Preserving the SPINE: How a Few Critical Tokens Prevent Global Entropy Collapse in LLM Reasoning

Jaeeun Jang*, Hansle Lee*, Sangmin Kim*
*AI Research Team, Hanwha Systems, Seongnam-si, Gyeonggi-do, Republic of Korea

***Abstract*—Reinforcement Learning with Verifiable Rewards (RLVR) and Reinforcement Learning from Internal Feedback (RLIF) often fail to benefit from test-time compute due to *entropy collapse* and the resulting loss of reasoning diversity. Through a token-level analysis of GRPO-style policy updates, we show that this collapse is driven not by uniform entropy decay, but by premature overconfidence at a small set of structurally critical, high-entropy decision points where reasoning trajectories branch (i.e., *forking tokens*). Based on this insight, we propose SPINE (Structural Preservation of Information at Nodal Elements), a structurally targeted entropy-preserving optimization method that selectively preserves entropy at these high-impact tokens via partial KL regularization. SPINE identifies forking tokens via offline entropy profiling and then applies online collapse-pressure scoring to regularize the policy only for the most collapse-prone subset, resulting in entropy control over just ∼2–3% of tokens. Across multiple scales of Qwen2.5 Base models and math reasoning benchmarks, SPINE consistently improves single-sample accuracy and test-time scaling performance (e.g., multi-sample inference) under RLVR and RLIF, demonstrating that targeted entropy preservation at key branching points sustains reasoning diversity without sacrificing optimization efficiency.**

***Index Terms*—Large Reasoning Model, Reinforcement Learning with Verifiable Rewards (RLVR), Reinforcement Learning from Internal Feedback (RLIF), Entropy Regularization**

Reproducibility

The source code and training configurations are publicly available at https://github.com/wkdwodms0779-ai/SPINE.

I. Introduction

The recent emergence of OpenAI's O1 [1] and DeepSeek-R1 [2] has shifted the dominant paradigm for improving large language model (LLM) performance from scaling training-time compute [3] to scaling test-time compute [1], [2]. This shift has renewed interest in Reinforcement Learning with Verifiable Rewards (RLVR), where outcome-level supervision from automatically checkable signals enables large-scale optimization of reasoning behavior. In parallel, supervised fine-tuning (SFT) has been widely used to distill high-quality reasoning trajectories from strong models [4], [5]. While RL often achieves stronger robustness and improved single-sample performance, it tends to reduce reasoning-path diversity compared to SFT [6]. This loss of diversity limits the effectiveness of test-time scaling methods such as multi-sample decoding, majority voting, or self-consistency, making it crucial to preserve output diversity while maintaining accuracy.

While RLVR achieves strong reasoning performance using verifiable, externally defined rewards, its dependence on task- and domain-specific supervision limits scalability and portability. To address this, Reinforcement Learning from Internal Feedback (RLIF) has emerged as a paradigm that optimizes reasoning using intrinsic signals generated by the model itself, without requiring external reward design [7]–[9]. For example, INTUITOR leverages self-certainty as a reward signal to achieve performance comparable to RLVR while exhibiting stronger out-of-distribution generalization [7]. However, intrinsic signals are inherently noisy and imperfectly calibrated, and confidence-based optimization can amplify spurious correlations, often accelerating entropy collapse and premature convergence. As a result, RLIF still lacks a stable mechanism to sustain effective test-time scaling.

Building on these observations, we show that entropy collapse in both RLVR and RLIF arises not from uniform entropy decay over the entire sequence, but from premature overconfidence at a small set of structurally critical decision points. While prior work studies entropy or confidence control within either RLVR or RLIF, no unified approach consistently mitigates entropy collapse across both paradigms. We demonstrate that **explicitly controlling the entropy of only a vanishingly small subset of tokens (approximately 2–3%)** is sufficient to prevent global entropy collapse while improving reasoning performance in both settings. Our approach selectively preserves uncertainty at high-impact *forking tokens*—tokens that govern branching in reasoning trajectories—thereby sustaining constructive exploration while allowing confident convergence elsewhere. Notably, despite the fundamentally different reward sources and optimization dynamics underlying RLVR and RLIF, this targeted entropy control yields consistent gains across both settings. These results suggest that entropy collapse is a shared structural failure mode of post-training alignment. We refer to our method as **SPINE** (**S**tructural **P**reservation of **I**nformation at **N**odal **E**lements).

