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

Zengjie Hu[1,2*], Jiantao Qiu[2‡], Tianyi Bai[3*], Haojin Yang[1],
Binhang Yuan[3], Qi Jing[1†], Conghui He[2†], Wentao Zhang[1†]

[1]Peking University, [2]Shanghai AI Lab, [3]HKUST

* Equal contribution. ‡ Project leader. † Corresponding authors.
{wentao.zhang, jingqi}@pku.edu.cn, {qiujiantao, heconghui}@pjlab.org.cn

November 25, 2025

Abstract

Group-based policy optimization methods like GRPO and GSPO have become standard for training multimodal models, leveraging group-wise rollouts and relative advantage estimation. However, they suffer from a critical *gradient vanishing* problem when all responses within a group receive identical rewards, causing advantage estimates to collapse and training signals to diminish. Existing attempts to mitigate this issue fall into two paradigms: filtering-based and sampling-based methods. Filtering-based methods first generate rollouts broadly and then retroactively filter out uninformative groups, leading to substantial computational overhead. Sampling-based methods proactively select effective samples before rollout but rely on static criteria or prior dataset knowledge, lacking real-time adaptability. To address these issues, we propose **VADE**, a **V**ariance-**A**ware **D**ynamic sampling framework via online sample-level difficulty **E**stimation. Our framework integrates three key components: online sample-level difficulty estimation using Beta distributions, a Thompson sampler that maximizes information gain through the estimated correctness probability, and a two-scale prior decay mechanism that maintains robust estimation under policy evolution. This three components design enables VADE to dynamically select the most informative samples, thereby amplifying training signals while eliminating extra rollout costs. Extensive experiments on multimodal reasoning benchmarks show that VADE consistently outperforms strong baselines in both performance and sample efficiency, while achieving a dramatic reduction in computational overhead. More importantly, our framework can serves as a plug-and-play component to be seamlessly integrated into existing group-based RL algorithms. Code and models are available at https://VADE-RL.github.io.

1 Introduction

Recent advancements in Reinforcement Learning (RL) have markedly improved the performance of large language models (LLMs) and multimodal large language models (MLLMs) on complex tasks involving mathematical reasoning, code generation, and scientific problem-solving[1, 2, 3, 4] . Reinforcement Learning with Verifiable Rewards (RLVR)[5] plays a central role in these developments and group-based policy optimization algorithms (e.g. GRPO[6], GSPO[7]) have gained widespread adoption due to their efficiency and scalability. These methods generate multiple response rollouts for each training sample and compute relative advantages within the group for policy optimization, eliminating the need for additional value function approximation.

Despite their success, group-based RL methods suffer from a critical limitation known as the *gradient vanishing* problem[8]: when all responses within a group receive identical rewards—a common scenario when the model

either solves all problems correctly or fails entirely—the relative advantages collapse to zero. This effectively eliminates the training signal for the entire group, severely degrading sample efficiency. As illustrated in the left half of Figure 1, during standard GRPO training with random sampling, the proportion of data yielding effective gradients progressively decreases, with zero-gradient groups dominating the training batches. In later stages, only around 10% of samples provide useful learning signals, creating a fundamental bottleneck for efficient training.

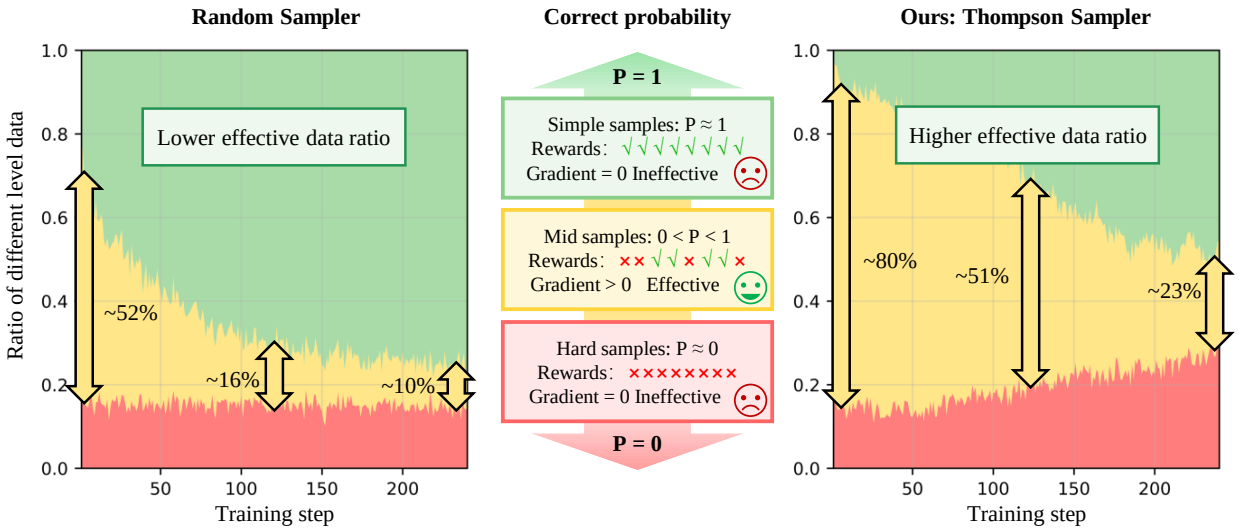


Figure 1: Proportion of data yielding effective training signals throughout training. **Left:** Under standard GRPO with random sampling, the proportion of informative samples providing non-zero gradients diminishes over time. **Right:** Our proposed Thompson sampling strategy consistently maintains a high ratio of effective data by proactively selecting informative samples.

