VI-CuRL: Stabilizing Verifier-Independent Reinforcement Learning Reasoning via Confidence-Guided Variance Reduction

Xin-Qiang Cai[1] **Masashi Sugiyama**[1,2]
[1]RIKEN AIP, Tokyo, Japan
[2]The University of Tokyo, Tokyo, Japan

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a dominant paradigm for enhancing Large Language Models (LLMs) reasoning, yet its reliance on external verifiers limits its scalability. Recent findings suggest that RLVR primarily functions by eliciting latent capabilities, motivating the development of verifier-free algorithms. However, in such settings, standard methods like Group Relative Policy Optimization face a critical challenge: destructive gradient variance that often leads to training collapse. To address this issue, we introduce **Verifier-Independent Curriculum Reinforcement Learning (VI-CuRL)**, a framework that leverages the model's intrinsic confidence to construct a curriculum independent from external verifiers. By prioritizing high-confidence samples, VI-CuRL effectively manages the bias-variance trade-off, specifically targeting the reduction of *action* and *problem variance*. We provide a rigorous theoretical analysis, proving that our estimator guarantees asymptotic unbiasedness. Empirically, VI-CuRL promotes stability and consistently outperforms verifier-dependent/independent baselines across math and general reasoning benchmarks with/without verifiers.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has recently emerged as a dominant paradigm for enhancing the reasoning capabilities of large language models (LLMs) in post-training [Guo et al., 2025]. Typically, RLVR relies on an external verifier to provide reward signals that guide the model toward correct reasoning paths [Lightman et al., 2024]. However, obtaining reliable verifiers is often prohibitively expensive or infeasible for many open-ended tasks [Kou et al., 2026]. Meanwhile, recent studies suggest that the success of RLVR may be largely attributed to its ability to elicit the model's latent potential rather than injecting new knowledge [Chen et al., 2024, Guo et al., 2025]. Consequently, there is a growing interest in verifier-free algorithms that can unlock these capabilities without reliance on ground-truth supervision [Liu et al., 2025, Zhou et al., 2025].

Despite its promise, RLVR, like many RL approaches, faces a significant challenge: training instability due to high gradient variance [Yao et al., 2025, Ye et al., 2025]. Traditional RL methods employ various techniques to reduce variance, often introducing a controlled amount of bias to stabilize training, known as the bias-variance trade-off [Geman et al., 1992, Schulman et al., 2016]. In the context of RLVR, particularly in verifier-free settings, this issue is exacerbated. Without effective external supervision to prune incorrect trajectories, the variance in LLM training becomes destructively large, often leading to performance collapse [Zhang et al., 2025b].

To address this instability, curriculum learning presents a promising direction, proposing to train models on a sequence of progressively harder examples to improve convergence [Bengio et al., 2009, Shi et al., 2025a]. While curriculum-based approaches for RLVR, such as Adaptive Curriculum

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

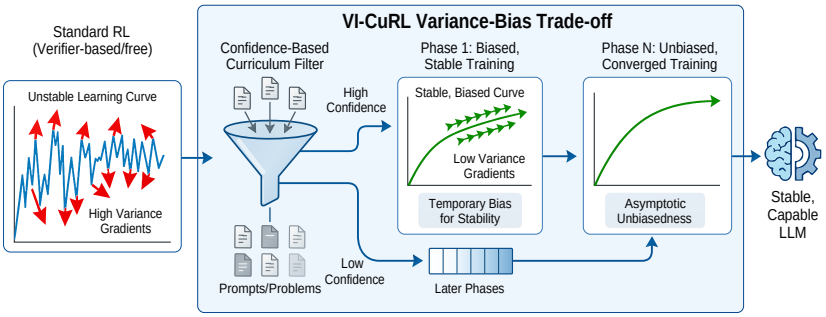


Figure 1: Conceptual overview of VI-CuRL. Unlike standard RL that treats all samples equally, VI-CuRL dynamically selects high-confidence samples to stabilize training via a principled bias-variance trade-off, without accessing external verifiers.

Reinforcement Finetuning (AdaRFT) [Shi et al., 2025b] and Variance-based Curriculum Reinforcement Learning (VCRL) [Jiang et al., 2025], have been proposed, they are verifier-dependent and rely on verifier-provided information to construct the curriculum. While in verifier-free scenarios, however, the destructive impact of variance is far more severe, and such external guidance is absent [Zhang et al., 2025b]. Therefore, we propose **Verifier-Independent Curriculum Reinforcement Learning (VI-CuRL)**, as illustrated in Figure 1. VI-CuRL uses an annealed curriculum to stabilize RL process with confidence as the curriculum metric. Instead of relying on the variance of external training data [Zhang et al., 2025b] or signals from verifiers [Lightman et al., 2024, Wen et al., 2025], VI-CuRL leverages an intrinsic prompt-level curriculum score, which instantiated as entropy-derived confidence, as a verifier-free curriculum statistic for data selection. By prioritizing high-confidence samples that form a lower-uncertainty subset in the early stages, our method achieves a favorable bias-variance balance. We guarantee that, as the curriculum progresses, the surrogate objective asymptotically converges to the true objective.

We formalize the curriculum bias–noise trade-off and evaluate VI-CuRL on mathematical reasoning and general-knowledge benchmarks. Oracle rewards test composability; intrinsic rewards test stability when reliable external reward supervision is absent. Our contribution concerns prompt selection, independently of the chosen reward source.

2 Preliminaries and Problem Formulation

In this section, we introduce the preliminaries and problem formulation.

Notation. We consider a prompt x drawn from a data distribution $p(x)$. At each iteration, we maintain a trainable policy π_θ parameterized by $\theta \in \mathbb{R}^d$, and a behavior policy $\pi_{\theta_{\text{old}}}$ (the policy from the current outer iteration). We also assume a fixed reference policy π_{ref} (typically the initial Supervised Fine-Tuning (SFT) policy) for Kullback-Leibler (KL) regularization. Given a prompt x, we sample G trajectories $\{y_1, \ldots, y_G\}$ from $\pi_{\theta_{\text{old}}}(\cdot|x)$, where each $y_i = a_{i,1:T_i}$ consists of T_i tokens. We denote the state at step t as $s_{i,t} = (x, a_{i,<t})$. The group-relative advantage A_i for trajectory i is:

$$A_i = \frac{R(x, y_i) - \bar{R}}{\sigma_R + \epsilon}, \tag{1}$$

where $\bar{R}$ and σ_R are the mean and standard deviation of rewards within the group, and ϵ is a small constant to prevent division by zero.

Base Surrogate Objective. We adopt the Group Relative Policy Optimization (GRPO) framework [Shao et al., 2024], which we view as optimizing a Proximal Policy Optimization (PPO) style surrogate objective with group-relative advantages and explicit KL regularization. Specifically, we define the log-ratio $r_{i,t}(\theta)$ and probability ratio $\rho_{i,t}(\theta)$ as:

$$r_{i,t}(\theta) := \log \pi_\theta(a_{i,t}|s_{i,t}) - \log \pi_{\theta_{\text{old}}}(a_{i,t}|s_{i,t}),$$
$$\rho_{i,t}(\theta) := \exp(r_{i,t}(\theta)).$$

The per-prompt, per-group surrogate loss averages over rollouts and tokens to ensure boundedness:

$$\ell(\theta; x, y_{1:G}) := \frac{1}{G}\sum_{i=1}^{G}\frac{1}{T_i}\sum_{t=1}^{T_i}\Big[L_{\text{clip}}(i,t;\theta) - \beta_{\text{KL}}\text{KL}(\pi_\theta(\cdot|s_{i,t})||\pi_{\text{ref}}(\cdot|s_{i,t}))\Big], \tag{2}$$

where the standard PPO-Clip term is:

$$L_{\text{clip}}(i,t;\theta) := \min\Big(\rho_{i,t}(\theta)A_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i\Big).$$

This normalization makes the loss scale approximately length-invariant (independent of T_i) in practice. The global training objective is to maximize the expected value of this surrogate:

$$\mathcal{L}(\theta) := \mathbb{E}_{x\sim p(x)}\ \mathbb{E}_{y_{1:G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\big[\,\ell(\theta; x, y_{1:G})\,\big]. \tag{3}$$

In standard GRPO, we estimate the gradient of $\mathcal{L}(\theta)$ by sampling prompts and groups from $\pi_{\theta_{\text{old}}}$ and performing multiple epochs of stochastic gradient ascent on ℓ.

Curriculum RL. Standard RL treats all samples from $p(x)$ equally. Curriculum methods effectively optimize a sequence of objectives $\mathcal{L}_t(\theta)$, where the data distribution is re-weighted. The goal is to start with a low-uncertainty distribution where gradient variance is low, and gradually anneal towards the target distribution $p(x)$. To this end, a prompt-level curriculum score is introduced and denoted as $c(x) \in [0,1]$. Then the curriculum mask is defined as $w_t(x) = \mathbf{1}[c(x) \geq \tau_t]$ [Kadavath et al., 2022, Huang et al., 2024, Ren et al., 2023]. We formalize this signal as follows:

Definition 2.1 (Curriculum Score)**.** Quantifying confidence in LLMs is achieved by aggregating token-level probabilities. To obtain a stable signal $c(x) \in [0,1]$ that depends only on the prompt, we first define the *normalized uncertainty* $u(x)$ as the expectation over trajectories sampled from the behavior policy $\pi_{\theta_{\text{old}}}$:

$$u(x) := \mathbb{E}_{y\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{H(\pi_{\theta_{\text{old}}}(\cdot|s_t))}{\log|\mathcal{V}|}\right], \tag{4}$$

where H is the Shannon entropy and $|\mathcal{V}|$ is the vocabulary size. The curriculum score is then instantiated as:

$$c(x) = 1 - u(x) \in [0,1]. \tag{5}$$

The confidence of a LLM on a prompt is a verifier-free curriculum statistic for data selection; it does not verify trajectories but gates prompt-level exposure to control gradient variance. Therefore, we leverage it here as an effective proxy for *variance control* to stabilize training, rather than as a ground-truth verification signal.

3 Verifier-Independent Curriculum Reinforcement Learning

We propose **Verifier-Independent Curriculum Reinforcement Learning (VI-CuRL)**, a novel framework for solving the high-variance challenge independent from verifiers.

3.1 The VI-CuRL Objective

To formally operationalize our curriculum, we define a sequence of curriculum phases indexed by $t = 1, \ldots, T$. Each phase is associated with a confidence threshold τ_t that determines which samples are included in the training process.

Curriculum Mask and Retention Rate. For a fixed behavior policy and threshold within curriculum phase t, define the population retention probability of the curriculum mask as

$$\beta_t = \mathbb{E}_{x\sim p(x)}[w_t(x)] \in (0,1]. \tag{6}$$

This probability is the mass of the retained region under $p(x)$ and normalizes the population objective and gradient estimator in Eqs. (8) and (9).

Let $B \geq 1$ be the number of prompts sampled at one outer step. The schedule specifies $\beta_t^{\text{target}} \in (0,1]$, the desired fraction of these prompts to retain. The integer retention count is $k_t = \lceil \beta_t^{\text{target}} B \rceil$,

so $1 \leq k_t \leq B$. Exact top-k_t selection produces batch masks $w_j \in \{0,1\}$, with $j \in \{1,\ldots,B\}$ indexing prompt occurrences in the sampled batch. The realized batch retention rate is

$$\hat{\beta}_t = \frac{1}{B}\sum_{j=1}^{B} w_j = \frac{k_t}{B}. \tag{7}$$

We normalize the empirical batch objective by $\hat{\beta}_t$. The batch masks are constructed jointly by ranking the confidence scores, as detailed below.

The Weighted Surrogate Objective. For each phase t, we optimize a re-weighted version of the standard surrogate $\mathcal{L}(\theta)$ that restricts learning to the selected subspace. To maintain the scale of the original objective, we introduce the importance sampling weight $1/\beta_t$:

$$\mathcal{L}_t(\theta) := \mathbb{E}_{x\sim p(x)}\mathbb{E}_{y_{1:G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\Big[\frac{w_t(x)}{\beta_t}\,\ell(\theta;x,y_{1:G})\Big]. \tag{8}$$

This weighted objective trades bias for variance. By focusing on the high-confidence region, it reduces the effective variance of the gradient estimator, albeit by optimizing a biased data distribution.