II. Related Work

A. Forking Tokens as Drivers of Reasoning Diversity

Recent work has examined RL-based reasoning through the lens of **token-level entropy**, revealing that a small subset of high-entropy tokens can exert a disproportionate influence on reasoning outcomes. Wang et al. [10] show that, in Chain-of-Thought (CoT) reasoning, the majority of tokens exhibit near-deterministic distributions (i.e., near-zero entropy), whereas a *high-entropy minority* functions as *forking tokens* that govern branching among alternative reasoning trajectories. Importantly, these forking tokens often correspond to logical connectors, intermediate assumptions, and structural pivot decisions,

rather than low-level lexical completions. Their results further indicate that reasoning performance is particularly sensitive to the entropy at these high-entropy tokens: aggressively reducing their entropy causes rapid performance degradation, whereas preserving—or modestly increasing—entropy at these points yields substantial gains. Complementary to this entropy-centric perspective, CURE [11] treats *high-entropy critical tokens* as explicit indicators of genuine uncertainty in the policy distribution and leverages them to drive structured exploration. Specifically, CURE identifies peak-entropy decision points, truncates CoT sequences at these positions, and re-concatenates preceding clauses to encourage exploration of novel yet coherent reasoning states. By explicitly promoting critical-token–guided branching, CURE mitigates early **entropy collapse**, sustains diverse reasoning trajectories during RL, and achieves state-of-the-art performance on mathematical reasoning benchmarks. Notably, preserving uncertainty at such *forking tokens* enables **constructive competition** among alternative solution paths, whereas entropy collapse at these points induces premature convergence and erodes the benefits of test-time scaling.

B. Entropy Control for Preserving Reasoning Diversity

Reinforcement learning (RL) is widely adopted for post-training alignment of large language models (LLMs) in verifiable reasoning tasks [1], [2], yet it is well known to suffer from *entropy collapse* at critical decision points, where probability mass prematurely concentrates on a single reasoning trajectory [6], [8]. One useful perspective is to view RL-based post-training as a process of *trading (or "selling") entropy for performance*: as optimization progresses, the policy systematically converts uncertainty into reward, gradually depleting its capacity for exploration [12]. When entropy is not explicitly regulated—as in methods such as INTUITOR [7]—RL optimization indiscriminately consumes uncertainty across the entire sequence, causing exploration to collapse well before the model reaches its attainable performance ceiling. To counteract this effect, Diversity-Aware Policy Optimization (abbreviated here as DA-PO) introduces selective entropy regularization on reward-positive outputs, flattening the policy distribution along successful traces and mitigating premature convergence to a single reasoning path [13]. Subsequent analyses further reveal that entropy erosion is mechanistically driven by persistently positive covariance between log-probabilities and advantages, which enforces monotonic entropy decay and ultimately limits scalability in RL-based reasoning [12]. To address this failure mode, they propose **KL-Cov**, which selectively constrains updates on collapse-prone actions, thereby mitigating exploration loss and improving scalability in RL-trained LLMs.

Despite their partial effectiveness, existing entropy-control methods exhibit fundamental structural limitations when applied to LLM-based reasoning. **First**, unlike conventional reinforcement learning environments with small and well-structured action spaces, LLMs operate over an extremely large and heterogeneous token space. In this regime, uniformly applying entropy bonuses at every generation step fails to promote reasoning-oriented exploration and instead degenerates into uninformed stochasticity. For example, approaches such as DA-PO preserve entropy relatively uniformly across the entire sequence, which can interfere with confident convergence at positions where uncertainty should be decisively resolved [13].

Second, prior methods often concentrate their regularization on **surface-level low-entropy tokens**—tokens that merely extend established textual templates without exerting meaningful influence over the underlying reasoning trajectory [10], [11]. Entropy collapse at such already-determined positions has only a marginal impact on global reasoning diversity, as the continuation is effectively predetermined. As a result, these approaches fail to address the **primary bottleneck of exploration**: preserving uncertainty at branching tokens that govern transitions between alternative reasoning paths. Even selective entropy-control techniques such as KL-Cov remain limited in this respect, as they do not sufficiently discriminate between positions where entropy must be preserved and those where it can be safely reduced, thereby failing to realize the optimal trade-off between entropy preservation and performance [12].

