2E training set. Frames were first strictly filtered, retaining only those that guaranteed four preceding time steps were available for consistent extraction of the ego-vehicle's historical states. The final subset was then created by uniformly sampling 20% of these valid frame sequences from all contexts. This dataset was then randomly split into training and validation sets using an $85/15$ ratio. The input images were resized to

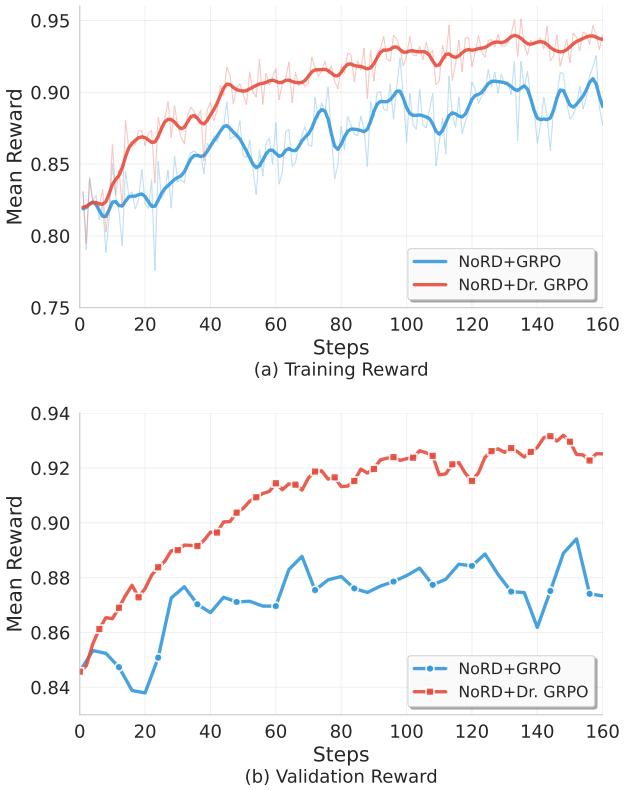


Figure 11. **Training and validation curves for RL fine-tuning with GRPO and Dr.GRPO.** Dr.GRPO (in red) consistently outperforms GRPO (in blue) on the (a) training and (b) validation sets by a significant margin.

Table 6. Detailed results on WaymoE2E Test Set.

Metric Name	**Value↑**
Construction Score	8.072616
Intersection Score	7.9252014
Pedestrian Score	7.7775736
Cyclist Score	7.8055406
Multi Lane Maneuver Score	7.8262477
Single Lane Maneuver Score	8.308635
Cut In Score	7.734755
Foreign Object Debris Score	7.6988134
Special Vehicle Score	7.7961473
Spotlight Score	6.5309787
Others Score	7.322814
ADE at 3 seconds	1.250462
ADE at 5 seconds	2.8928785
Average Score	7.709029

ensure the total number of pixels lies between 784 and 401,408, following the Qwen vision encoder's constraints.

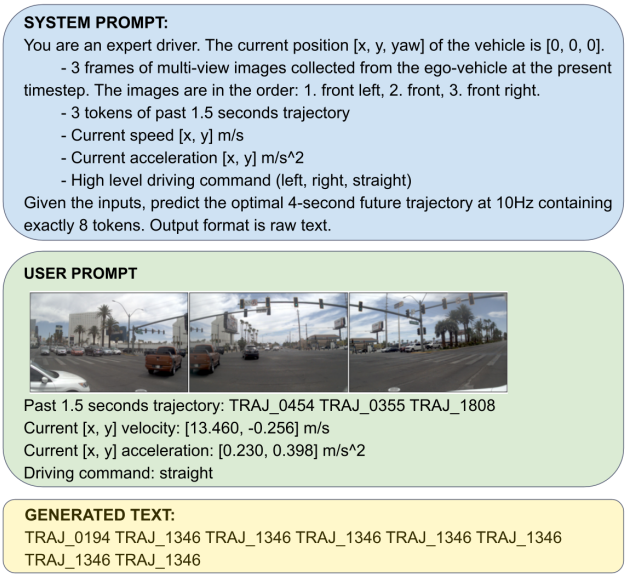


Figure 12. **Example of NoRD inference.** Given multi-view images, past trajectory, and the current velocity, acceleration, and driving command, NoRD directly predicts the trajectory tokens without explicit reasoning.

RL Finetuning: We use the official WaymoE2E validation set, for which preference annotations are provided for a single frame per scenario. Consequently, we extract one sample per scenario and randomly split the resulting set into training and validation sets using an $85/15$ ratio.