The Curriculum Gradient Estimator. To optimize $\mathcal{L}_t(\theta)$, we use the standard auto-differentiation subgradient of the clipped PPO surrogate. Critically, the advantages A_i, masks $w_t(x)$, and weights β_t are treated as constants ("stop-gradients") during the inner loop updates:

$$\hat{g}_t := \frac{w_t(x)}{\beta_t}\,\nabla_\theta \ell(\theta;x,y_{1:G}). \tag{9}$$

3.2 Curriculum Schedule and Algorithm

We now detail the practical implementation of the VI-CuRL framework. The optimization process is divided into T steps. At each step t, we seek to effectively filter the training batch to retain only the samples where the model exhibits high confidence.

Dynamic Quantile Thresholding. We implement the target retention schedule through exact top-k_t selection, adapting the confidence threshold to the current batch. For $\mathcal{B} = \{x_j\}_{j=1}^{B}$, write $c_j = c(x_j)$ using Eq. (5), and let $c_{(1)} \leq \cdots \leq c_{(B)}$ be the ordered scores, with $c_{(j)}$ denoting the j-th smallest score. The retention count and threshold are

$$k_t = \left\lceil \beta_t^{\text{target}} B \right\rceil, \qquad \tau_t = c_{(B-k_t+1)} = c_{(\lfloor(1-\beta_t^{\text{target}})B\rfloor+1)}. \tag{10}$$

The index $B-k_t+1$ lies in $\{1,\ldots,B\}$ and identifies the smallest score among the k_t highest-scoring prompts. The two threshold expressions agree because $B - \lceil \beta_t^{\text{target}} B\rceil = \lfloor(1-\beta_t^{\text{target}})B\rfloor$.

To obtain exactly k_t retained prompts even when scores coincide, let $n_t^{>} = |\{j : c_j > \tau_t\}|$ be the number of prompts strictly above the threshold, where $|\cdot|$ denotes set cardinality. We retain all $n_t^{>}$ such prompts and select the remaining $k_t - n_t^{>}$ from those with $c_j = \tau_t$, taking smaller batch indices first. Denote the resulting set of retained indices by $\mathcal{I}_t \subseteq \{1,\ldots,B\}$ and define $w_j = \mathbf{1}[j \in \mathcal{I}_t]$. With distinct scores, this rule reduces to $w_j = \mathbf{1}[c_j \geq \tau_t]$. In all cases, it guarantees

$$\sum_{j=1}^{B} w_j = k_t, \qquad \hat{\beta}_t = \frac{k_t}{B}, \qquad 0 \leq \hat{\beta}_t - \beta_t^{\text{target}} < \frac{1}{B}. \tag{11}$$

Algorithm. VI-CuRL follows the GRPO outer/inner-loop structure. At outer step t, we set $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$, sample B prompts, and collect G rollouts per prompt to compute rewards, advantages, and confidence scores. The linear schedule determines β_t^{target}, after which we form $\mathcal{I}_t$ using the ranking and tie-breaking rule above and set $\hat{\beta}_t = k_t/B$. Using the base surrogate ℓ and the collected rollout groups $y_{j,1:G}$, define the empirical curriculum objective as

$$\widehat{\mathcal{L}}_t(\theta) = \frac{1}{B}\sum_{j=1}^{B}\frac{w_j}{\hat{\beta}_t}\ell(\theta;x_j,y_{j,1:G}) = \frac{1}{k_t}\sum_{j\in\mathcal{I}_t}\ell(\theta;x_j,y_{j,1:G}). \tag{12}$$

Algorithm 1 Verifier-Independent Curriculum RL

Input: Dataset $\mathcal{D}$, Initial Policy π_{ref}, Total steps T, Curriculum steps T_{curr}, Batch size B, Number of epochs K, Learning rate η.
Initialize $\pi_\theta \leftarrow \pi_{\text{ref}}$.
for $t = 1$ **to** T **do**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$.
  Sample batch $\mathcal{B} = \{x_j\}_{j=1}^{B} \sim \mathcal{D}$.
  for each prompt $x_j \in \mathcal{B}$ **do**
    Sample group $\{y_{j,k}\}_{k=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x_j)$.
    Compute rewards $R_{j,k}$ and advantages $A_{j,k}$ using the chosen reward source.
    Compute curriculum score c_j from $\pi_{\theta_{\text{old}}}$. ▷ *Used only to form w_j; receives no gradient.*
  end for
  Determine β_t^{target} via the schedule; set $k_t = \lceil \beta_t^{\text{target}} B \rceil$.
  Sort the scores in ascending order; set $\tau_t = c_{(B-k_t+1)}$ as in Eq. (10).
  Form $\mathcal{I}_t$ from the k_t highest-scoring prompt indices, breaking ties by ascending batch index.
  Set $w_j = \mathbf{1}[j \in \mathcal{I}_t]$ and $\hat{\beta}_t = k_t/B$ as in Eq. (7).
  for $k = 1$ **to** K **do**
    Sample $\mathcal{M} \subseteq \{1, \ldots, B\}$ uniformly with fixed size $m \in \{1, \ldots, B\}$.
    Compute $\hat{g} = \hat{g}_t^{\mathcal{M}}$ using Eq. (13), with normalization $w_j/\hat{\beta}_t$.
    Update $\theta \leftarrow \theta + \eta \hat{g}$.
  end for
end for

During the inner loop, we fix the outer-batch data and apply the stop-gradient convention to the advantages, masks, and normalization. At each update, let $m \in \{1, \ldots, B\}$ denote the minibatch size and draw a fresh set of indices $\mathcal{M} \subseteq \{1, \ldots, B\}$ uniformly from all subsets of size m of the full outer batch. Thus $|\mathcal{M}| = m$, and every prompt index has inclusion probability m/B. The minibatch gradient estimator is

$$\hat{g}_t^{\mathcal{M}} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \frac{w_j}{\hat{\beta}_t} \nabla_\theta \ell(\theta; x_j, y_{j,1:G}). \tag{13}$$

Conditional on the outer-batch data, its fixed curriculum mask, and the current parameters, the expectation over the fresh minibatch draw satisfies

$$\mathbb{E}_{\mathcal{M}}[\hat{g}_t^{\mathcal{M}}] = \frac{1}{m} \sum_{j=1}^{B} \frac{m}{B} \frac{w_j}{\hat{\beta}_t} \nabla_\theta \ell(\theta; x_j, y_{j,1:G}) = \nabla_\theta \widehat{\mathcal{L}}_t(\theta). \tag{14}$$

Early in training, the schedule concentrates the empirical objective on high-confidence prompts, allowing the policy to start from a low-uncertainty training subset; At full retention, it recovers $B^{-1} \sum_{j=1}^{B} \ell(\theta; x_j, y_{j,1:G})$ exactly, ensuring unbiased optimization of the full distribution. Algorithm 1 summarizes the pseudocode for VI-CuRL.

4 Theoretical Analysis of the Bias-Variance Trade-off

In this section, we formalize the properties of VI-CuRL. Our analysis first addresses the "bias" component of the trade-off. We then dissect the "variance" component to reveal the source of the training stability.

4.1 Consistency of the Objective

A foundational property of our framework is that the normalized gradient estimator $\hat{g}_t(x, y; \theta)$ is, by construction, an unbiased estimator of the gradient of the surrogate objective $\mathcal{L}_t(\theta)$. That is, $\mathbb{E}[\hat{g}_t(x, y; \theta)] = \nabla_\theta \mathcal{L}_t(\theta)$. This ensures our algorithm correctly optimizes the objective for the current curriculum phase. The critical theoretical question, however, is whether maximizing these time-dependent surrogate objectives consistently leads to maximizing the base surrogate. Our first theorem confirms that it does:

Theorem 4.1 (Consistency of the Objective)**.** *Assuming the per-prompt surrogate loss is bounded,* $|\ell(\theta; x, y_{1:G})| \leq L_{\max}$*, the absolute difference between the true surrogate objective* $\mathcal{L}(\theta)$ *and the weighted objective* $\mathcal{L}_t(\theta)$ *is bounded by* $(1 - \beta_t)$*:*

$$|\mathcal{L}(\theta) - \mathcal{L}_t(\theta)| \leq 2L_{\max}(1 - \beta_t). \tag{15}$$

The proof can be found in Appendix C.1. The bounded-surrogate assumption $|\ell(\theta; x, y_{1:G})| \leq L_{\max}$ is mild in practice with bounded rewards, group-normalized advantage, ratio clipping, and per-token KL penalty [Ouyang et al., 2022, Shao et al., 2024]. Meanwhile, as the curriculum is annealed ($\tau_t \to 0$), the retention rate $\beta_t \to 1$, and the bound converges to zero, ensuring that $\mathcal{L}_t(\theta) \to \mathcal{L}(\theta)$. This theorem guarantees that the approximation error diminishes over time. Meanwhile, this result does not depend on the correctness or calibration of the confidence score. It only requires the retention rate to anneal to one.

4.2 A Deeper Investigation of Variance

Having established that our objective is asymptotically correct, we now analyze how the curriculum tames the variance that destabilizes standard RL. To this end, here we decompose the variance into its constituent parts.

Definition 4.2 (Vector Variance)**.** For a random vector $\mathbf{Z} \in \mathbb{R}^d$, we define its variance as the trace of its covariance matrix:

$$\mathrm{Var}(\mathbf{Z}) := \mathbb{E}[\|\mathbf{Z} - \mathbb{E}[\mathbf{Z}]\|^2].$$

Theorem 4.3 (Variance Decomposition)**.** *Let* $g(x, y_{1:G}; \theta) := \nabla_\theta \ell(\theta; x, y_{1:G})$ *be the gradient of the surrogate loss, and let* $\bar{g}(x; \theta)(x; \theta) = \mathbb{E}_{y_{1:G}|\pi_{\theta_{\text{old}}}}[g]$ *be the expected gradient for a fixed prompt* x*. The variance of the VI-CuRL estimator* $\hat{g}_t(x, y; \theta)$ *can be decomposed into three sources:*

$$\mathrm{Var}(\hat{g}_t(x, y; \theta)) = \underbrace{\frac{\sigma^2_{g,t}}{\beta_t}}_{\textit{Action Variance}} + \underbrace{\frac{V_{\text{prob},t}}{\beta_t}}_{\textit{Problem Variance}} + \underbrace{\frac{1 - \beta_t}{\beta_t}\|\nabla_\theta \mathcal{L}_t(\theta)\|^2}_{\textit{Masking Variance}}, \tag{16}$$

where $\sigma^2_{g,t} := \mathbb{E}_{x|w_t=1}[\mathrm{Var}_{y_{1:G}|\pi_{\theta_{\text{old}}}}(g)]$ *is the average action variance within the selected set of problems, and* $V_{\text{prob},t} := \mathrm{Var}_{x|w_t=1}(\bar{g}(x; \theta))$ *is the variance of mean gradients within that same set.*

The proof can be found in Appendix C.2.

This decomposition motivates our focus on action and problem variance, which capture within-prompt rollout variability and across-prompt gradient heterogeneity, respectively [Tucker et al., 2018, Yao et al., 2026]. These intrinsic sources of gradient noise can persist even at full retention and can perturb the direction of policy updates. For the population estimator analyzed above, the masking covariance $\frac{1-\beta_t}{\beta_t}\mu_t\mu_t^\top$, with $\mu_t = \nabla_\theta \mathcal{L}_t(\theta)$, contributes only along the selected mean-gradient direction. We thus prioritize reducing action and problem variability within the selected subset, while treating masking as an additional cost of partial retention which becomes negligible as $\beta_t \to 1$.

Beyond this exact identity, Appendix C.3 and Appendix C.4 contain two conditional bounds. They formalize regimes in which gradient norms or selected-subset variances shrink sufficiently fast as confidence decreases.

