VADE: Variance-Aware Dynamic Sampling via Online Sample-Level Difficulty Estimation for Multimodal Reinforcement Learning

Supplementary Material

6. Dataset Details

We now detail the construction of our training dataset. We select 9 tasks from LLaVA-OneVision-Data[8] and incorporate the training sets of ChartQA[12] and ScienceQA[10], ensuring no overlap between the training data and the evaluation benchmarks. The constructed dataset spans three distinct domains: mathematics, chart understanding, and scientific reasoning, comprising a total of 9,471 samples. The detailed data composition is presented in Table 3. All entries except ChartQA and ScienceQA correspond to subset names from LLaVA-OneVision-Data.

7. Implementation Details

Table 4 specifies the hyperparameter configuration used in our experiments. All training procedures employed identical parameter settings. All models were trained on NVIDIA H200 140GB GPUs.

Table 4. Hyperparameters for training Qwen2.5VL-7B/3B-Instruct models.

Hyperparameter	Value
Train batch size	512
Max prompt length	8192
Max response length	2048
Filter overlong prompts	True
Rollouts per prompt	8
Total epoches	15
Learning rate	1e-6
ppo_mini_batch_size	128
ppo_micro_batch_size_per_gpu	16
kl_loss_coef	0.01
tensor_model_parallel_size	2
Rollout engine	vLLM
GPU	NVIDIA H200 140G
Train machines numbers	1
GPU numbers per machine	4

8. More Experimental Results

8.1. More Training Dynamics

This section presents additional training dynamics results.

Figure 7 shows the validation score, effective gradient ratio, and actor gradient norm curves comparing our method with vanilla method on Qwen2.5VL-7B-Instruct trained with GSPO. Similarly, Figure 8 presents corresponding results on Qwen2.5VL-3B-Instruct with GRPO, while Figure 9 shows results on Qwen2.5VL-3B-Instruct with GSPO. These results demonstrate that VADE consistently outperforms vanilla baselines across different model scales and algorithms, confirming its effectiveness and general applicability.

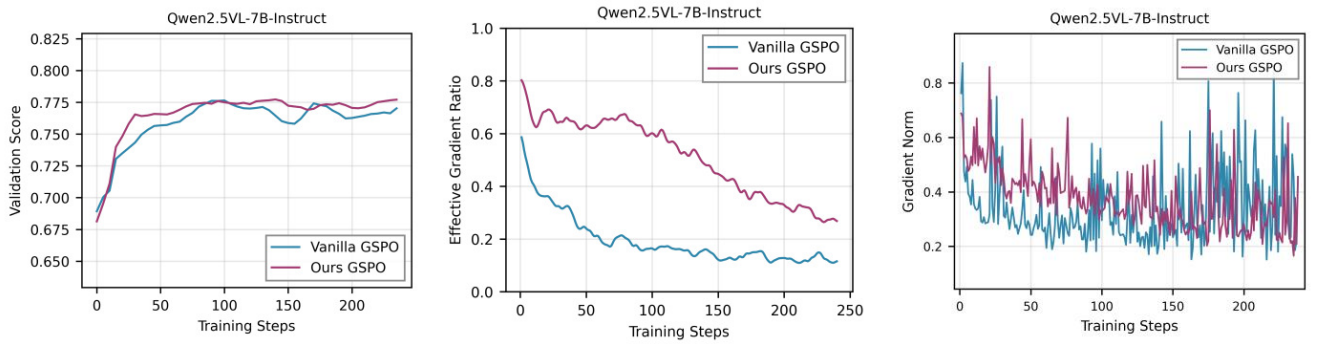
Furthermore, Figure 10 extends the computational efficiency comparison between VADE and DAPO across multiple configurations: Qwen2.5VL-7B-Instruct with GRPO, Qwen2.5VL-7B-Instruct with GSPO, and Qwen2.5VL-3B-Instruct with GSPO. Our method consistently achieves superior performance with significantly fewer rollout generations under all these settings, highlighting its substantial advantage in training efficiency and general applicability.