10.2. NAVSIM

Supervised Finetuning: We use the official NAVSIM's training set (`navtrain`) and split it into training and validation sets for SFT using an $80/20$ ratio. The input images were resized to ensure the total number of pixels lies between 784 and 401,408, following the Qwen vision encoder's constraints.

RL Finetuning: We construct a RLFT dataset from the NAVSIM validation split originally used for supervised fine-tuning. To remove trivial driving behaviors, we filter trajectories using a constant-velocity baseline and discard samples with a final-point displacement error below 0.2 m. For turning maneuvers, we additionally enforce a minimum average heading change of 0.01 rad per timestep to eliminate mild curvature and drift. Straight trajectories are exempt from the heading filter and are filtered solely using the displacement criterion. After filtering, the remaining samples are balanced across three driving intents—straight, left, and right—by uniformly subsampling each class. The resulting dataset contains only non-trivial and dynamically diverse trajectories, providing a more rigorous training signal for reinforcement learning-based trajectory prediction

and decision-making models.

11. Implementation Details

11.1. Supervised Finetuning

We perform supervised fine-tuning of NoRD on the NAVSIM and WaymoE2E datasets using the Qwen2.5-VL-3B-Instruct backbone, adapted to predict discretized trajectory tokens from multi-view images, past trajectories, and the ego-vehicle's current kinematic states. For NAVSIM, inputs consist of three camera frames (Front-Left, Front, Front-Right), three past trajectory tokens covering the previous 1.5 seconds, current velocity and acceleration, and a high-level driving command, with the model predicting 8 future trajectory tokens over a 4-second horizon at 10Hz. For WaymoE2E, inputs include six past trajectory tokens spanning 3 seconds, and the model predicts 10 future tokens over a 5-second horizon. In both cases, trajectory tokens are incorporated into the model vocabulary. All components of the model, including the vision encoder, multimodal MLP, and language model, are fine-tuned using mixed-precision training with bf16 and gradient checkpointing to reduce memory footprint. We train the model across 16 A100 GPUs, applying DeepSpeed ZeRO Stage 3 optimization for WaymoE2E and standard distributed training for NAVSIM. We use consistent hyperparameters across datasets, including a learning rate of 5×10^{-5}, a batch size of 8 per device with 4 gradient accumulation steps, a cosine learning rate scheduler, a warmup ratio of 0.03, and gradient clipping at 1. We evaluate the model every 50 steps on the validation sets and select the best model based on minimum evaluation loss.

11.2. RL Finetuning

We perform RL fine-tuning of NoRD using Dr. GRPO to optimize task-specific rewards. We generate 8 rollouts per input to estimate group-relative advantages and update the policy accordingly. We use a batch size of 128 trajectories for NAVSIM and 256 trajectories for WaymoE2E, applying asymmetric clipping with a high clip of 0.1 and a low clip of -0.2 to stabilize policy updates. We train across 32 A100 GPUs for WaymoE2E and 30 A100 GPUs for NAVSIM, leveraging mixed-precision and gradient checkpointing for memory efficiency. We periodically evaluate the policy on validation sets and retain the checkpoint achieving the highest reward.

12. Dataset Scale Estimation

To visualize the performance-efficiency frontier in Fig. 6, we estimated the total number of training samples for all evaluated models based on their reported configurations. Across both NAVSIM and WaymoE2E, baseline methods frequently employ complex multi-dataset mixtures or utilize varying fractions of the available data. To standardize these counts, we explicitly aggregated the reported dataset percentages and official splits detailed in the respective papers' training sections. For example, on WaymoE2E, we calculate HMVLM and DiffusionLTF at approx. 500k and 730k samples based on the train and val splits in the Waymo Open Dataset for end-to-end driving and perception. Similarly, for Poutine and AutoVLA, we aggregate their reported multi-dataset percentages to approx. 700k and 210k samples, respectively. These standardizations ensure a fair relative comparison of data efficiency on the x-axis.

Figure 13. **Failure cases of NoRD.** The predicted trajectory is shown in red and the violations marked in red circle.

13. Failure Cases

While NoRD achieves strong performance, it remains susceptible to failure in certain scenarios. We present representative examples in Fig. 13. These cases can be attributed, in part, to the fact that Dr. GRPO remains susceptible to difficulty bias, which still affects the policy optimization dynamics. We therefore believe that targeted interventions to better account for task difficulty could further push the performance frontier.