From this perspective, our approach moves beyond token-agnostic entropy control by explicitly focusing on *forking tokens*. We show that preserving entropy over only a *vanishingly small subset* of these critical positions, while permitting entropy decay elsewhere, is sufficient to sustain reasoning diversity. Under an entropy–performance trade-off view, this strategy retains uncertainty where it is necessary while selectively "selling" entropy at non-critical locations, thereby allowing RL-trained LLMs to achieve higher asymptotic performance than uniform entropy-regularization approaches. Crucially, although RLVR and RLIF differ substantially in their feedback signals and optimization procedures, this targeted entropy control yields reliable gains in both regimes. Together, these observations point to entropy collapse as a common structural failure mechanism in post-training alignment.

III. Proposed Method: SPINE

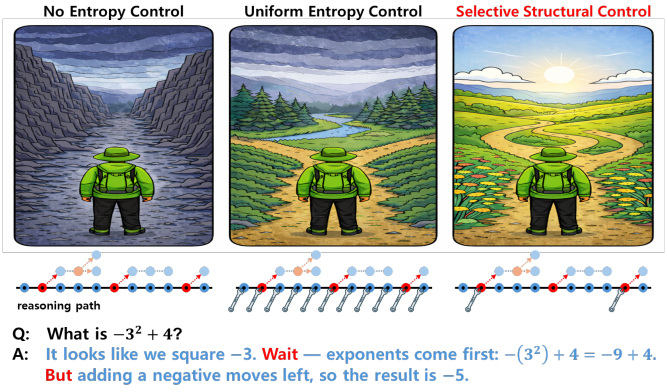


Fig. 1: SPINE selectively preserves entropy at critical branching tokens while allowing confident convergence elsewhere.

A. Origins of Entropy Collapse in RL-based Reasoning

We derive the mechanistic origin of entropy collapse by analyzing how the GRPO clipped surrogate drives token-level logit updates and entropy dynamics at each token position.

Setup. Fix a context $c_t = (x, y_{<t})$ and write $\pi(v) \equiv \pi_\theta(v \mid c_t)$, $z = (z_v)_{v\in\mathcal{V}}$. The policy is parameterized via softmax, $\pi(v) = e^{z_v}/\sum_{u\in\mathcal{V}} e^{z_u}$. The clipped surrogate is $L_t(\theta) = \min\big(U_t(\theta),\ C_t(\theta)\big)$ with $r_t = \pi_\theta(y_t \mid c_t)/\pi_{\text{old}}(y_t \mid c_t)$:

$$U_t(\theta) := r_t(\theta)\hat{A}_t, \qquad C_t(\theta) := \text{clip}(r_t(\theta))\hat{A}_t, \tag{1}$$

where $\hat{A}_t$ is a stop-gradient scalar. Let $g_t \in \{0, 1\}$ be the branch-activity indicator satisfying $\nabla_\theta L_t = g_t\, \nabla_\theta(r_t\hat{A}_t)$ a.e.: $g_t = 1$ when the unclipped branch is active and the gradient flows normally; $g_t = 0$ when clipping suppresses the update.

Logit gradient of the clipped surrogate. By the stop-gradient assumption, $\partial L_t/\partial z_v = g_t\, \hat{A}_t\, \partial r_t/\partial z_v$ a.e.

$$\frac{\partial r_t}{\partial z_v} = \frac{1}{\pi_{\text{old}}(y_t \mid c_t)} \frac{\partial \pi_\theta(y_t \mid c_t)}{\partial z_v}. \tag{2}$$

The softmax derivative gives $\partial\pi(i)/\partial z_v = \pi(i)\big(\mathbf{1}[i{=}v] - \pi(v)\big)$ for any $i, v \in \mathcal{V}$. Setting $i = y_t$:

$$\frac{\partial \pi_\theta(y_t \mid c_t)}{\partial z_v} = \pi(y_t)\big(\mathbf{1}[v{=}y_t] - \pi(v)\big). \tag{3}$$

Substituting into Eq. (2) and using $\pi(y_t)/\pi_{\text{old}}(y_t \mid c_t) = r_t$:

$$\frac{\partial r_t}{\partial z_v} = r_t\big(\mathbf{1}[v{=}y_t] - \pi(v)\big). \tag{4}$$

Therefore the logit gradient and local ascent contribution are

$$\frac{\partial L_t}{\partial z_v} = g_t\, r_t\, \hat{A}_t\big(\mathbf{1}[v{=}y_t] - \pi(v)\big), \tag{5}$$

$$\Delta z_v = \eta\, g_t\, r_t\, \hat{A}_t\big(\mathbf{1}[v{=}y_t] - \pi(v)\big). \tag{6}$$

Note that Δz_v is the local logit-coordinate ascent contribution at fixed c_t, not the actual optimizer update.