Existing attempts to mitigate this issue can be categorized into two paradigms: filtering-based and sampling-based methods, both of which exhibit critical limitations. Filtering-based methods, such as DAPO[9] and VCRL[10], first generate rollouts for a broad set of samples and then retroactively filter out groups with uninformative reward signals. This "generate-then-filter" approach inevitably leads to substantial computational waste and requires costly mechanisms like repeated rollouts or replay buffers to assemble valid batches. In contrast, sampling-based methods, such as SEC[11] and MMR1[12], aim to proactively select samples of appropriate difficulty before rollout generation. However, they rely on static criteria or prior dataset knowledge: SEC depends on pre-defined, coarse-grained difficulty categories, while MMR1 requires periodic and expensive re-rollouts to refresh its sample-level estimates. Consequently, both strategies lack the fine-grained, real-time adaptability necessary to keep pace with an evolving policy, leaving an open need for a dynamic and efficient sample selection mechanism.

To address these limitations, we propose VADE, a Variance-Aware Dynamic sampling framework via online difficulty Estimation, designed to overcome the gradient vanishing problem in group-based RL. VADE introduces an efficient, sample-level data selection mechanism that operates without additional rollout inference overhead. We formulate data selection as a non-stationary Multi-Armed Bandit problem and solve it with a dynamic Thompson sampling strategy that continuously adapts to the policy's evolution. Our framework integrates three key components: an online sample-level difficulty estimation module that employs a Beta distribution as the posterior estimate of each sample's correctness probability; a Thompson sampler which leverages these estimates to maximize information gain and balance exploration with exploitation; and a two-scale prior decay mechanism that robustly maintains estimation accuracy under policy evolution. This integrated approach enables VADE to proactively identify the most informative samples, thereby maximizing the proportion of data yielding effective gradients, as

visualized in the right half of Figure 1. Notably, VADE requires no prior difficulty annotations, incurs no extra rollout cost, and serves as a plug-and-play component that can be seamlessly integrated into existing group-based RL algorithms such as GRPO and GSPO. Comprehensive experiments on multimodal reasoning benchmarks across various model scales and policy optimization algorithms show that VADE consistently achieves higher sample efficiency and better evaluation performance compared to strong baselines, demonstrating its robustness and generalization. Furthermore, ablation studies validate the effectiveness of each component in our framework.

Our main contributions are summarized as follows:

- We propose VADE, a novel and efficient framework designed to tackle the gradient vanishing problem in group-based RL. It enables dynamic selection of highly informative samples during training without introducing extra rollouts, thereby achieving higher performance and training efficiency.
- We establish a theoretical formulation by modeling data selection as a non-stationary Multi-Armed Bandit problem and then solve it using a dynamic Thompson Sampling strategy. This formulation requires no prior difficulty knowledge or annotated categories and can serve as a plug-and-play component for existing group-based RL algorithms.
- We conduct systematic experiments on multimodal reasoning tasks across various model scales and policy optimization algorithms. The experimental results show that VADE consistently achieves higher sample efficiency and better evaluation performance compared to strong baselines, demonstrating its robustness and generalization.

2 Related Work

The *gradient vanishing* problem presents a fundamental bottleneck for group-based policy optimization algorithms. Existing approaches to address this issue can be categorized into two main paradigms based on when and how they select informative data.

Filtering-based Methods. This line of work operates by generating rollouts first and filtering out non-informative samples afterward[9, 10, 13, 14]. DAPO [9] exemplifies this approach by discarding groups with zero reward variance and performing repeated rollouts to assemble valid batches. VCRL [10] extends this idea by using reward variance as a difficulty proxy, filtering out samples with variance below a pre-defined threshold and replenishing the batch from a replay buffer. While effective in ensuring gradient quality, these methods inherently follow a "generate-then-filter" paradigm that leads to substantial computational waste due to discarded rollouts. In contrast, our VADE framework avoids this overhead by proactively selecting samples before the rollout stage, eliminating the need for costly filtering and re-rollout mechanisms.

Sampling-based Methods. In contrast, this category aims to select high-utility samples *before* rollout generation[11, 12, 15]. Curriculum learning methods like SEC [11] formulate sample selection as a Multi-Armed Bandit problem over pre-defined difficulty categories. Meanwhile, MMR1 [12] employs a variance-aware sampling strategy that combines reward variance and trajectory diversity. However, these approaches rely on static criteria: SEC depends on coarse-grained, pre-defined difficulty hierarchies, while MMR1 requires periodic and expensive re-rollouts to refresh its estimates. Consequently, they lack the fine-grained, real-time adaptability needed to keep pace with an evolving policy. Unlike these methods, VADE performs fully online, sample-level difficulty estimation without relying on pre-defined categories or expensive re-rollouts, enabling real-time adaptation to policy evolution through its novel two-scale prior decay mechanism.