5 Experiments

To validate the effectiveness of VI-CuRL, we conducted experiments on several reasoning tasks. In addition to the main results, we provide additional experiments in Appendix B, including learning curve analysis across different reward settings, curriculum control experiments, confidence signal analysis, long-output robustness, and mechanism analysis. Code is provided in `https://github.com/caixq1996/VI-CURL`.

5.1 Experimental Setup

We trained on four backbones, *Qwen2.5-Math-1.5B*, *DeepSeek-R1-Distill-Qwen-1.5B*, *Llama-3.2-3B-Instruct*, and *Qwen2.5-Math-7B*. Rewards for the oracle baselines came from a rule-based checker that

Table 1: Main Results (Part I): **Pass@1** performance for **1.5B parameter models**. The training dataset is DeepScaleR. Blue rows highlight our method.

Dataset	AIME 2024	AIME 2025	AMC 2023	Math500	Minerva MATH	Olympiad Bench	Average
Qwen2.5-Math-1.5B							
Base (No RL)	6.0 ± 1.9	4.0 ± 0.6	34.2 ± 0.2	47.5 ± 0.3	5.1 ± 0.4	25.1 ± 0.4	20.3 ± 0.6
Oracle Reward (w. Verifier)							
No Curriculum	15.0 ± 0.4	5.6 ± 0.6	50.3 ± 0.6	69.4 ± 0.4	17.8 ± 0.6	31.6 ± 0.0	31.6 ± 0.4
VCRL	12.9 ± 2.0	5.5 ± 1.0	48.8 ± 2.0	69.2 ± 0.2	18.7 ± 0.0	31.3 ± 0.2	31.1 ± 0.9
AdaRFT	13.2 ± 1.5	6.2 ± 1.4	48.6 ± 2.0	68.7 ± 0.4	20.1 ± 0.5	29.0 ± 0.3	31.0 ± 1.0
SENT	15.6 ± 1.3	8.1 ± 1.9	52.1 ± 1.1	70.0 ± 1.0	14.3 ± 1.9	32.0 ± 0.7	32.0 ± 1.3
VI-CuRL	**15.9 ± 1.2**	**9.6 ± 0.7**	**52.9 ± 1.3**	**70.1 ± 1.5**	**24.3 ± 1.4**	**34.8 ± 3.9**	**34.6 ± 1.7**
Verifier-Free: Majority Vote							
TTRL	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.2 ± 0.0	0.4 ± 0.0	0.0 ± 0.0	0.1 ± 0.0
VI-CuRL	**15.9 ± 1.1**	**8.7 ± 1.1**	**53.1 ± 1.3**	**72.3 ± 0.6**	**21.6 ± 4.0**	**33.5 ± 4.5**	**34.2 ± 2.1**
Verifier-Free: Entropy							
RENT	0.0 ± 0.0	0.0 ± 0.0	10.8 ± 0.5	9.4 ± 0.1	4.6 ± 0.1	3.7 ± 0.2	4.8 ± 0.2
EMPO	5.2 ± 1.0	3.1 ± 0.8	15.4 ± 2.1	18.2 ± 1.5	6.5 ± 1.2	8.1 ± 0.9	9.4 ± 1.3
VI-CuRL	**13.8 ± 0.8**	**6.0 ± 0.7**	**48.6 ± 0.9**	**69.9 ± 1.9**	**19.7 ± 4.0**	**32.0 ± 0.3**	**31.7 ± 1.4**
DeepSeek-R1-Distill-Qwen-1.5B							
Base (No RL)	29.5 ± 0.4	20.6 ± 0.9	69.3 ± 0.6	83.1 ± 0.3	24.9 ± 0.6	43.6 ± 0.5	45.2 ± 0.6
Oracle Reward (w. Verifier)							
No Curriculum	31.0 ± 0.7	26.3 ± 0.5	79.1 ± 0.9	88.7 ± 0.4	29.5 ± 0.1	50.1 ± 0.2	50.8 ± 0.5
VCRL	33.2 ± 1.4	26.8 ± 1.6	80.2 ± 0.6	90.8 ± 0.8	30.1 ± 2.0	55.5 ± 4.6	52.8 ± 1.8
AdaRFT	35.0 ± 0.9	**29.9 ± 1.7**	82.0 ± 1.6	91.1 ± 1.0	32.7 ± 1.0	56.3 ± 1.0	54.5 ± 1.2
SENT	34.7 ± 2.2	28.1 ± 1.3	81.6 ± 1.2	91.2 ± 1.1	33.0 ± 1.5	55.5 ± 1.2	54.0 ± 1.4
VI-CuRL	**37.1 ± 1.2**	29.6 ± 1.9	**82.7 ± 1.3**	**92.3 ± 1.1**	**33.8 ± 3.4**	**58.6 ± 2.5**	**55.7 ± 1.9**
Verifier-Free: Majority Vote							
TTRL	24.1 ± 1.2	21.1 ± 1.6	64.0 ± 1.7	75.4 ± 0.8	23.1 ± 0.9	39.5 ± 0.6	41.2 ± 1.1
VI-CuRL	**40.6 ± 1.7**	**34.0 ± 2.5**	**87.4 ± 0.9**	**93.5 ± 1.0**	**35.0 ± 0.5**	**60.9 ± 1.3**	**58.6 ± 1.3**
Verifier-Free: Entropy							
RENT	25.3 ± 0.5	23.2 ± 0.5	71.1 ± 1.0	85.3 ± 0.8	31.5 ± 2.8	48.0 ± 2.0	47.4 ± 1.3
EMPO	34.0 ± 2.4	28.0 ± 1.3	73.6 ± 1.5	83.7 ± 1.2	31.9 ± 2.4	41.6 ± 1.3	48.8 ± 1.7
VI-CuRL	**38.0 ± 1.4**	**31.0 ± 1.0**	**83.8 ± 1.0**	**92.8 ± 0.3**	**32.8 ± 1.1**	**57.8 ± 0.5**	**56.0 ± 0.9**

extracted the final `\boxed{\cdot}` answer and tested numeric/rational equivalence. For the verifier-free reward settings, we relied solely on intrinsic reward signals (e.g., majority vote or entropy); the VI-CuRL curriculum remains verifier-independent in all settings. We implemented VI-CuRL as a curriculum hook within `VERL` [Sheng et al., 2024], following a PPO-style GRPO recipe. In the math domain, we used DeepScaleR [Luo et al., 2025] as the training set, and the evaluation used six verifiable math suites AIME-2024, AIME-2025, AMC-2023, MATH500, MINERVA MATH, and OLYMPIADBENCH. To evaluate the non-mathematical general knowledge capabilities, we selected the Llama-3.2-3B-Instruct model for these experiments. In the general knowledge domain, we used WEBINSTRUCT-VERIFIED [Ma et al., 2025], and filtered out mathematics prompts from the data, and the evaluation used MMLU-PRO [Wang et al., 2024], GPQA-DIAMOND [Rein et al., 2024], THEOREMQA [Chen et al., 2023], and WEBINSTRUCT-VALIDATION [Ma et al., 2025]. We reported *Pass@1* and *Pass@8* with 16 samples and averaged over 5 random seeds. Compute was 8×A100 (40GB) GPUs servers. More details can be found in Appendix D.

Baselines We evaluated VI-CuRL against two categories of baselines:

- **Curriculum RL with Ground Truth**: We compared against **VCRL** [Jiang et al., 2025], **AdaRFT** [Shi et al., 2025b], and **SENT** [Cao et al., 2025]. They were curriculum-based RLVR algorithms. While VCRL and AdaRFT fundamentally relied on access to ground-truth labels (verifiers) to guide the curriculum, unlike VI-CuRL which is verifier-independent. SENT constructs an offline semantic-entropy curriculum. In contrast, VI-CuRL builds

Table 2: Main Results (Part II): **Pass@1** performance for **Llama3.2-3B-Instruct** and **Qwen2.5-Math-7B**. The training dataset is DeepScaleR. Blue rows highlight our method.

Dataset	AIME 2024	AIME 2025	AMC 2023	Math500	Minerva MATH	Olympiad Bench	Average
Llama3.2-3B-Instruct							
Base (No RL)	5.7 ± 1.2	0.6 ± 0.4	17.2 ± 1.5	34.8 ± 0.7	4.8 ± 0.3	12.7 ± 0.2	12.6 ± 0.7
Oracle Reward (w. Verifier)							
No Curriculum	6.0 ± 1.0	1.0 ± 0.4	24.3 ± 1.7	40.4 ± 0.5	8.9 ± 0.7	15.0 ± 0.5	15.9 ± 0.8
VCRL	4.0 ± 1.1	1.3 ± 0.9	24.1 ± 1.4	37.0 ± 0.3	7.8 ± 0.2	17.1 ± 0.8	15.2 ± 0.8
AdaRFT	4.5 ± 1.6	1.6 ± 0.7	25.7 ± 0.4	**44.0 ± 0.3**	8.5 ± 0.1	17.8 ± 0.7	17.0 ± 0.6
SENT	6.5 ± 0.8	1.4 ± 0.4	26.7 ± 1.5	39.9 ± 0.8	8.7 ± 1.5	18.6 ± 1.4	17.0 ± 1.1
VI-CuRL	**6.8 ± 0.9**	**1.9 ± 0.3**	**29.7 ± 1.1**	43.0 ± 0.7	**9.4 ± 1.1**	**19.5 ± 0.7**	**18.4 ± 0.8**
Verifier-Free: Majority Vote							
TTRL	**5.7 ± 1.3**	0.6 ± 0.6	18.5 ± 2.3	41.9 ± 1.0	8.8 ± 0.3	**15.4 ± 0.6**	15.2 ± 1.0
VI-CuRL	5.2 ± 1.1	**0.8 ± 0.5**	**19.6 ± 2.3**	**46.2 ± 1.5**	**10.8 ± 0.2**	15.1 ± 0.6	**16.3 ± 1.0**
Verifier-Free: Entropy							
RENT	0.6 ± 0.6	0.0 ± 0.0	3.0 ± 1.1	3.2 ± 0.0	2.1 ± 0.2	1.2 ± 0.1	1.7 ± 0.3
EMPO	2.1 ± 0.5	0.2 ± 0.1	8.5 ± 1.2	10.1 ± 1.1	4.5 ± 0.6	5.2 ± 0.8	5.1 ± 0.7
VI-CuRL	**6.2 ± 1.4**	**0.8 ± 0.5**	**21.6 ± 1.4**	**42.6 ± 0.9**	**9.7 ± 0.4**	**14.7 ± 0.5**	**15.9 ± 0.9**
Qwen2.5-Math-7B							
Base (No RL)	12.7 ± 0.9	5.8 ± 0.7	44.4 ± 2.2	52.0 ± 0.4	9.8 ± 0.7	26.4 ± 0.4	25.2 ± 0.9
Oracle Reward (w. Verifier)							
No Curriculum	29.2 ± 2.5	13.5 ± 1.2	62.8 ± 1.9	**78.9 ± 0.5**	24.2 ± 0.7	37.8 ± 0.5	41.1 ± 1.2
VCRL	22.0 ± 1.5	11.1 ± 0.7	57.0 ± 1.6	75.1 ± 0.5	22.1 ± 0.7	35.6 ± 0.5	37.2 ± 0.9
AdaRFT	26.9 ± 0.6	12.9 ± 0.7	63.0 ± 0.8	76.6 ± 1.9	22.5 ± 6.9	39.1 ± 2.6	40.2 ± 2.3
SENT	26.2 ± 0.7	10.8 ± 0.9	53.2 ± 1.3	69.2 ± 1.2	22.4 ± 2.0	31.2 ± 1.0	35.5 ± 1.2
VI-CuRL	**29.9 ± 0.9**	**13.9 ± 1.4**	**63.8 ± 0.5**	78.8 ± 0.6	**26.3 ± 5.5**	**45.0 ± 3.0**	**43.0 ± 2.0**
Verifier-Free: Majority Vote							
TTRL	29.0 ± 1.5	14.0 ± 1.0	64.8 ± 1.4	**80.5 ± 0.4**	22.8 ± 0.3	39.5 ± 0.4	41.8 ± 0.8
VI-CuRL	**33.9 ± 0.7**	**14.4 ± 0.9**	**65.4 ± 0.4**	79.8 ± 0.5	**25.9 ± 0.6**	**41.2 ± 0.4**	**43.4 ± 0.6**
Verifier-Free: Entropy							
RENT	13.7 ± 0.6	7.2 ± 0.4	55.7 ± 1.1	74.4 ± 2.1	**24.8 ± 4.5**	32.9 ± 4.4	34.8 ± 2.2
EMPO	18.5 ± 1.5	9.4 ± 0.8	58.2 ± 2.0	75.1 ± 1.2	21.0 ± 2.5	34.5 ± 1.8	36.1 ± 1.6
VI-CuRL	**25.0 ± 1.8**	**13.1 ± 1.1**	**63.0 ± 2.8**	**76.8 ± 0.6**	19.1 ± 5.0	**37.0 ± 0.7**	**39.0 ± 2.0**

the curriculum online at each policy update, so the curriculum explicitly trades transient selection bias for lower gradient variance and anneals back to the full distribution as $\beta_t \to 1$.