8.2. Detailed Ablation Study Results

This section provides complete ablation results across all benchmarks in Table 5. Following the same format as Table 1, we report average@4 scores for each benchmark. Our full configuration consistently outperforms all ablation variants, demonstrating the effectiveness of each component in VADE's design.

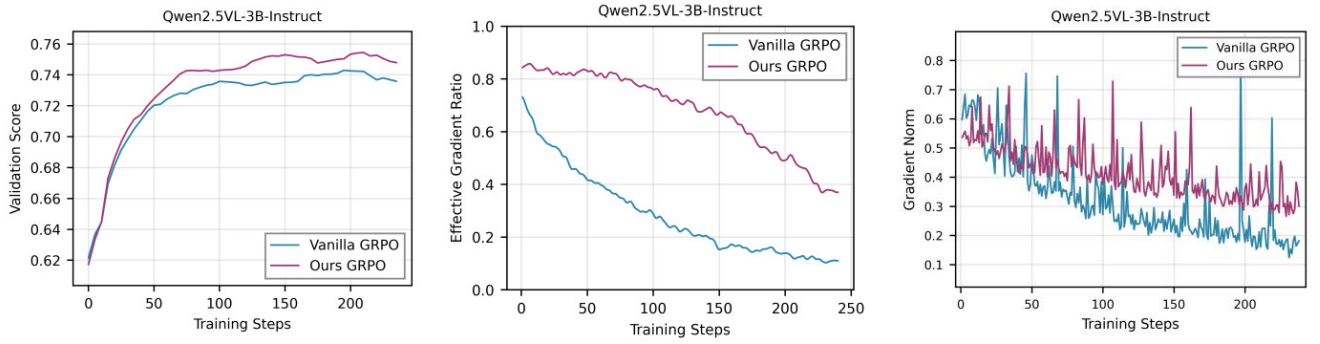
Table 3. Composition of the constructed training dataset. All entries except ChartQA and ScienceQA correspond to subset from LLaVA-OneVision-Data. The dataset spans mathematical, chart-based, and scientific reasoning tasks.

Domain	#Total Samples	Tasks	#Samples Per Task
Math	5670	geo3k	800
		geo170k(qa)	1,778
		CLEVR-Math(MathV360K)	321
		GEOS(MathV360K)	498
		GeoQA+(MathV360K)	1,923
		Geometry3K(MathV360K)	350
Chart	2476	figureqa(cauldron,llava_format)	957
		ChartQA[12]	1,300
		tabmwp(cauldron)	219
Science	1325	ScienceQA[10]	1,100
		ai2d(cauldron,llava_format)	225



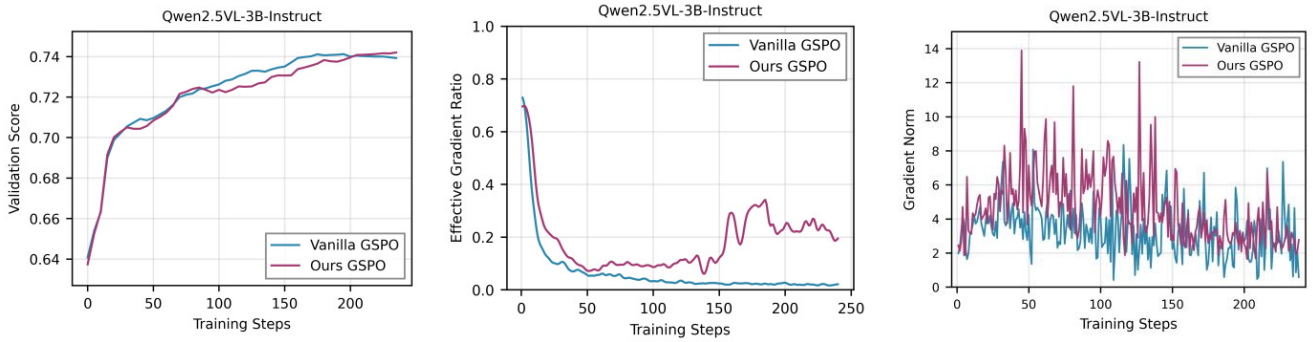
(a) Validation score (b) Effective gradient ratio (c) Gradient norm

Figure 7. Training dynamics of Qwen2.5VL-7B-Instruct trained with the GSPO algorithm. (a) Validation score, illustrating convergence speed and final performance; (b) Effective gradient ratio, representing the proportion of data with non-uniform rewards (neither all-zero nor all-one) in each training batch, reflecting data efficiency; (c) Actor gradient norm, indicating the magnitude of gradient signals throughout training.