3 Method

In this section, we present the VADE framework, which tackles the *gradient vanishing* problem by selecting the most informative samples to maximize learning signals. An overview of the framework is illustrated in Figure 2. This section is organized as follows: In Section 3.1, we introduce the background of group-based policy optimization and formalize the sample selection objective from an information gain perspective. In

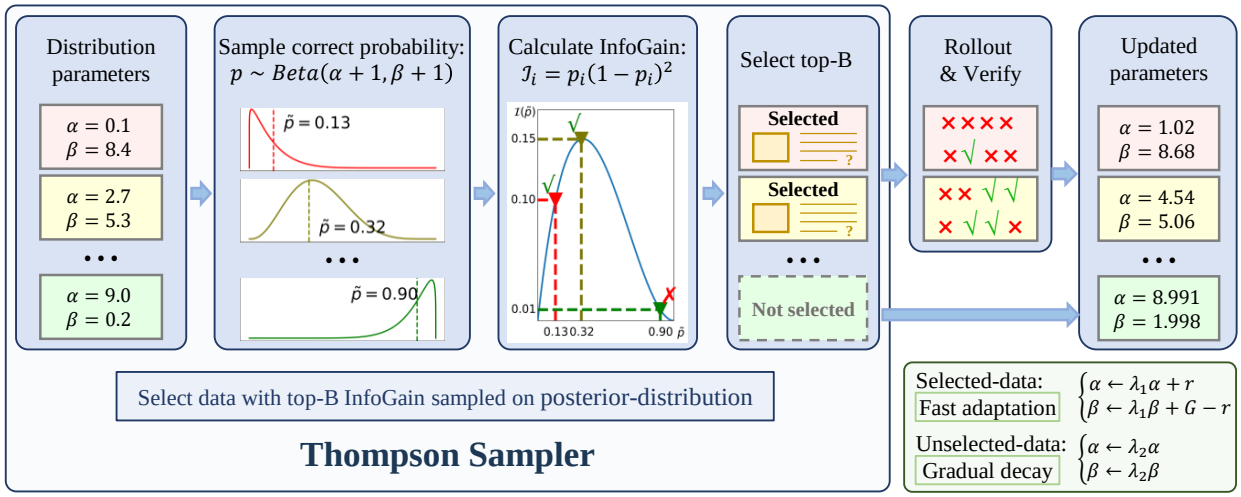


Figure 2: **Overview of the VADE framework.** Our method maintains distributions Beta(α_i+1,β_i+1) for each sample to enable online difficulty estimation. Through Thompson sampling and InfoGain $\mathcal{I}_i=p_t(1-p_t)^2$ maximization, VADE dynamically selects informative batches for group-wise rollouts. The two-scale prior decay mechanism ensures estimates remain accurate throughout policy evolution.

Section 3.2, we formulate the data selection process as a non-stationary Multi-Armed Bandit problem and details its solution via dynamic Thompson sampling strategy. Finally, in Section 3.3, we introduce the two-scale prior decay mechanism for maintaining robust online sample-level estimates under policy evolution. The complete RL training pipeline of our VADE framework is illustrated in Algorithm 1.

3.1 Group-wise Policy Optimization and Information Gain Theory

Current mainstream RLVR methods for training vision-language models follow this setup: each training data point x_i consists of an image and a related question, along with a reward function $R(x_i,y_i)\rightarrow\{0,1\}$ that verifies the answer and assigns a reward score. For a question x_i and its corresponding answer y_i, $R(x_i,y_i)=1$ indicates that the verification function considers the answer correct, while $R(x_i,y_i)=0$ indicates otherwise.

For a policy function $\pi_\theta(y|x)$ with parameters θ, the optimization objective on the training set $\mathcal{D}$ is:

$$J(\theta)=\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[R(x,y)] \tag{1}$$

The most popular group-wise policy optimization methods, such as GRPO and GSPO, employ group-wise rollouts to sample data and compute advantages based on the reward distribution within each rollout group:

$$A_{i,j}=\frac{R_{i,j}-mean(\{R_{i,j}\}_{j=1}^{G})}{std(\{R_{i,j}\}_{j=1}^{G})} \tag{2}$$

This approach eliminates the need for an additional value network, enabling efficient training with a reliable reward baseline.

A widely observed phenomenon is that the performance gain $\mathcal{I}_i$ from training on a specific data point x_i is related to its current difficulty[15]. According to the theory of sample information gain in group-wise policy optimization methods, as presented in *Knapsack RL*[15], for a problem x_i, let $p_i(\theta)$ denote the probability of correct reasoning under the current policy model π_θ:

$$p_i(\theta)=P_{y\sim\pi_\theta(\cdot|x_i)}(R(x_i,y)=1) \tag{3}$$

The information gain $\mathcal{I}_i(\theta)$ from selecting this data point for policy gradient optimization is approximately:

$$\mathcal{I}_i(\theta) \approx p_i(\theta)(1-p_i(\theta))^2 \tag{4}$$

Under the constraints of a fixed batch size B per training step and a rollout budget G per x_i, selecting the top B data points with the highest $\mathcal{I}_i(\theta)$ theoretically maximizes the single-step training gain.

3.2 Non-stationary Multi-Armed Bandit with Thompson Sampling

However, in practice, directly computing $p_i(\theta)$ precisely is infeasible due to the complexity of neural networks. To address this, we characterize its distribution $f(p_i(\theta))$ using rollout results and reward verification during training. Given that group-wise policy optimization inherently relies on rollout results from the current training batch, we naturally model the online data sampling and training process as a *Non-stationary Multi-Armed Bandit*:

1. At training step t, estimate the distribution of correctness probabilities $f(p_i(\theta_t)|h_{i,1:t})$ for all data points based on historical rollout records $h_{i,1:t}$, where $h_{i,1:t}$ represents the rollout history of x_i.
2. Compute $\mathcal{I}_i(\theta_t)$ using $f(p_i(\theta_t)|h_{i,1:t})$ and sample a batch of training data $\mathcal{B}_t = [x_{1,t}, x_{2,t}, \cdots, x_{B,t}]$. Perform rollouts on these data points and update the model parameters.
3. Update the rollout results for the current training step into $h_{i,1:t+1}$.