Entropy gradient. For the Shannon entropy $H(z) = -\sum_v \pi(v)\log\pi(v)$, differentiating via the product rule:

$$\frac{\partial H}{\partial z_v} = -\sum_{i\in\mathcal{V}} \frac{\partial \pi(i)}{\partial z_v}\big(\log\pi(i) + 1\big). \tag{7}$$

Substituting the softmax derivative:

$$\begin{aligned} &= -\sum_{i\in\mathcal{V}} \pi(i)\big(\mathbf{1}[i{=}v] - \pi(v)\big)\big(\log\pi(i) + 1\big) \\ &= -\pi(v)\big(\log\pi(v) + 1\big) + \pi(v)\sum_{i\in\mathcal{V}} \pi(i)\big(\log\pi(i) + 1\big) \end{aligned} \tag{8}$$

Defining $\mu_t := \sum_v \pi(v)\log\pi(v)$:

$$\begin{aligned} \frac{\partial H}{\partial z_v} &= -\pi(v)(\log\pi(v) + 1) + \pi(v)(\mu_t + 1) \\ &= -\pi(v)\big(\log\pi(v) - \mu_t\big). \end{aligned} \tag{9}$$

Entropy change decomposition. Since $H(z)$ is C^∞ on finite logits, Taylor's theorem with remainder gives exactly:

$$\Delta H_t := H(z{+}\Delta z) - H(z) = \sum_v \frac{\partial H}{\partial z_v}\Delta z_v + R_{2,t}\,, \tag{10}$$

where the remainder admits the exact integral form

$$R_{2,t} = \int_0^1 (1-s)\, \Delta z^\top \nabla_z^2 H(z{+}s\Delta z)\, \Delta z\, ds. \tag{11}$$

Letting $M_{H,t} := \sup_{s\in[0,1]} \|\nabla_z^2 H(z{+}s\Delta z)\|_{\text{op}}$ bound the Hessian along the path and using $\int_0^1 (1-s)\, ds = \frac{1}{2}$:

$$|R_{2,t}| \le \tfrac{1}{2}\, M_{H,t}\, \|\Delta z\|_2^2. \tag{12}$$

Expanding the first-order term by substituting Eqs. (6) and (9):

$$\begin{aligned} \sum_v \frac{\partial H}{\partial z_v}\Delta z_v &= \eta\, g_t\, r_t\, \hat{A}_t \sum_v \frac{\partial H}{\partial z_v}\big(\mathbf{1}[v{=}y_t] - \pi(v)\big) \\ &= \eta\, g_t\, r_t\, \hat{A}_t\left(\frac{\partial H}{\partial z_{y_t}} - \sum_v \pi(v)\frac{\partial H}{\partial z_v}\right). \end{aligned} \tag{13}$$

Recognizing $\sum_v \pi(v)\, \partial H/\partial z_v = \mathbb{E}_{y_t\sim\pi}[\partial H/\partial z_{y_t} \mid c_t]$ and taking the conditional expectation over $y_t \sim \pi(\cdot \mid c_t)$:

$$\begin{aligned} \mathbb{E}[\Delta H_t \mid c_t] = \eta\, \mathbb{E}\left[g_t r_t \hat{A}_t \left(\frac{\partial H}{\partial z_{y_t}} - \mathbb{E}\left[\frac{\partial H}{\partial z_{y_t}} \middle| c_t \right] \right) \middle| c_t \right] \\ + \mathbb{E}[R_{2,t} \mid c_t]. \end{aligned} \tag{14}$$

Applying the identity $\mathbb{E}[X(Y{-}\mathbb{E}[Y])] = \text{Cov}(X, Y)$ with $X := g_t r_t \hat{A}_t$ and $Y := \partial H/\partial z_{y_t}$:

$$\begin{aligned} \mathbb{E}[\Delta H_t \mid c_t] = \eta\, \text{Cov}_{y_t\sim\pi(\cdot|c_t)}\Big(g_t r_t \hat{A}_t,\ \frac{\partial H}{\partial z_{y_t}} \Big) \\ + \mathbb{E}[R_{2,t} \mid c_t]. \end{aligned} \tag{15}$$