- **Standard RL (Oracle)**: Standard RLVR training using the true verifier reward. This served as a topline for performance.
- **TTRL**: A verifier-free baseline where the reward was determined by self-consistency (majority voting) among rollout samples [Zuo et al., 2025].
- **RENT**: A verifier-free baseline that used the prediction entropy of the model's outputs as the reward signal (penalizing high entropy) [Prabhudesai et al., 2025]. VI-CuRL used RENT's reward signal in the verifier-free entropy setting.
- **EMPO**: Entropy Minimization Policy Optimization [Zhang et al., 2025a], an entropy-based verifier-free baseline that combines entropy minimization with policy optimization.

5.2 Main Results

Mathematical Reasoning Performance. Table 1 and Table 2 summarize the mathematical reasoning performance of our method compared to standard baselines. We evaluated the models across two distinct settings: Oracle Reward (with verifier) and Verifier-Free (intrinsic reward). In the Oracle setting, VI-CuRL consistently matches or outperforms standard curriculum RL methods (such as VCRL, AdaRFT, and SENT) across all model scales, from 1.5B to 7B parameters. Notably, on harder benchmarks like AIME and OlympiadBench, VI-CuRL achieves better prformance over the baselines, demonstrating that confidence-based curriculum selection can stabilize updates and improve the

Table 3: General Knowledge Results: **Pass@1** performance for **Llama3.2-3B-Instruct**. The training dataset is WebInstruct-Verified. Blue rows highlight our method.

Dataset	MMLU-Pro	GPQA-Diamond	TheoremQA	WebInstruct-Val	Average
Base (No RL)	34.8 ± 0.4	25.3 ± 0.3	5.1 ± 0.2	20.6 ± 0.4	21.5 ± 0.3
Oracle Reward (w. Verifier)					
No Curriculum	36.2 ± 0.5	26.3 ± 0.2	10.4 ± 0.4	26.0 ± 0.5	24.7 ± 0.4
VCRL	36.5 ± 0.6	26.6 ± 0.2	11.4 ± 0.4	22.6 ± 0.5	24.3 ± 0.4
AdaRFT	36.9 ± 0.3	25.3 ± 0.5	11.5 ± 0.6	22.4 ± 0.3	24.0 ± 0.4
SENT	37.5 ± 0.7	25.8 ± 0.4	9.0 ± 0.3	$\mathbf{28.6 \pm 0.6}$	25.2 ± 0.5
VI-CuRL	$\mathbf{37.8 \pm 0.5}$	$\mathbf{27.8 \pm 0.6}$	$\mathbf{13.2 \pm 0.4}$	28.4 ± 0.5	$\mathbf{26.8 \pm 0.5}$
Verifier-Free: Majority Vote					
TTRL	24.5 ± 0.3	24.3 ± 0.2	1.5 ± 0.2	0.6 ± 0.2	12.7 ± 0.2
VI-CuRL	$\mathbf{35.6 \pm 0.5}$	$\mathbf{26.7 \pm 0.4}$	$\mathbf{9.5 \pm 0.3}$	$\mathbf{21.4 \pm 0.5}$	$\mathbf{23.3 \pm 0.4}$
Verifier-Free: Entropy					
RENT	0.0 ± 0.0	8.3 ± 0.4	0.0 ± 0.0	0.0 ± 0.0	2.1 ± 0.1
EMPO	10.3 ± 0.5	17.5 ± 0.6	0.0 ± 0.0	11.0 ± 0.4	9.7 ± 0.4
VI-CuRL	$\mathbf{35.6 \pm 0.4}$	$\mathbf{25.7 \pm 0.5}$	$\mathbf{9.5 \pm 0.3}$	$\mathbf{21.4 \pm 0.4}$	$\mathbf{23.1 \pm 0.4}$

optimization trajectory even when external rewards are available. The benefits of VI-CuRL become critically pronounced in the Verifier-Free settings (Majority Vote and Entropy). While standard verifier-free methods frequently succumb to high variance and suffer from severe model collapse (often degrading to near 0.0% accuracy on harder datasets), VI-CuRL effectively stabilizes the policy updates and prevents this degradation. By dynamically filtering out high-variance samples in the early stages, VI-CuRL recovers performance comparable to the Oracle topline across all model scales. This phenomenon highlights that managing gradient variance via our annealed confidence curriculum is absolutely essential for stable learning from noisy, intrinsic signals. Additional results for Pass@8 are detailed in Appendix B.

General Knowledge Capabilities. Table 3 extends our evaluation to general knowledge benchmarks using the Llama3.2-3B-Instruct model, assessing whether VI-CuRL generalizes beyond formal mathematical reasoning domains. The results demonstrate that VI-CuRL maintains consistent performance advantages across a diverse array of tasks. Similar to the math domain, VI-CuRL establishes strong performance in the Oracle setting (achieving an average of 26.8% compared to 25.2% for the strongest baseline, SENT). More importantly, in the challenging Verifier-Free settings, VI-CuRL prevents the catastrophic policy collapse universally observed in verifier-free baselines. These findings provide compelling evidence that the variance-bias trade-off curriculum generalizes seamlessly to broad, open-ended applications, effectively stabilizing training and maintaining policy integrity without relying on domain-specific structures or ground-truth verifiers.

5.3 Does VI-CuRL Really Reduce Variance?

A key theoretical advantage of VI-CuRL is variance reduction through curriculum selection. According to Theorem 4.3, the variance of the curriculum-selected estimator depends on:

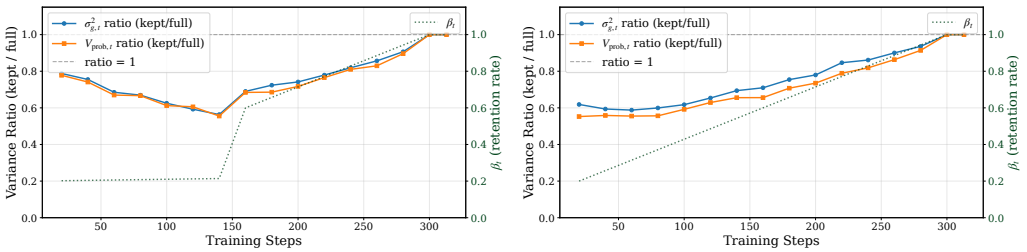
Action Variance: $\sigma^2_{g,t} := \mathbb{E}_{x|w_t=1}[\mathrm{Var}_{y|x}(g)]$: the expected gradient variance due to sampling different rollouts for the *same* prompt within the selected set.

Problem Variance: $V_{\mathrm{prob},t} := \mathrm{Var}_{x|w_t=1}(\bar{g})$: the variance of mean gradients *across* different prompts within the selected set.

To empirically validate this, we estimated these variance components at each training step t using Monte-Carlo approximations. For K independent rollouts per-prompt,

$$\hat{\sigma}^2_{g,t} = \frac{1}{|\mathcal{S}_t|} \sum_{x \in \mathcal{S}_t} \widehat{\mathrm{Var}}_{y|x}(g(x,y)), \quad \hat{V}_{\mathrm{prob},t} = \widehat{\mathrm{Var}}_{x \in \mathcal{S}_t}(\bar{g}(x)),$$

where $\mathcal{S}_t$ denotes either the curriculum-selected subset ("kept") or the full dataset ("full").



(a) Qwen2.5-Math-1.5B (b) DeepSeek-R1-Distill-Qwen-1.5B

Figure 2: **Variance Ratio Analysis.** Variance ratio (kept/full) for Action Variance ($\sigma^2_{g,t}$) and Problem Variance ($V_{\text{prob},t}$) alongside the retention rate (β_t, right axis). When $\beta_t < 1$, both ratios are consistently below 1, confirming that curriculum selection reduces variance as predicted by Theorem 4.3. As $\beta_t \to 1$, the ratios approach 1 since the selected set converges to the full dataset.

Validating Theorem 4.3. The theorem predicted that the curriculum-selected subset $\mathcal{S}_t^{\text{kept}}$ should have lower action/problem variance than the full dataset $\mathcal{S}_t^{\text{full}}$ when $\beta_t < 1$. To test this, we computed

$$\text{Ratio}_{\sigma,t} = \frac{\hat{\sigma}^2_{g,t}(\text{kept})}{\hat{\sigma}^2_{g,t}(\text{full})}, \quad \text{Ratio}_{V_t} = \frac{\hat{V}_{\text{prob},t}(\text{kept})}{\hat{V}_{\text{prob},t}(\text{full})}.$$

If the curriculum effectively reduces variance, we expected $\text{Ratio}_{\sigma,t}, \text{Ratio}_{V_t} < 1$ when $\beta_t < 1$, and $\text{Ratio}_{\sigma,t}, \text{Ratio}_{V_t} \approx 1$ when $\beta_t \to 1$.

Figure 2 presented the variance ratios for both Qwen2.5-Math-1.5B and DeepSeek-R1-Distill-Qwen-1.5B under the Oracle (ground-truth verifier) setting. We can oberve that when $\beta_t \approx 0.2$ (early training), the variance ratios ranged from 0.55 to 0.80, indicating a 20–45% reduction in variance compared to using the full dataset. Meanwhile, as β_t increased towards 1, the variance ratios monotonically increased toward 1, as predicted by the theorem. Besides, action Variance and Problem Variance exhibited similar reduction patterns, confirming that the curriculum simultaneously reduced sampling variance (within prompts) and heterogeneity (across prompts). Therefore, the curriculum selection strategy successfully reduced variance as predicted by Theorem 4.3. Additional analysis of the absolute variance values is provided in Appendix B.

6 Conclusion

In this work, we presented **Verifier-Independent Curriculum Reinforcement Learning (VI-CuRL)**, a theoretically grounded framework designed to stabilize Reinforcement Learning with Verifiable Rewards (RLVR) especially when reliable verifiers are absent. Recognizing that the primary bottleneck in RL is the extreme problem-level variance, VI-CuRL introduces a dynamic, confidence-based curriculum that allows the policy to "start low-uncertainty" by prioritizing high-confidence generations. We formalized this strategy as a principled bias-variance trade-off, proving that our curriculum estimator significantly reduces variance bounds during early training while guaranteeing asymptotic convergence to the true objective. Empirically, VI-CuRL acts as a robust stabilizer for Group Relative Policy Optimization (GRPO), outperforming verifier-based/free baselines across math and general knowledge benchmarks. By effectively filtering out high-variance gradients through intrinsic confidence, VI-CuRL paves the way for scalable, self-improving reasoning models in open-ended domains where ground-truth supervision is scarce. So far VI-CuRL has been verified successfully on reasoning benchmarks, and we plan to extend it to other RL domains in the future.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. We also thank Zikang Shan for his valuable discussions. X-Q C was supported by JSPS KAKENHI Grant Number JP26K21323. MS was supported by JST ASPIRE Grant Number JPMJAP2405.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48, 2009.

Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional pareto-optimal multi-objective reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15593–15613. Curran Associates, Inc., 2023. doi: 10.52202/075280-0686.

Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. Efficient reinforcement learning with semantic and token entropy for llm reasoning. *arXiv preprint arXiv:2512.04359*, 2025.

Haolin Chen, Yihao Feng, Zuxin Liu, Weiran Yao, Akshara Prabhakar, Shelby Heinecke, Ricky Ho, Phil Mui, Silvio Savarese, Caiming Xiong, and Huan Wang. Language models are hidden reasoners: Unlocking latent reasoning capabilities via self-rewarding. *arXiv preprint arXiv:2411.04282*, 2024.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, December 2023.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Thomas Foster, Anya Sims, Johannes Forkel, Mattie Fellows, and Jakob Foerster. Learning to reason at the frontier of learnability, 2026. URL `https://arxiv.org/abs/2502.12272`.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL `http://dx.doi.org/10.1038/s41586-025-09422-z`.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850. Association for Computational Linguistics, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024.

HuggingFaceH4. Aime 2024 (dataset card). Hugging Face, 2024. URL `https://huggingface.co/datasets/HuggingFaceH4/aime_2024`.

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.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Zhiqiang Kou, Junyang Chen, Xin-Qiang Cai, Xiaobo Xia, Ming-Kun Xie, Dong-Dong Wu, Biao Liu, Yuheng Jia, Xin Geng, Masashi Sugiyama, and Tat-Seng Chua. Positive–unlabeled reinforcement learning distillation for on-premise small models. In *Forty-third International Conference on Machine Learning*, 2026.

M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay V. Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Wei Liu, Siya Qi, Xinyu Wang, Chen Qian, Yali Du, and Yulan He. Nover: Incentive training for language models via verifier-free reinforcement learning. *arXiv preprint arXiv:2505.16022*, 2025.

Weiting Liu, Jieyi Bi, Wanqi Zhou, Jianfeng Feng, Yining Ma, Ai Han, and Wenlian Lu. Toolanchor: Anchoring counterfactual context to boost agentic tool-use capability. *arXiv preprint arXiv:2607.14145*, 2026.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. `https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2`, 2025. Notion Blog.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun MA, and Wenhu Chen. General-Reasoner: Advancing LLM reasoning across all domains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

math-ai. Amc 2023 (dataset card). Hugging Face, 2025. URL `https://huggingface.co/datasets/math-ai/amc23`.

OpenCompass. Aime 2025 (dataset card). Hugging Face, 2025. URL `https://huggingface.co/datasets/opencompass/AIME2025`.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, pages 49–64, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025a.

Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. *arXiv preprint arXiv:2504.05520*, 2025b.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Songjun Tu, Chengdong Xu, Qichao Zhang, Yaocheng Zhang, Xiangyuan Lan, Linjing Li, Dong Li, and Dongbin Zhao. Dynamic dual-granularity skill bank for agentic rl. *arXiv preprint arXiv:2603.28716*, 2026.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *International conference on machine learning*, pages 5015–5024, 2018.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, et al. Staying vigil ant: Mitigating visual laziness via counterfactual visual alignment in mllms. In *European Conference on Computer Vision*, pages 287–305. Springer, 2026a.

Xi Xiao, Tianchen Zhao, Youngeun Kim, Zhuowei Li, Linghan Xu, Jiaye Wu, Zheng Zhang, Xiang Xu, Xuanbai Chen, Farhan Tejani, et al. Rethinking latent visual reasoning: Grounding latent reasoning in visual evidence. *arXiv preprint arXiv:2609.34563*, 2026b.

Bangji Yang, Hongbo Ma, Jiajun Fan, and Ge Liu. Batched contextual reinforcement. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8Oc3Mx754M.

Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, and Tong Zhang. Optimizing chain-of-thought reasoners via gradient variance minimization in rejection sampling and rl. *arXiv preprint arXiv:2505.02391*, 2025.

Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, and Tong Zhang. Optimizing chain-of-thought reasoners via gradient variance minimization in rejection sampling and rl. *Advances in Neural Information Processing Systems*, 38:163245–163284, 2026.

Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, and Chengchun Shi. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*, 2025a.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*, 2025b.

Qinjian Zhao, Zhihao Dou, Dinggen Zhang, Xiangyu Li, Chaoda Song, Zhongwei Wan, Xinpeng Li, Yanyan Zhang, Kaijie Chen, Qingtao Pan, et al. Stride: Strategic trajectory reasoning via discriminative estimation for verifiable reinforcement learning. *arXiv preprint arXiv:2606.15866*, 2026.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*, 2025.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A Related Work

LLM Reasoning. Recent advances in LLM reasoning have been significantly propelled by RL [Schulman et al., 2017, Cai et al., 2023], especially for the development of LLM-based agents [Tu et al., 2026, Liu et al., 2026, Yang et al., 2026]. Early iterative approaches like Self-Taught Reasoner (STaR) [Zelikman et al., 2022] demonstrated that models could self-improve by bootstrapping from their own verified correct solutions. Shao et al. [2024] introduced GRPO to achieve strong mathematical reasoning with improved sample efficiency, while DeepSeek-R1 [Guo et al., 2025] demonstrated that sophisticated behaviors can emerge purely from RL without supervised fine-tuning. The RLVR paradigm [Wen et al., 2025, Zhao et al., 2026] formalized this approach. Parallel work on Process Reward Models (PRMs) [Lightman et al., 2024] has explored dense, step-level supervision to guide reasoning, though this often requires expensive human annotation. Recent verifier-free approaches have also gained traction: Reinforcement Learning via Entropy Minimization (RENT) [Prabhudesai et al., 2025] and Entropy Minimization Policy Optimization (EMPO) [Zhang et al., 2025a] leverage entropy minimization as a fully unsupervised reward signal, while the Entropy Mechanism [Cui et al., 2025] stabilizes training by managing policy collapse. Test-Time Reinforcement Learning (TTRL) [Zuo et al., 2025] further explores converting test-time compute (e.g., majority voting) into training signals. However, most of these methods fundamentally rely on access to ground-truth verifiers or correctness labels, while verifier-free methods tend to be unstable and fragile during training. VI-CuRL addresses the complementary data-selection problem: it uses intrinsic confidence to rank prompts, while the chosen oracle or intrinsic reward still supplies the optimization signal.

Bias-Variance Trade-off in Policy Gradients. The high variance of policy gradient estimators has long been recognized as a central challenge in RL [Williams, 1992]. Sutton et al. [1999] established the theoretical foundations for variance reduction through baselines. Modern LLM reasoning pipelines typically employ Group Relative Policy Optimization (GRPO) [Shao et al., 2024], which estimates a baseline from the mean reward of a group of sampled outputs. A group-relative baseline and prompt selection act on different aspects of the update; changing the optimizer does not eliminate the need to evaluate the prompt-distribution intervention separately. In verifier-independent settings, the problem-level variance can be extreme due to the lack of accurate reward signals, necessitating a curriculum approach like VI-CuRL to curate the training distribution itself.

Curriculum Learning. Curriculum learning, first formalized by Bengio et al. [2009], proposes training models on progressively harder examples to improve convergence and generalization. Self-paced learning [Kumar et al., 2010] extends this by allowing the model itself to determine example difficulty based on training loss, with self-paced curriculum learning [Jiang et al., 2015] unifying both paradigms. In the context of RL for LLMs, AdaRFT [Shi et al., 2025b] dynamically adjusts problem difficulty based on recent reward signals, while VCRL [Jiang et al., 2025] uses reward variance to identify moderately difficult samples for training. However, both AdaRFT and VCRL fundamentally require ground-truth reward signals from external verifiers to compute their curriculum metrics. Our VI-CuRL is verifier-independent: it uses the model's confidence as an intrinsic proxy for variance control, enabling curriculum-based training in domains where verifiers are unavailable. SENT [Cao et al., 2025] organizes training data with a semantic-entropy curriculum computed offline from the initial policy. However, the curriculum ranking induced by the initial policy may become mismatched after RL updates shift the policy distribution. In contrast, VI-CuRL computes prompt-level predictive confidence online from the current behavior policy and anneals an importance-weighted selection mask, allowing the curriculum to adaptively trade transient bias for variance reduction during optimization. LILO [Foster et al., 2026] prioritizes learnability through binary-success variance $p(x)(1-p(x))$. That outcome statistic differs from the covariance of vector surrogate gradients analyzed here. A prompt can offer a nonzero learning signal while its rollout gradients remain directionally incoherent. Likewise, high-entropy minority tokens [Wang et al., 2025] locate informative decisions within a trajectory; VI-CuRL ranks prompts using average uncertainty and retains all valid response tokens by default.

Table 4: Comparison with **Oracle (w. Verifier)** reward. Mean and standard deviation **(pass@8)** with 16 samples and 5 random seeds. We compare VI-CuRL against Curriculum baselines (VCRL, AdaRFT, SENT) and No Curriculum. Blue rows highlight our method.

Dataset	AIME 2024	AIME 2025	AMC 2023	Math500	Minerva MATH	Olympiad Bench	Average
Qwen2.5-Math-1.5B							
Base (No RL)	32.4 ± 0.5	16.7 ± 0.7	79.3 ± 1.4	62.8 ± 1.3	15.5 ± 1.1	31.3 ± 0.8	39.7 ± 1.0
Oracle (w. Verifier)							
No Curriculum	34.3 ± 1.2	19.9 ± 1.1	80.7 ± 1.4	69.6 ± 0.5	18.0 ± 0.6	32.8 ± 1.2	42.6 ± 1.0
VCRL	32.2 ± 0.5	19.8 ± 1.0	79.0 ± 0.9	70.2 ± 1.1	20.4 ± 1.0	34.0 ± 1.1	42.6 ± 0.9
AdaRFT	32.1 ± 0.4	17.9 ± 1.5	77.5 ± 1.0	**72.4 ± 0.9**	21.7 ± 0.9	34.4 ± 0.5	42.7 ± 0.9
SENT	35.0 ± 0.4	**26.7 ± 0.4**	**82.5 ± 0.7**	71.4 ± 0.7	15.3 ± 1.4	33.2 ± 0.9	44.0 ± 0.8
VI-CuRL	**35.2 ± 1.4**	24.6 ± 0.2	**82.5 ± 1.2**	71.3 ± 1.8	**26.9 ± 1.3**	**37.9 ± 4.0**	**46.4 ± 1.7**
DeepSeek-R1-Distill-Qwen-1.5B							
Base (No RL)	44.8 ± 0.8	32.8 ± 0.7	88.2 ± 0.1	85.9 ± 0.6	27.8 ± 0.3	48.6 ± 0.3	54.7 ± 0.5
Oracle (w. Verifier)							
No Curriculum	55.1 ± 0.6	40.1 ± 0.4	96.1 ± 0.2	96.6 ± 0.1	44.6 ± 0.1	64.9 ± 0.1	66.2 ± 0.3
VCRL	55.0 ± 0.9	42.1 ± 0.9	96.8 ± 1.1	97.5 ± 0.5	42.3 ± 1.7	66.9 ± 3.5	66.8 ± 1.4
AdaRFT	57.4 ± 1.3	43.2 ± 1.9	97.9 ± 0.3	98.3 ± 0.7	46.9 ± 1.0	70.0 ± 1.2	69.0 ± 1.1
SENT	57.1 ± 0.7	39.9 ± 1.5	97.1 ± 0.9	98.1 ± 0.7	49.6 ± 1.3	71.4 ± 0.8	68.9 ± 1.0
VI-CuRL	**60.9 ± 2.3**	**45.5 ± 2.6**	**98.9 ± 1.3**	**98.8 ± 0.4**	**51.3 ± 1.3**	**72.8 ± 1.5**	**71.4 ± 1.6**
Llama3.2-3B-Instruct							
Base (No RL)	18.8 ± 1.1	3.3 ± 0.2	54.7 ± 1.4	40.8 ± 0.8	5.2 ± 1.1	16.4 ± 0.5	23.2 ± 0.9
Oracle (w. Verifier)							
No Curriculum	22.4 ± 1.0	6.7 ± 1.4	60.0 ± 1.3	44.8 ± 0.6	13.3 ± 0.7	18.7 ± 1.5	27.7 ± 1.1
VCRL	23.5 ± 0.4	8.2 ± 1.1	61.4 ± 1.3	45.9 ± 0.9	15.3 ± 0.7	17.5 ± 0.5	28.6 ± 0.8
AdaRFT	24.9 ± 1.8	8.0 ± 0.7	**67.9 ± 1.4**	45.6 ± 0.2	16.3 ± 1.9	18.1 ± 1.6	30.1 ± 1.3
SENT	25.1 ± 0.8	8.1 ± 0.7	62.2 ± 1.0	48.0 ± 1.5	14.6 ± 0.4	20.1 ± 1.0	29.7 ± 0.9
VI-CuRL	**26.1 ± 0.8**	**9.3 ± 0.7**	63.1 ± 1.4	**49.6 ± 1.4**	**17.1 ± 0.8**	**22.9 ± 1.2**	**31.4 ± 1.1**
Qwen2.5-Math-7B							
Base (No RL)	36.2 ± 1.2	23.3 ± 1.0	81.9 ± 1.3	66.6 ± 0.5	11.8 ± 0.6	33.9 ± 0.7	42.3 ± 0.9
Oracle (w. Verifier)							
No Curriculum	50.0 ± 0.1	28.7 ± 0.8	83.4 ± 0.6	82.8 ± 0.5	29.0 ± 1.1	45.8 ± 1.3	53.3 ± 0.7
VCRL	48.3 ± 1.4	23.3 ± 0.5	84.7 ± 0.6	79.8 ± 1.5	24.3 ± 0.4	41.0 ± 0.2	50.2 ± 0.8
AdaRFT	50.0 ± 3.1	27.9 ± 3.0	85.0 ± 1.5	78.0 ± 2.0	22.8 ± 6.6	44.5 ± 1.1	51.4 ± 2.9
SENT	55.7 ± 1.9	17.5 ± 0.7	64.7 ± 0.8	**85.8 ± 0.2**	35.7 ± 0.7	50.2 ± 0.8	51.6 ± 0.9
VI-CuRL	**56.4 ± 1.3**	**30.8 ± 0.2**	**88.8 ± 0.7**	82.6 ± 0.4	**37.1 ± 0.9**	**51.4 ± 1.2**	**57.9 ± 0.8**