For a data point x_i at training step t, two scenarios are possible: a) It is sampled and generates G trajectories, with $a_{i,t}$ correct responses; b) It is not sampled, denoted as $\emptyset$. Thus, the historical record $h_{i,1:t}$ can be represented as a sequence containing $a_{i,t}$ and $\emptyset$, e.g., $[a_{i,1}, \emptyset, a_{i,3}, \cdots]$.

Once the posterior distribution $f(p_i(\theta_t)|h_{i,1:t})$ is estimated, sampling data according to the Thompson Sampling[16] method theoretically achieves high sample efficiency. Specifically, we sample a candidate probability $\tilde{p}_i(\theta_t)$ from the posterior distribution $f(p_i(\theta_t)|h_{i,1:t})$ and compute the corresponding information gain $\tilde{\mathcal{I}}_i(\theta_t)$. We then select the top B data points with the highest $\tilde{\mathcal{I}}_i(\theta_t)$ to form the current training batch. This strategy naturally balances exploration and exploitation: while the information gain objective $\tilde{\mathcal{I}}_i(\theta_t)$ favors samples with high expected utility, the random sampling from the posterior distribution $f(p_i(\theta_t)|h_{i,1:t})$ gives those uncertain yet potentially valuable samples a chance of being selected.

3.3 Online Estimation with Two-Scale Prior Decay

To implement the Thompson sampling strategy described above, we need a probabilistic model for the correctness probability p_i. A natural choice is the Beta distribution[17], which is a conjugate prior for the Bernoulli process. We maintain parameters (α_i, β_i) for each data point x_i, where α_i and β_i represent the cumulative counts of correct and incorrect responses for x_i, respectively, and we model the correctness probability with distribution Beta(α_i+1, β_i+1) to ensure numerical stability. After rolling out G responses and observing r_i correct and $G-r_i$ incorrect answers, the posterior distribution becomes Beta$(\alpha_i+1+r_i, \beta_i+1+G-r_i)$ under a stationary environment.

However, the model's continuous evolution during training makes the correctness probability non-stationary. If we simply updated the Beta parameters by adding new outcomes (e.g. $\alpha_i = \alpha_i + r_i$), the historical data would eventually overwhelm the new evidence, leading to outdated estimates. This is particularly problematic in two scenarios: (1) When a sample is used for training, its response distribution shifts significantly, diminishing the value of past rollouts; (2) Conversely, even unused samples are affected by generalization from trained data, gradually making their historical estimates less relevant.

Based on this analysis, we propose a ***two-scale prior decay model*** with two decay parameters $0 < \lambda_1 \ll \lambda_2 < 1$ to account for the two scenarios:

- If x_i is sampled:

$$\begin{cases} \alpha_{i,t} \leftarrow \lambda_1 \alpha_{i,t-1} + r_t \\ \beta_{i,t} \leftarrow \lambda_1 \beta_{i,t-1} + G - r_t \end{cases} \tag{5}$$

- If x_i is not sampled:

$$\begin{cases} \alpha_{i,t} \leftarrow \lambda_2 \alpha_{i,t-1} \\ \beta_{i,t} \leftarrow \lambda_2 \beta_{i,t-1} \end{cases} \tag{6}$$

This design accounts for: (1) significant distribution shift when data is directly used for training (λ_1), and (2) gradual generalization effects when data is not selected (λ_2).

Algorithm 1 provides the complete training pipeline for our VADE framework.

Algorithm 1 Reinforcement Learning with VADE.

Require: Dataset $\mathcal{D}$, batch size B, rollouts per sample G, decay rates λ_1,λ_2, training steps T, policy model π_θ

for each sample $x_i \in \mathcal{D}$ **do**
  Generate G rollouts $\{y_i\}_{i=1}^{G}$ from $\pi_\theta(\cdot|x)$
  Initial α_i,β_i according to the rollouts results.
end for
for training step t=1, ..., T **do**
  for each sample $x_i \in \mathcal{D}$ **do**
    Compute p_i using Thompson sampling.
    Compute $\mathcal{I}_i(\theta)$ according equation (4)
  end for
  Select the top-B $\mathcal{I}_i(\theta)$ to form the training batch $\mathcal{B}$.
  Apply RL update using the Training Batch $\mathcal{B}$
  for each sample $x_i \in \mathcal{D}$ **do**
    Update α_i,β_i according to equation (5) and equation (6).
  end for
end for

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness and efficiency of our proposed VADE framework across a range of visual reasoning tasks. This section is organized as follows: In Section 4.1, we describe our experimental setup, including model configurations, training details, and evaluation benchmarks. Section 4.2 presents the main results comparing our method against strong baselines across multiple visual reasoning benchmarks. Section 4.3 presents the training dynamics to demonstrate the efficiency of our approach during the training process. Finally, in Section 4.4, we provide ablation studies to validate the design choices of our method.