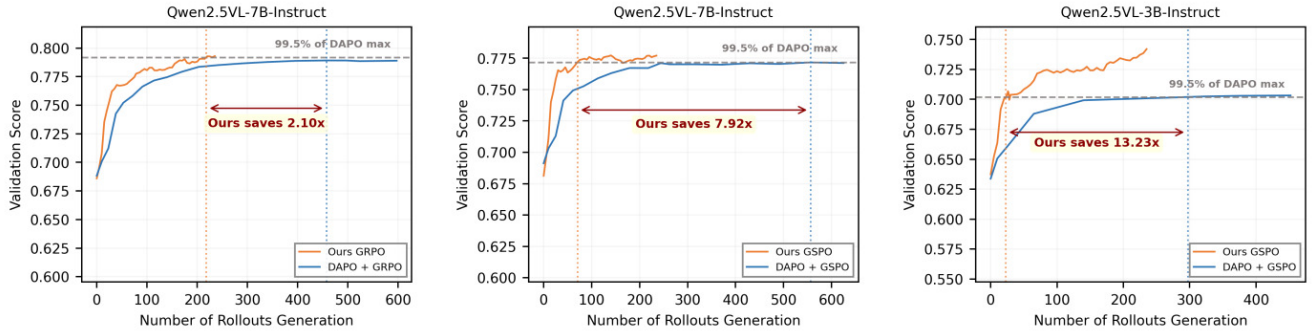
(a) Validation score (b) Effective gradient ratio (c) Gradient norm

Figure 8. Training dynamics of Qwen2.5VL-3B-Instruct trained with the GRPO algorithm. (a) Validation score, illustrating convergence speed and final performance; (b) Effective gradient ratio, representing the proportion of data with non-uniform rewards (neither all-zero nor all-one) in each training batch, reflecting data efficiency; (c) Actor gradient norm, indicating the magnitude of gradient signals throughout training.



(a) Validation score (b) Effective gradient ratio (c) Gradient norm

Figure 9. Training dynamics of Qwen2.5VL-3B-Instruct trained with the GSPO algorithm. (a) Validation score, illustrating convergence speed and final performance; (b) Effective gradient ratio, representing the proportion of data with non-uniform rewards (neither all-zero nor all-one) in each training batch, reflecting data efficiency; (c) Actor gradient norm, indicating the magnitude of gradient signals throughout training.



(a) Qwen2.5VL-7B with GRPO (b) Qwen2.5VL-7B with GSPO (c) Qwen2.5VL-3B with GSPO

Figure 10. Extended comparison between VADE and DAPO across multiple configurations: (a) Qwen2.5VL-7B-Instruct with GRPO; (b) Qwen2.5VL-7B-Instruct with GSPO; (c) Qwen2.5VL-3B-Instruct with GSPO. VADE achieves competitive performance with significantly fewer rollout generations, demonstrating consistent efficiency gains across model scales and algorithms.

Table 5. Detailed ablation study results on Qwen2.5VL-7B-Instruct with GRPO, reporting average@4 performance across all evaluation benchmarks. Highlighted cells indicate whether each variant improves or degrades performance compared to our full VADE design.

Method	MathVista	MathVerse	MathVision	ScienceQA	ChartQA	AVG
Ours	75.10	45.65	25.79	91.67	85.56	64.75
$p(1-p)$	73.07	45.57	24.37	90.38	84.92	63.66
greedy	73.15	44.72	24.74	90.32	84.30	63.25
last update	71.33	45.11	22.25	91.22	85.60	63.10