B Algorithm Details and Additional Experimental Results

B.1 Learning Curve Analysis: with/without Verifiers

We performed a comprehensive comparison of learning curves across three distinct reward settings: **Oracle** (Ground Truth), **Majority Vote**, and **Entropy**. Figure 3 presented a 6×6 grid of results.

Oracle Setting (Rows 1-2): Even with a perfect verifier, VI-CuRL (blue) often matched or exceeded standard RL (orange) in stability and final performance.

Majority Vote (Rows 3-4): This verifier-independent baseline (purple) was notoriously unstable, as seen in the fluctuating curves. VI-CuRL significantly stabilized training, yielding smooth, and consistent improvements.

Entropy (Rows 5-6): Using entropy as a proxy also benefited from VI-CuRL's curriculum, which filtered out high-entropy (uncertain) samples early on to prevent policy collapse.

VI-CuRL demonstrated effective variance reduction and robust learning across all settings, validating its utility as a general-purpose stabilizer for RLVR.

Table 5: Comparison in **Verifier-Free** settings. Mean and standard deviation **(pass@8)** with 16 samples and 5 random seeds. We compare VI-CuRL against baselines using **Majority Vote** and **Entropy** as intrinsic reward signals. Note that VCRL, AdaRFT, and SENT are excluded as they require ground-truth verifiers or offline initial policy distribution. Blue rows highlight our method.

Dataset	AIME 2024	AIME 2025	AMC 2023	Math500	Minerva MATH	Olympiad Bench	Average
Qwen2.5-Math-1.5B							
Base (No RL)	32.4 ± 0.5	16.7 ± 0.7	79.3 ± 1.4	62.8 ± 1.3	15.5 ± 1.1	31.3 ± 0.8	39.7 ± 1.0
Majority Vote (w/o. Verifier)							
TTRL	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.6 ± 0.2	1.1 ± 0.4	0.0 ± 0.0	0.3 ± 0.1
VI-CuRL	**35.0 ± 0.2**	**26.1 ± 2.0**	**81.2 ± 0.6**	**73.9 ± 1.3**	**22.1 ± 4.6**	**37.1 ± 1.0**	**45.9 ± 1.6**
Entropy (w/o. Verifier)							
RENT	0.0 ± 0.0	0.0 ± 0.0	12.9 ± 0.7	11.7 ± 1.3	8.6 ± 0.9	5.7 ± 0.8	6.5 ± 0.6
EMPO	12.5 ± 1.5	8.2 ± 1.1	25.4 ± 2.5	28.6 ± 2.1	10.2 ± 1.8	14.5 ± 1.5	16.6 ± 1.8
VI-CuRL	**33.1 ± 0.1**	**21.4 ± 1.7**	**81.7 ± 0.5**	**71.7 ± 0.4**	**26.2 ± 1.0**	**35.5 ± 0.8**	**44.9 ± 0.8**
DeepSeek-R1-Distill-Qwen-1.5B							
Base (No RL)	44.8 ± 0.8	32.8 ± 0.7	88.2 ± 0.1	85.9 ± 0.6	27.8 ± 0.3	48.6 ± 0.3	54.7 ± 0.5
Majority Vote (w/o. Verifier)							
TTRL	38.7 ± 0.6	30.7 ± 0.9	90.8 ± 0.8	91.9 ± 0.2	36.9 ± 0.9	52.5 ± 1.7	56.9 ± 0.9
VI-CuRL	**62.8 ± 0.8**	**50.5 ± 0.8**	**99.3 ± 0.5**	**98.4 ± 1.1**	**47.6 ± 0.7**	**71.9 ± 1.5**	**71.8 ± 0.9**
Entropy (w/o. Verifier)							
RENT	38.0 ± 1.0	25.1 ± 0.2	90.5 ± 0.8	95.0 ± 0.8	43.6 ± 1.8	59.9 ± 1.1	58.7 ± 1.0
EMPO	48.6 ± 2.5	35.1 ± 2.2	91.7 ± 1.5	96.4 ± 0.4	44.0 ± 1.6	57.4 ± 1.1	62.2 ± 1.6
VI-CuRL	**62.1 ± 1.1**	**47.3 ± 1.0**	**97.6 ± 0.2**	**98.8 ± 0.1**	**45.8 ± 0.3**	**70.4 ± 1.0**	**70.3 ± 0.6**
Llama3.2-3B-Instruct							
Base (No RL)	18.8 ± 1.1	3.3 ± 0.2	54.7 ± 1.4	40.8 ± 0.8	5.2 ± 1.1	16.4 ± 0.5	23.2 ± 0.9
Majority Vote (w/o. Verifier)							
TTRL	**23.3 ± 0.3**	3.3 ± 0.7	53.8 ± 0.5	50.6 ± 0.8	10.7 ± 1.0	**24.0 ± 1.3**	27.6 ± 0.8
VI-CuRL	20.4 ± 0.8	**4.6 ± 0.6**	**58.1 ± 0.9**	**54.2 ± 0.7**	**14.0 ± 0.8**	23.3 ± 1.4	**29.1 ± 0.9**
Entropy (w/o. Verifier)							
RENT	2.4 ± 0.9	0.0 ± 0.0	15.9 ± 1.2	5.4 ± 1.6	4.2 ± 0.5	1.5 ± 0.6	4.9 ± 0.8
EMPO	8.5 ± 1.2	1.1 ± 0.3	25.4 ± 2.0	18.5 ± 1.5	6.5 ± 1.1	8.4 ± 1.0	11.4 ± 1.2
VI-CuRL	**20.8 ± 0.9**	**5.2 ± 1.0**	**56.9 ± 1.3**	**50.2 ± 0.1**	**12.9 ± 0.7**	**21.8 ± 0.7**	**28.0 ± 0.8**
Qwen2.5-Math-7B							
Base (No RL)	36.2 ± 1.2	23.3 ± 1.0	81.9 ± 1.3	66.6 ± 0.5	11.8 ± 0.6	33.9 ± 0.7	42.3 ± 0.9
Majority Vote (w/o. Verifier)							
TTRL	53.8 ± 0.7	28.7 ± 0.9	82.8 ± 1.3	81.4 ± 0.1	23.2 ± 0.6	41.9 ± 0.7	52.0 ± 0.7
VI-CuRL	**55.5 ± 2.1**	**31.7 ± 0.1**	**88.2 ± 1.0**	**81.6 ± 0.8**	**27.9 ± 0.6**	**49.1 ± 3.7**	**55.7 ± 1.4**
Entropy (w/o. Verifier)							
RENT	28.8 ± 1.4	17.9 ± 2.5	71.5 ± 3.2	74.6 ± 2.3	**25.4 ± 5.0**	35.2 ± 4.2	42.2 ± 3.1
EMPO	35.2 ± 2.5	22.4 ± 1.8	76.5 ± 2.2	78.2 ± 1.5	22.5 ± 3.1	39.5 ± 2.0	45.7 ± 2.2
VI-CuRL	**48.4 ± 3.9**	**30.0 ± 1.4**	**85.0 ± 0.3**	**80.6 ± 0.3**	24.6 ± 0.7	**43.4 ± 0.8**	**52.0 ± 1.2**

B.2 Pass@8 Performance

While Pass@1 roughly measures the mode of the policy distribution (the ability to generate the best answer), Pass@8 evaluates the breadth and coverage of the model's reasoning capabilities, measuring the probability that at least one solution is correct across 8 independent samples. Table 4 and Table 5 present the detailed Pass@8 results for the Oracle and Verifier-Free settings, respectively. The results closely mirror the trends observed in the Pass@1 analysis: VI-CuRL improves coverage across almost all benchmarks. Crucially, in the verifier-independent setting, where baselines often collapse to repetitive, incorrect outputs (leading to low Pass@8), VI-CuRL maintains high diversity and solvability. This consistency between Pass@1 and Pass@8 confirms that our method improves the fundamental quality of the policy without trading off diversity or overfitting to narrow solution paths.

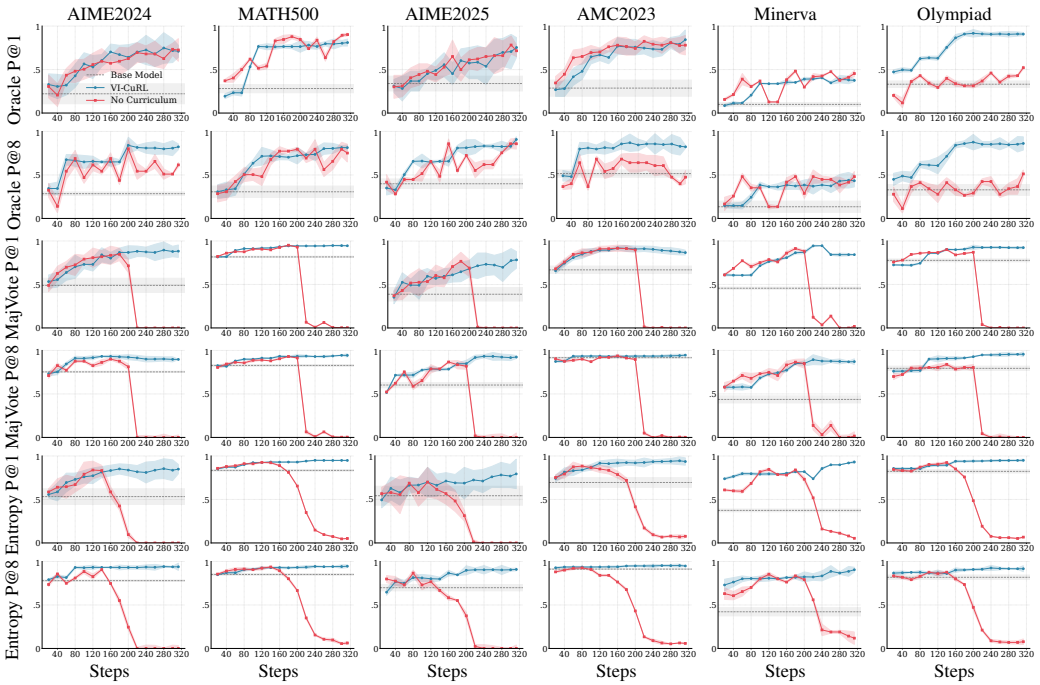


Figure 3: The learning curves comparing VI-CuRL against the baseline No Curriculum across verifier-based (Oracle) and verifier-free (Majority Vote and Entropy) settings. Rows alternate between Pass@1 and Pass@8 for each setting. For each subplot, the y-axis is normalized to $[0, 1]$.