4.1 Experimental Setup

Implementation Details: We conduct reinforcement learning training on both Qwen2.5-VL-7B-Instruct and Qwen2.5-VL-3B-Instruct models [18], implementing all experiments using the verl [19] RL training framework. We implement our proposed VADE method using both GRPO [6] and GSPO [7] algorithms, with comparative analysis against vanilla GRPO/GSPO and DAPO [9] as baselines. It should be noted that, to ensure a fair comparison, we only utilize the dynamic sampling component of DAPO without incorporating its other configurations. Therefore, we denote this implementation as DAPO + GRPO/GSPO. The experimental configuration is standardized across all methods: training batch size is set to 512 and we sample 8 rollouts per prompt. The epoch of training is 15. In the implementation of our VADE method, the update parameters are configured as $\lambda_1 = 0.2$ and $\lambda_2 = 0.999$. For vanilla GRPO/GSPO and our proposed method, we complete all scheduled training steps. For DAPO, due to its need to over-sample from the training data, we limit the maximum number of rollout generation to three times the total training steps to ensure fair comparison. For further implementation details and experimental results, please refer to Appendix.

Table 1: Main performance comparison of VADE against other RL baselines on Qwen2.5-VL models. Highlighted cells show whether using the method improves or degrades the performance compared to the vanilla method.

Algo.	Method	Average@4					AVG
		MathVista	MathVerse	MathVision	ScienceQA	ChartQA	
Qwen2.5-VL-7B-Instruct							
Base Model		69.30	40.36	18.55	84.53	80.40	58.63
GRPO	vanilla	72.85	41.36	20.71	90.08	84.12	61.82
	DAPO	71.80	**47.11**	24.16	89.54	**85.72**	63.67
	Ours	**75.10**	45.65	**25.79**	91.67	85.56	**64.75**
GSPO	vanilla	71.88	41.09	23.29	**91.92**	84.60	62.56
	DAPO	73.35	**47.01**	24.13	91.32	86.04	64.37
	Ours	**73.45**	45.70	**25.68**	91.77	**86.24**	**64.57**
Qwen2.5-VL-3B-Instruct							
Base Model		63.15	23.27	11.29	76.70	78.84	50.65
GRPO	vanilla	63.37	33.40	19.16	85.08	76.80	55.56
	DAPO	**66.88**	**38.03**	**21.82**	83.64	**82.20**	**58.51**
	Ours	63.62	37.22	18.92	**85.37**	81.48	57.32
GSPO	vanilla	62.80	**33.47**	17.55	84.09	**82.12**	56.01
	DAPO	**64.70**	32.08	16.00	81.36	81.36	55.10
	Ours	63.82	32.89	**19.20**	**84.58**	81.40	**56.38**

Dataset: For the training dataset, we select 9 tasks from LLaVA-OneVision-Data[20] and incorporate the training sets of ChartQA[21] and ScienceQA[22], 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. During training, we split the dataset into 90% for training and 10% for validation. Further details regarding the dataset are provided in Appendix. To specifically demonstrate the effectiveness of our proposed method, all models and approaches in our experiments are exclusively trained using this dataset through Reinforcement Learning.

Evaluation Benchmarks: We conduct comprehensive evaluation across five challenging benchmarks: MathVista [23], MathVerse [24] and MathVision [25] for mathematical reasoning, ScienceQA [22] for scientific question answering, and ChartQA [21] for chart comprehension. For the MathVista, MathVerse and MathVision benchmarks, we employ the official evaluation code provided in their original papers, which utilizes the `LLM-as-a-judge` approach for answer assessment. In our implementation, we use the Qwen2.5-72B-Instruct[26] model deployed with vLLM[27] as the judge model. For remaining benchmarks, we utilize the lmms-eval [28] framework for evaluation. To ensure stable and reliable evaluation results, we report the average performance across 4 independent runs (average@4) for all experiments.

4.2 Main Results

We evaluate ours VADE method on multimodal reasoning benchmarks by training Qwen2.5-VL-3B/7B-Instruct models on our constructed dataset and comparing against vanilla GRPO/GSPO and DAPO baselines. As shown in Table 1, ours VADE consistently outperforms baseline methods, demonstrating its both effectiveness and efficiency.

Our approach demonstrates stable improvements over vanilla methods. After training, our method achieves higher average scores than vanilla baselines across all configurations. Specifically, with Qwen2.5VL-7B-Instruct,

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

Figure 3: Training dynamics of Qwen2.5VL-7B-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.

our method outperforms vanilla GRPO by 2.93% and vanilla GSPO by 2.01% in average performance. Across 20 specific measurements (5 benchmarks $\times$ 4 configurations), our method surpasses vanilla approaches in 16 cases, with minimal performance gaps in the remaining 4 cases. These results consistently validate the effectiveness of our sample selection strategy.

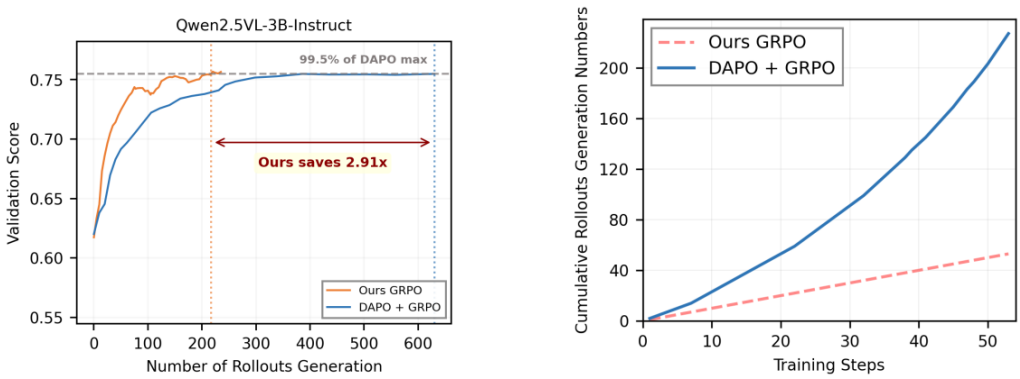
Our method achieves better performance than DAPO with significantly less computational overhead. While DAPO relies on an "over-sample and filter" paradigm that incurs growing computational overhead, VADE achieves better or comparable performance with only one-third of DAPO's rollout budget. Under this constrained setting, VADE achieves higher average scores in 3 out of 4 configurations. More importantly, as shown in Figure 4(a), our method achieves higher validation scores with far fewer rollout generations. Remarkably, VADE attains 99.5% of DAPO's peak performance while using nearly three times fewer rollout inferences, highlighting its superior sampling efficiency and practical scalability.

Our performance gains stem from selecting more informative training data. VADE achieves consistent performance gains by dynamically estimating sample difficulty and selecting the most informative data at each training step, thereby enhancing the learning signal and leading to more effective policy updates. As evidenced in Figure 1 and Figure 3(b), our method maintains a significantly higher proportion of data yielding effective gradients by consistently selecting informative samples. Moreover, unlike DAPO, which wastes computation on repeated rollouts for filtering, VADE leverages each rollout result to update online difficulty estimates, avoiding extra inference costs. This design enables more efficient parameter updates and faster convergence compared to both vanilla optimization algorithms and filtering-based method.

4.3 Training Dynamics

To further validate the efficiency and effectiveness of our VADE, we analyze key training dynamics against baseline methods. Figure 3 illustrates the training dynamics of Qwen2.5VL-7B-Instruct using GRPO algorithm. Additional results for other model configurations are provided in the Appendix. We focus on three critical aspects: validation score, effective gradient ratio, and actor gradient norm:

- The **validation score** serves as the primary indicator of model capability during training. As shown in Figure 3(a), VADE not only achieves higher final performance but also demonstrates accelerated convergence compared to vanilla GRPO, coonfirming its ability to improve final model performance.
- The **effective gradient ratio** measures the proportion of samples that yeilds effective gradients in each



(a) Validation score

(b) Cumulative Rollouts gen numbers

Figure 4: Training dynamics comparison between our method and DAPO. (a) **Validation Score.** The x-axis represents the total count of forward passes through the model to sample responses for all training batches. Our method achieves competitive or superior validation performance with significantly fewer rollout generations, demonstrating substantially higher training efficiency compared to DAPO. (b)**Cumulative Rollout Generations throughout training.** This figure plots the total number of forward passes performed to sample responses for all training batches up to a given step. DAPO incurs significantly more rollout generations than our VADE due to its over-sampling and filtering strategy, demonstrating the superior computational efficiency of our approach.

training batch. Figure 3(b) reveals a striking difference: while vanilla GRPO suffers from rapidly diminishing effective gradients, VADE maintains a consistently high ratio throughout training. This fundamental improvement directly addresses the core *gradient vanishing* problem by ensuring that most training batches contain diverse reward outcomes. The sustained high ratio demonstrates VADE's ability to continuously identify and select samples that challenge the current policy without being trivially easy or impossibly difficult.

- The **actor gradient norm** reflects the magnitude of parameter updates to the policy model during training. As shown in Figure 3(c), VADE produces significantly higher gradient norms compared to vanilla GRPO, indicating more substantial and directed policy updates. This enhancement is achieved through strategic prioritization of samples that provide informative gradient signals, effectively amplifying the learning signal per parameter update.

Furthermore, we compare the computational efficiency of VADE against DAPO. As DAPO relies on repeated rollouts to filter uninformative groups through its "generate-then-filter" paradigm, we use the total number of rollout generations as a fair efficiency metric. Figure 4(b) clearly shows that DAPO incurs substantially more rollout inferences due to this repetitive sampling strategy. The efficiency gap stems from their fundamental difference: DAPO wastes computation on generating and discarding rollouts, while VADE's proactive selection ensures most rollouts contribute to learning. Remarkably, as shown in Figure 4(a), VADE achieves 99.5% of DAPO's peak performance while using several times fewer inference computations. This demonstrates that VADE provides a highly inference-efficient alternative, offering strong performance without the computational burden of oversampling.

4.4 Ablation Study

We conduct ablation studies to validate the design of VADE's three core components: the online difficulty estimator based on information gain objective, the Thompson sampler for balancing exploration and exploitation, and the two-scale prior decay mechanism for robust parameter updates. Table 2 compares variants of our method when training Qwen2.5VL-7B-Instruct with GRPO, reporting the average performance across our five evaluation benchmarks.

Our asymmetric information gain $p(1-p)^2$ outperforms symmetric formulation $p(1-p)$. We first evaluate the information gain objective that guides the online estimator, comparing our asymmetric formulation $\mathcal{I}_i = p(1-p)^2$ against the symmetric alternative $p(1-p)$ used in MMR1[12]. The former actively prioritizes challenging samples, while the latter treats all non-extreme cases uniformly. Results show that our formulation improves average performance by 1.09%, confirming that emphasizing harder problems is beneficial for model advancement.
Our Thompson sampling strategy surpasses greedy selection. We next ablate the Thompson sampler by replacing it with a greedy strategy that selects samples based on the mean estimate $\hat{p}_i = \alpha_i/(\alpha_i+\beta_i)$. The greedy approach overly exploits current knowledge and lacks exploration. In contrast, Thompson sampling naturally balances exploration and exploitation by stochastically sampling from the posterior to periodically select uncertain yet potentially valuable samples. Experimental results show that the greedy variant underperforms our Thompson sampling strategy by 1.50% on average, underscoring the importance of maintaining exploration and exploitation balance for effective long-term learning.
Ours two-scale prior decay exceeds last-update strategy. Finally, we evaluate the two-scale prior decay mechanism against a "last-update" baseline that discards all historical data and uses only the most recent rollout to update (α_i, β_i). By contrast, our method retains historical information with time-aware decay, allowing it to robustly track sample difficulty in non-stationary environments. The performance gain of 1.65% demonstrates the superiority of our proposed update strategy.