B.3 Curriculum Control Ablation

To validate that VI-CuRL's gains stem from its specific annealed confidence curriculum rather than from any form of data filtering, we conduct a systematic ablation study on DeepSeek-R1-Distill-Qwen-1.5B under verifier-free RENT reward. Table 6 compares VI-CuRL against several alternative filtering strategies: (1) **Length-First Filter**, which retains the shortest responses; (2) **Random Filter**, which randomly retains prompts at the same rate β_t; and (3) **Reverse-Confidence Filter**, which retains the *lowest*-confidence prompts (the opposite of VI-CuRL).

Table 6: Curriculum-specific control ablation for DeepSeek-R1-Distill-Qwen-1.5B under verifier-free GRPO with RENT reward. Each cell reports Pass@1 / Pass@8. Blue rows highlight our method.

Setting	AIME 2024	AIME 2025	AMC 2023	Math500	Minerva MATH	Olympiad Bench	Average
RENT / No Curriculum	25.3 / 38.0	23.2 / 25.1	71.1 / 90.5	85.3 / 95.0	31.5 / 43.6	48.0 / 59.9	47.4 / 58.7
RENT / Length-First Filter	29.6 / 45.4	23.6 / 30.5	77.0 / 93.6	88.7 / 97.1	30.1 / 42.8	51.1 / 63.1	50.0 / 62.1
RENT / Random Filter	25.5 / 38.5	22.6 / 25.1	71.9 / 90.8	85.2 / 95.4	31.4 / 44.0	48.2 / 60.1	47.5 / 59.0
RENT / Reverse-Confidence	24.9 / 37.5	21.6 / 25.6	70.1 / 90.0	85.1 / 95.0	29.9 / 43.3	47.5 / 59.4	46.5 / 58.5
VI-CuRL / RENT	**38.0 / 62.1**	**31.0 / 47.3**	**83.8 / 97.6**	**92.8 / 98.8**	**32.8 / 45.8**	**57.8 / 70.4**	**56.0 / 70.3**

B.4 Long-Output Ablation

A potential concern is that truncation artifacts (when responses hit the decoding cap) may affect the confidence signal. Table 7 compares `max_new_tokens`$= 3072$ (our default) with 8192.

B.5 Mechanism Analysis: Why Annealing Matters

Table 8 examines why the annealed curriculum outperforms static alternatives. We compare four strategies with 16 rollout samples: (1) RENT / No Curriculum, (2) Static Confidence Filter (fixed $\beta = 0.2$), (3) Soft Confidence Weighting (using confidence as sample weights without hard filtering), and (4) VI-CuRL's annealed schedule. The key metric is *Upgration* (change from reference step

Table 7: Long-output ablation for DeepSeek-R1-Distill-Qwen-1.5B (RENT reward).

Setting	max_tokens	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg
RENT	3072	25.3	23.2	71.1	85.3	31.5	48.0	47.4
RENT	8192	25.8	23.2	71.9	86.1	31.5	48.2	47.8
VI-CuRL / RENT	3072	38.0	31.0	83.8	92.8	32.8	57.8	56.0
VI-CuRL / RENT	8192	**38.9**	**31.6**	**84.2**	**93.4**	**33.5**	**58.6**	**56.7**

Setting	max_tokens	Cap-hit	No-boxed	Mean len	P90 len
RENT	3072	3.9%	10.0%	1895	3004
RENT	8192	1.7%	7.9%	2136	4251
VI-CuRL / RENT	3072	2.9%	8.1%	1056	2805
VI-CuRL / RENT	8192	1.2%	3.9%	2311	4511

T_{ref} to long-horizon T_{long}): Static filtering achieves high initial performance (30.2) but collapses long-term (-22.4), while Soft Weighting decays less severely (-10.2) but still degrades. Only VI-CuRL maintains stability ($+0.8$) with the highest AULC (13.0 vs. 1.0 baseline). Table 9 provides the per-dataset breakdown, confirming that VI-CuRL's advantage is consistent across all benchmarks. Besides, some other reasoning works studies highlight the importance of grounding predictions in visual evidence [Xiao et al., 2026b,a]. Such signals could complement confidence-based selection when extending VI-CuRL beyond text.

Table 8: Mechanism analysis with 16 rollout samples. T_{ref}: step 313; T_{long}: step 626. AULC is normalized with RENT/No Curriculum $= 1.0$. Blue rows highlight our method.

Setting	**Avg P@1 @ T_{ref}**	**Avg P@1 @ T_{long}**	**Best Avg P@1**	**Upgration**	**AULC**
RENT / No Curriculum	4.8	2.6	32.8	-2.2	1.0
RENT / Static Conf. Filter	30.2	7.8	30.2	-22.4	3.3
RENT / Soft Conf. Weighting	13.8	3.6	30.5	-10.2	1.9
VI-CuRL / RENT	**31.7**	**32.5**	**36.7**	**+0.8**	**13.0**

Table 9: Per-dataset Pass@1 breakdown at T_{long} (step 626). Blue rows highlight our method.

Setting	**AIME2024**	**AIME2025**	**AMC2023**	**MATH500**	**Minerva**	**OlympiadBench**
RENT / No Curriculum	0.0	0.0	5.1	4.0	4.2	2.5
RENT / Static Conf. Filter	0.3	1.2	22.6	8.6	9.5	4.3
RENT / Soft Conf. Weighting	0.0	0.0	7.8	7.2	4.1	2.6
VI-CuRL / RENT	**14.3**	**5.4**	**51.3**	**70.4**	**20.5**	**33.1**

C Proofs and Additional Analysis

C.1 Proof of Theorem 4.1

Proof. The goal is to bound the difference between the true surrogate objective $\mathcal{L}(\theta)$ and the VI-CuRL surrogate $\mathcal{L}_t(\theta)$. Recall the definitions:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim p(x)}\mathbb{E}_{y_{1:G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}[\ell(\theta; x, y_{1:G})]$$

$$\mathcal{L}_t(\theta) = \mathbb{E}_{x\sim p(x)}\mathbb{E}_{y_{1:G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{w_t(x)}{\beta_t}\ell(\theta; x, y_{1:G})\right]$$

We partitions the objective using the law of total expectation conditioned on the mask $w_t(x)$. Let $\mathcal{S}_t = \{x \mid w_t(x) = 1\}$ be the selected subspace.

$$\begin{aligned}\mathcal{L}(\theta) &= \int \ell(\theta; x, y_{1:G})p(x)\pi_{\theta_{\text{old}}}(y_{1:G}|x)dy_{1:G}dx \\ &= \beta_t\mathcal{L}_t(\theta) + (1-\beta_t)\mathcal{L}_{\text{discard}}(\theta).\end{aligned}$$

Thus, we have the exact decomposition:

$$\mathcal{L}(\theta) = \beta_t \mathcal{L}_t(\theta) + (1 - \beta_t)\mathcal{L}_{\text{discard}}(\theta).$$

Now, we examine the difference:

$$\mathcal{L}(\theta) - \mathcal{L}_t(\theta) = (1 - \beta_t)(\mathcal{L}_{\text{discard}}(\theta) - \mathcal{L}_t(\theta)).$$

Assuming the surrogate loss is bounded $|\ell(\theta; x, y_{1:G})| \leq L_{\max}$, then:

$$\begin{aligned} |\mathcal{L}(\theta) - \mathcal{L}_t(\theta)| &\leq (1 - \beta_t)(L_{\max} + L_{\max}) \\ &= 2L_{\max}(1 - \beta_t). \end{aligned}$$

This completes the proof. □

C.2 Proof of Theorem 4.3

Proof. Let $\hat{g}_t$ be the VI-CuRL gradient estimator:

$$\hat{g}_t = \frac{w_t(x)}{\beta_t} g(x, y_{1:G}; \theta),$$

where $g(x, y_{1:G}; \theta) = \nabla_\theta \ell(\theta; x, y_{1:G})$ is the gradient of the surrogate loss. We use the Law of Total Variance, conditioning on the prompt x:

$$\text{Var}(\hat{g}_t) = \mathbb{E}_x[\text{Var}_{\mathbf{y}}(\hat{g}_t|x)] + \text{Var}_x(\mathbb{E}_{\mathbf{y}}[\hat{g}_t|x]).$$

We analyze these two terms separately.

1. The Expected Action Variance Term: Conditioned on a specific prompt x, the mask $w_t(x)$ and scalar β_t are constants. Thus:

$$\begin{aligned} \text{Var}_{y_{1:G}}(\hat{g}_t|x) &= \text{Var}_{y_{1:G}}\left(\frac{w_t(x)}{\beta_t} g(x, y_{1:G}; \theta)\middle| x\right) \\ &= \left(\frac{w_t(x)}{\beta_t}\right)^2 \text{Var}_{y_{1:G}}(g(x, y_{1:G}; \theta)|x), \\ &\qquad \text{(quadratic scaling of covariance trace)} \end{aligned}$$

Taking the expectation over x:

$$\mathbb{E}_x[\text{Var}_{y_{1:G}}(\hat{g}_t|x)] = \mathbb{E}_x\left[\frac{w_t(x)^2}{\beta_t^2} \text{Var}_{y_{1:G}}(g(x, y_{1:G}; \theta)|x)\right].$$

Since $w_t(x)$ is an indicator, $w_t(x)^2 = w_t(x)$.

$$\begin{aligned} \mathbb{E}_x[\text{Var}_{y_{1:G}}(\hat{g}_t|x)] &= \frac{1}{\beta_t^2}\mathbb{E}_x\left[w_t(x) \text{Var}_{y_{1:G}}(g(x, y_{1:G}; \theta)|x)\right] \\ &= \frac{1}{\beta_t} \cdot \frac{\mathbb{E}_x[w_t(x) \text{Var}_{y_{1:G}}(g(x, y_{1:G}; \theta)|x)]}{\beta_t} \\ &= \frac{1}{\beta_t}\mathbb{E}_{x|w_t(x)=1}[\text{Var}_{y_{1:G}}(g(x, y_{1:G}; \theta)|x)]. \end{aligned}$$

We define $\sigma_{g,t}^2 \triangleq \mathbb{E}_{x|w_t(x)=1}[\text{Var}_{y_{1:G}}(g|x)]$ as the average **Action Variance**. Thus, the first term is $\frac{\sigma_{g,t}^2}{\beta_t}$.