In summary, these ablation studies systematically validate the necessity of each component in VADE, confirming that the information gain objective, Thompson sampling, and two-scale prior decay collectively contribute to its strong performance.

Table 2: Ablation study on Qwen2.5VL-7B-Instruct with GRPO, evaluating the impact of key design choices on average performance across five benchmarks.

Method	AVG
Ours	**64.75**
$p(1-p)$	63.66 (-1.09%)
greedy	63.25 (-1.50%)
last update	63.10 (-1.65%)

5 Conclusions

In this work, we identified and addressed the critical gradient vanishing problem in group-based RL, where identical rewards within groups cause advantage estimates to collapse, eliminating effective training signals. We proposed VADE, a variance-aware dynamic sampling framework via online sample-level difficulty estimation. By integrating online difficulty estimation using Beta distributions, Thompson sampling guided by an information-theoretic objective, and a two-scale prior decay mechanism, VADE proactively selects the most informative samples for each training batch without incurring extra rollout costs.

Extensive experiments on multimodal reasoning benchmarks show that VADE consistently outperforms strong baselines in both final performance and sample efficiency. Ablation studies validate the contribution of each component in our framework, confirming that the information gain objective, Thompson sampling, and two-scale prior decay collectively contribute to its strong performance.

Notably, VADE can serves as a plug-and-play module compatible with existing group-based RL pipelines. For example, it can replace the dynamic sampling module in DAPO while preserving other configurations such as token-level policy optimization or clip-higher mechanisms. Future work will extend VADE to broader applications including mathematical reasoning and code generation, and explore integration with other policy optimization frameworks.

References

[1] OpenAI. Learning to reason with llms. OpenAI Technical Report, 2024. Accessed: 2024-01-01.

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[3] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[4] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[5] Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

[6] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[7] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[8] Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua Susskind, and Etai Littwin. Vanishing gradients in reinforcement finetuning of language models. *arXiv preprint arXiv:2310.20703*, 2023.

[9] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[10] Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. Vcrl: Variance-based curriculum reinforcement learning for large language models. *arXiv preprint arXiv:2509.19803*, 2025.

[11] Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*, 2025.

[12] Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, et al. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*, 2025.

[13] 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.

[14] Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, et al. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*, 2025.

[15] Ziniu Li, Congliang Chen, Tianyun Yang, Tian Ding, Ruoyu Sun, Ge Zhang, Wenhao Huang, and Zhi-Quan Luo. Knapsack rl: Unlocking exploration of llms via optimizing budget allocation. *arXiv preprint arXiv:2509.25849*, 2025.

[16] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[17] Arjun K Gupta and Saralees Nadarajah. *Handbook of beta distribution and its applications*. CRC press, 2004.

[18] 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.

[19] 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. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[21] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[22] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[23] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[24] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

[25] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[26] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[27] 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.

[28] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

Appendices

A Dataset Details

We now detail the construction of our training dataset. We select 9 tasks from LLaVA-OneVision-Data[20] and incorporate the training sets of ChartQA[21] and ScienceQA[22], 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.

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[21]	1,300
		tabmwp(cauldron)	219
Science	1325	ScienceQA[22]	1,100
		ai2d(cauldron,llava_format)	225

B 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.

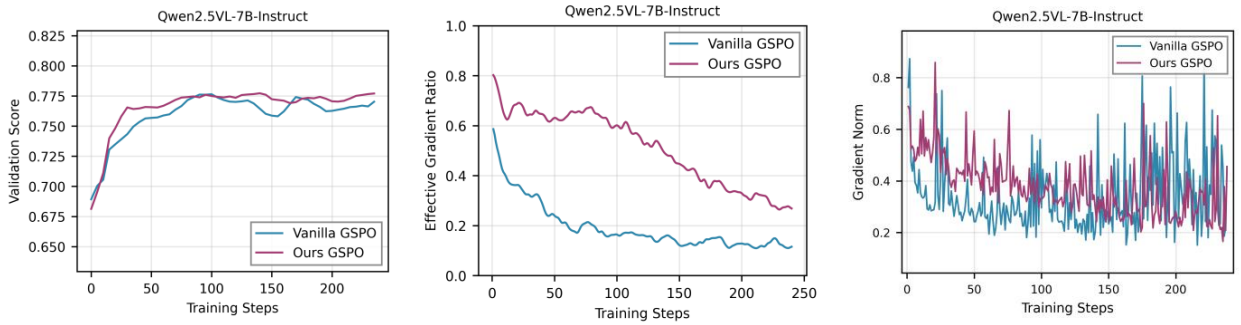
Hyperparameter	Value
Train batch size	512
Max prompt length	8192
Max response length	2048
Filter overlong prompts	True
Rollouts per prompt	8
Total epoches	15
Learning rate	1e-6
ppo_mini_batch_size	128
ppo_micro_batch_size_per_gpu	16
kl_loss_coef	0.01
tensor_model_parallel_size	2
Rollout engine	vLLM
GPU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

C More Experimental Results

C.1 More Training Dynamics

This section presents additional training dynamics results.

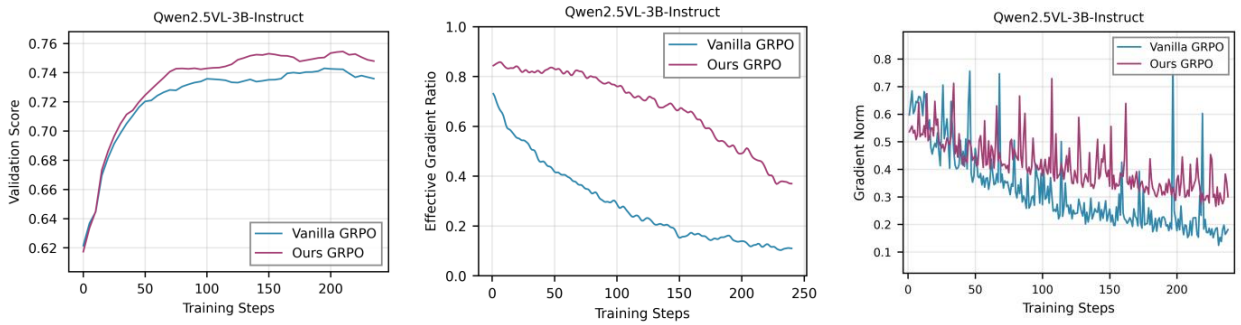
Figure 5 shows the validation score, effective gradient ratio, and actor gradient norm curves comparing our method with vanilla method on Qwen2.5VL-7B-Instruct trained with GSPO. Similarly, Figure 6 presents corresponding results on Qwen2.5VL-3B-Instruct with GRPO, while Figure 7 shows results on Qwen2.5VL-3B-Instruct with GSPO. These results demonstrate that VADE consistently outperforms vanilla baselines across different model scales and algorithms, confirming its effectiveness and general applicability.



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

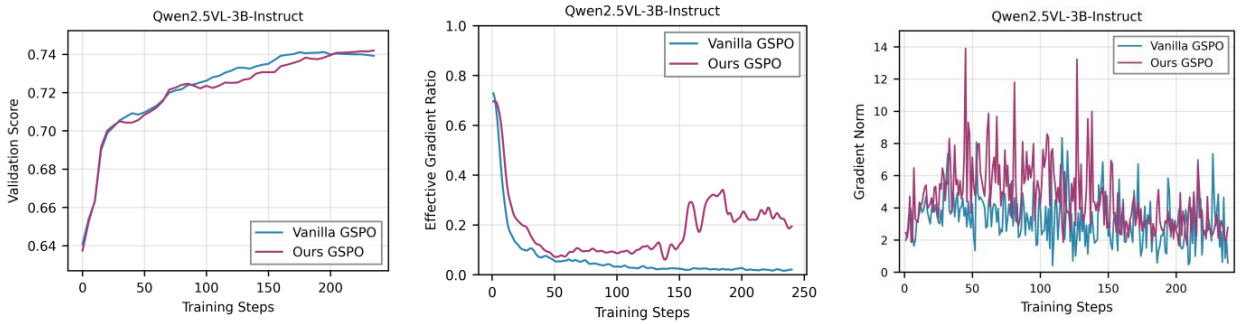
Figure 5: Training dynamics of Qwen2.5VL-7B-Instruct trained with the GSPO algorithm. (a) Validation score, illustrating convergence speed and final performance; (b) Effective gradient ratio, representing the proportion of data with non-uniform rewards (neither all-zero nor all-one) in each training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

Furthermore, Figure 8 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



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

Figure 6: 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.



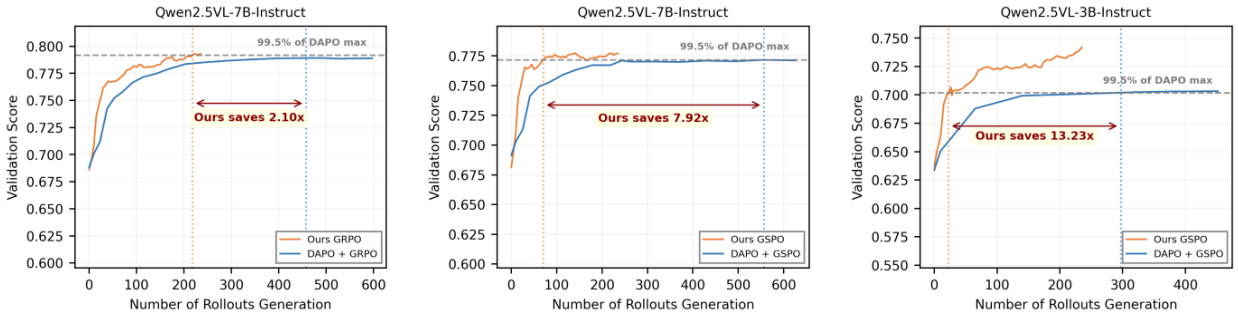
(a) Validation score (b) Effective gradient ratio (c) Gradient norm

Figure 7: 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.

generations under all these settings, highlighting its substantial advantage in training efficiency and general applicability.

C.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.



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

Figure 8: 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