2. The Problem Variance Term: First, consider the inner expectation $\mathbb{E}_{y_{1:G}}[\hat{g}_t|x]$. Let $\bar{g}(x) = \mathbb{E}_{y_{1:G}|\pi_{\theta_{\text{old}}}}[g|x]$ be the expected gradient for prompt x. Then:

$$\mathbb{E}_{y_{1:G}}[\hat{g}_t|x] = \frac{w_t(x)}{\beta_t}\bar{g}(x).$$

The variance of this quantity over x is:

$$\text{Var}_x\left(\frac{w_t(x)}{\beta_t}\bar{g}(x)\right) = \mathbb{E}_x\left[\left\|\frac{w_t(x)}{\beta_t}\bar{g}(x)\right\|^2\right] - \left\|\mathbb{E}_x\left[\frac{w_t(x)}{\beta_t}\bar{g}(x)\right]\right\|^2.$$

The second part is simply the squared norm of the true gradient of the surrogate objective $\nabla_\theta \mathcal{L}_t(\theta)$. Expanding the first part and using the definition of $V_{\text{prob},t} = \text{Var}_{x|w_t(x)=1}(\bar{g}(x))$:

$$\text{Var}_x(\mathbb{E}_{y_{1:G}}[\hat{g}_t|x]) = \frac{V_{\text{prob},t}}{\beta_t} + \frac{1-\beta_t}{\beta_t}\|\nabla_\theta \mathcal{L}_t(\theta)\|^2.$$

Combining the Action Variance term and the Problem Variance term yields the final decomposition:

$$\text{Var}(\hat{g}_t) = \frac{\sigma_{g,t}^2}{\beta_t} + \frac{V_{\text{prob},t}}{\beta_t} + \frac{1-\beta_t}{\beta_t}\|\nabla_\theta \mathcal{L}_t(\theta)\|^2.$$

□

C.3 Proof of Lemma C.1

Lemma C.1 (Conditional Entropy-Envelope Bound)**.** *Let $h(x) = 1 - c(x)$ and $a_t = 1 - \tau_t$ be normalized entropy and its selection cutoff. Suppose there exists a non-decreasing function $G_{\text{conf}} : [0,1] \to [0,\infty)$ such that for all selected prompts ($h(x) \le a_t$), the gradient norm obeys the following entropy envelope:*

$$\|\nabla_\theta \ell(\theta; x, y_{1:G})\| \le G_{\text{conf}}(h(x)). \tag{17}$$

Then

$$\text{Var}(\hat{g}_t) \le \frac{G_{\text{conf}}(a_t)^2}{\beta_t} - \|\nabla_\theta \mathcal{L}_t(\theta)\|^2. \tag{18}$$

This bound is conditional on the envelope assumption. It does not imply that the numerator compensates for inverse retention.

Proof. We derive a bound for the second moment $\mathbb{E}[\|\hat{g}_t\|^2]$ under the entropy-envelope assumption.

$$\hat{g}_t = \frac{w_t(x)}{\beta_t}\nabla_\theta \ell(\theta; x, y_{1:G}).$$

Taking the squared norm and expectation:

$$\begin{aligned}\mathbb{E}[\|\hat{g}_t\|^2] &= \mathbb{E}_{x,y_{1:G}}\left[\left\|\frac{w_t(x)}{\beta_t}\nabla_\theta \ell\right\|^2\right] \\ &= \frac{1}{\beta_t^2}\mathbb{E}_x\left[w_t(x)\mathbb{E}_{y_{1:G}|x}[\|\nabla_\theta \ell(\theta; x, y_{1:G})\|^2]\right].\end{aligned}$$

Using the assumption $\|\nabla_\theta \ell\| \le G_{\text{conf}}(h(x))$ applied to the prompt x:

$$\mathbb{E}_{y_{1:G}|x}[\|\nabla_\theta \ell\|^2] \le G_{\text{conf}}(h(x))^2.$$

Since $w_t(x) = 1$ implies $h(x) \le a_t$ and G_{conf} is non-decreasing, $G_{\text{conf}}(h(x)) \le G_{\text{conf}}(a_t)$.

$$\begin{aligned}\mathbb{E}[\|\hat{g}_t\|^2] &\le \frac{1}{\beta_t^2}\mathbb{E}_x[w_t(x)G_{\text{conf}}(a_t)^2] \\ &= \frac{G_{\text{conf}}(a_t)^2}{\beta_t^2}\mathbb{E}_x[w_t(x)] \\ &= \frac{G_{\text{conf}}(a_t)^2}{\beta_t}.\end{aligned}$$

Finally, using $\text{Var}(\hat{g}_t) = \mathbb{E}[\|\hat{g}_t\|^2] - \|\mathbb{E}[\hat{g}_t]\|^2$, we obtain:

$$\text{Var}(\hat{g}_t) \le \frac{G_{\text{conf}}(a_t)^2}{\beta_t} - \|\nabla_\theta \mathcal{L}_t(\theta)\|^2.$$

□

C.4 Proof of Theorem C.2

Theorem C.2 (Conditional Curriculum-Sensitive Variance Bound)**.** *Assume that for each phase t there exist constants $a_1, a_2 \geq 0$ and $\alpha > 0$ such that the variance numerators from Theorem 4.3 decay faster than the retention rate:*

$$\sigma_{g,t}^2 \leq a_1\, \beta_t^{1+\alpha}, \quad \textit{and} \quad V_{\mathrm{prob},t} \leq a_2\, \beta_t^{1+\alpha}. \tag{19}$$

Then the total variance is bounded by

$$\mathrm{Var}(\hat{g}_t) \leq (a_1 + a_2)\, \beta_t^{\alpha} + \frac{1-\beta_t}{\beta_t} \left\|\nabla_\theta \mathcal{L}_t(\theta)\right\|^2. \tag{20}$$

Assumption (19) is an additional sufficient condition, not a consequence of confidence ranking. The measured component ratios alone do not verify its proposed retention scaling. The masking contribution still depends on the selected mean gradient and may grow under aggressive retention; the bound does not establish universal variance reduction.

Proof. This proof follows directly from substituting the assumption bounds into the exact decomposition derived in Theorem 4.3. Recall:

$$\mathrm{Var}(\hat{g}_t) = \frac{\sigma_{g,t}^2}{\beta_t} + \frac{V_{\mathrm{prob},t}}{\beta_t} + \frac{1-\beta_t}{\beta_t} \|\nabla_\theta \mathcal{L}_t(\theta)\|^2.$$

Assumption (19) posits that the curriculum effectively selects easier samples such that:

$$\sigma_{g,t}^2 \leq a_1 \beta_t^{1+\alpha} \quad \text{and} \quad V_{\mathrm{prob},t} \leq a_2 \beta_t^{1+\alpha},$$

for some $\alpha > 0$. Substituting these inequalities:

$$\frac{\sigma_{g,t}^2}{\beta_t} \leq \frac{a_1 \beta_t^{1+\alpha}}{\beta_t} = a_1 \beta_t^{\alpha}$$

$$\frac{V_{\mathrm{prob},t}}{\beta_t} \leq \frac{a_2 \beta_t^{1+\alpha}}{\beta_t} = a_2 \beta_t^{\alpha}.$$

Summing these terms:

$$\mathrm{Var}(\hat{g}_t) \leq (a_1 + a_2)\beta_t^{\alpha} + \frac{1-\beta_t}{\beta_t} \|\nabla_\theta \mathcal{L}_t(\theta)\|^2.$$

As $\beta_t \to 0$, only the first contribution is guaranteed to vanish under the assumption. Bounded total variance additionally requires control of $\|\nabla_\theta \mathcal{L}_t\|^2/\beta_t$; the practical schedule starts at positive retention. □

D Implementation Details

This section details the prompt configurations, hyperparameters, and datasets used to ensure the reproducibility of our experiments. We adopt a structured presentation to distinguish experimental components clearly.

D.1 Prompt Templates

Training & Generation. We employ a consistent prompt structure for both training rollouts and evaluation, designed to elicit step-by-step reasoning (Chain-of-Thought) and strictly enforce the output format required for rule-based verification.

Standard Prompt Template

User:
`{QUESTION}`
Let's think step by step and enclose the reasoning process within `<think>` and `</think>` tags. The final result in the answer MUST BE within `\boxed{}`.

During preprocessing, this is formatted as a single-turn conversation. The ground-truth answer is attached separately for reward computation.

Verification Logic. The rule-based verifier parses the model's output to extract the content inside the last `\boxed{...}` tag. It then performs robust equality checking (handling fractions, decimals, and algebraic equivalents) against the ground truth to assign a binary reward $\tilde{R} \in \{0, 1\}$.

D.2 Configuration and Hyperparameters

We utilize Group Relative Policy Optimization (GRPO) as the underlying reinforcement learning algorithm. In the VI-CuRL setting, we introduce curriculum-specific parameters.

Table 10: Detailed hyperparameters for GRPO training and VI-CuRL curriculum.

Hyperparameter	Value
Optimization & Model	
Algorithm	GRPO
Training Strategy	Fully Sharded Data Parallel 2 (FSDP2) + Zero Redundancy Optimizer Stage 3 (ZeRO-3)
Optimizer	AdamW
Learning Rate	3×10^{-6}
Total Epochs	1
Global Batch Size	128
Gradient Accumulation	1
Rollout & GRPO	
Rollout Engine	vLLM
Context Window	512 (Input) + 3072 (Output)
Group Size (G)	8
Sampling Temperature	1.0
Advantage	Group-Normalized ($A_i = \frac{R_i - \bar{R}}{\sigma + \epsilon}$)
KL Coefficient (β_{KL})	0.001
Entropy Coefficient	0.0
Reference Model	Frozen Initial Policy
VI-CuRL Curriculum	
Initial Retention (β_{start})	0.2
Curriculum Duration (T_{curr})	100% of Training Steps
Curriculum Signal	Confidence ($c(x)$ from Eq. (5))
Retention Schedule	Linear Annealing

D.3 Dataset Composition

Our experiments leverage a combination of established benchmarks for training and comprehensive evaluation.

Training Source.

DeepScaleR [Luo et al., 2025]: A large-scale math reasoning corpus used for all outcome-supervised training runs.

WebInstruct-Verified [Ma et al., 2025]: A verified instruction-following dataset used for general knowledge RL training. For our general-domain experiments, we filter out mathematics prompts to focus on non-mathematical reasoning and knowledge-intensive tasks.

Mathematical Reasoning Evaluation Suite. We evaluate mathematical reasoning on six distinct datasets to assess robustness and generalization.

MATH500 [Lightman et al., 2024]: A subset of the MATH benchmark containing 500 representative problems.

AIME (2024 & 2025) [HuggingFaceH4, 2024, OpenCompass, 2025]: Recent sets from the American Invitational Mathematics Examination, testing high-difficulty competition math.

AMC 2023 [math-ai, 2025]: Problems from the American Mathematics Competitions, varying in difficulty.

Minerva Math [Lewkowycz et al., 2022]: A comprehensive suite evaluating quantitative reasoning capabilities.

OlympiadBench [He et al., 2024]: An advanced benchmark specifically designing to test olympiad-level problem solving.

General Knowledge Evaluation Suite. To evaluate whether VI-CuRL generalizes beyond formal mathematical reasoning, we additionally test on four general knowledge and reasoning benchmarks.

MMLU-Pro [Wang et al., 2024]: A more challenging version of MMLU designed to evaluate broad multitask language understanding with harder questions and expanded answer choices.

GPQA-Diamond [Rein et al., 2024]: A high-difficulty graduate-level question answering benchmark emphasizing scientific knowledge and expert reasoning.

TheoremQA [Chen et al., 2023]: A theorem-driven question answering benchmark that evaluates whether models can apply formal scientific and mathematical principles to solve problems.

WebInstruct-Validation [Ma et al., 2025]: The validation split associated with WebInstruct-Verified, used to measure general instruction-following and reasoning performance under the same data family as the general-domain training source.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accurately state our theoretical claims and experimental findings in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the current scope of evaluated domains and future extensions in Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All assumptions are clearly stated in Section 4, and complete proofs are provided in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully disclose training configurations, prompt templates, and evaluation datasets in Section 5 and Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our full source code in the supplemental material and discuss dataset access in Appendix D.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed experimental settings and hyperparameters are provided in Section 5 and Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean and standard deviation across random seeds for our main results in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources required (8x A100 GPUs) are specified in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work focuses on foundational RL optimization algorithms and does not introduce immediate direct negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not release models or scraped datasets that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We appropriately cite all the existing datasets and models used in our experiments in Section 5 and Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce new datasets or pre-trained models.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any original or non-standard components; they are solely the subject of optimization in Section 5.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.