Substituting $\partial H/\partial z_{y_t} = -\pi(y_t)(\log\pi(y_t) - \mu_t)$ from Eq. (9):

Proposition 1 (Entropy Variation). *For a fixed context c_t,*

$$\begin{aligned} &\mathbb{E}[\Delta H_t \mid c_t] \\ &= -\eta\, \text{Cov}_{y_t\sim\pi(\cdot|c_t)}\Big(g_t r_t \hat{A}_t,\ \pi(y_t)\big(\log\pi(y_t) - \mu_t\big) \Big) \\ &\quad + \mathbb{E}[R_{2,t} \mid c_t]. \end{aligned} \tag{16}$$

Strict entropy decrease $\mathbb{E}[\Delta H_t \mid c_t] < 0$ is guaranteed whenever the first-order drive exceeds the remainder. Since $\Delta z \propto \eta$ (Eq. (6)), the remainder scales as $O(M_{H,t}\,\eta^2)$ while the covariance term scales as $O(\eta)$, so the condition is satisfied whenever η is sufficiently small relative to $M_{H,t}$. In practice, GRPO's clipping ($g_t = 0$ when r_t exits $[1{-}\epsilon, 1{+}\epsilon]$) inherently bounds $\|\Delta z\|$, ensuring the small-update regime.

The covariance term reveals the mechanistic origin of entropy collapse. Specifically, entropy decreases when tokens that receive large positive update signals (i.e., high $g_t r_t \hat{A}_t$) are aligned with entropy-reducing directions, characterized by $\pi(y_t)(\log\pi(y_t) - \mu_t)$, which measures deviation from the average log-probability. This implies that entropy collapse is primarily driven by token instances that simultaneously exhibit strong reward-weighted updates and high entropy sensitivity.

Collapse pressure score. Computing the conditional covariance in Eq. (16) over the token distribution is intractable during training. For a batch of N token instances, define

$$X_i := g_i\, r_i\, \hat{A}_i\,, \tag{17}$$

$$Y_i := \pi(y_i \mid c_i)\big(\log\pi(y_i \mid c_i) - \mu_{c_i}\big), \tag{18}$$

with $\mu_{c_i} := \sum_v \pi(v \mid c_i)\log\pi(v \mid c_i)$. The empirical covariance estimator over the batch of token instances is

$$\widehat{\text{Cov}}(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}). \tag{19}$$

To better approximate the conditional covariance in Proposition 1, we center within each prompt group G rather than globally across the batch, reducing cross-prompt mean shifts:

$$\bar{X}_G := \frac{1}{|G|}\sum_{j\in G} X_j, \qquad \bar{Y}_G := \frac{1}{|G|}\sum_{j\in G} Y_j. \tag{20}$$

The *collapse pressure* of instance i is

$$s_i := \big(X_i - \bar{X}_G\big)\big(Y_i - \bar{Y}_G\big). \tag{21}$$

Instances with $s_i > 0$ contribute positively to the first-order entropy-decrease driver, reinforcing entropy-reducing updates; larger values indicate stronger collapse-inducing pressure.

B. Identifying Forking Tokens via Offline Entropy Profiling

We first perform an **offline entropy profiling** stage to characterize the uncertainty structure of the base model π_{base} and identify *forking tokens*—token positions where alternative reasoning branches plausibly emerge. For each question q, we sample $K = 8$ reasoning rollouts $o^{(k)} = (o_1^{(k)}, \ldots, o_{T_k}^{(k)})$ from base policy π_{base} under a fixed decoding temperature τ:

$$\pi_{\text{base}}(o^{(k)} \mid q) = \prod_{t=1}^{T_k} \pi_{\text{base}}(o_t^{(k)} \mid q, o_{<t}^{(k)}) \tag{22}$$

For every rollout k and generation step t, we compute the entropy of the base policy's next-token distribution as:

$$H_t^{(k)} := -\sum_{j=1}^{|\mathcal{V}|} p_{t,j}^{(k)} \log p_{t,j}^{(k)}, \quad p_t^{(k)} = \text{Softmax}(z_t^{(k)}/\tau) \tag{23}$$

Here, $z_t^{(k)}$ denotes the pre-softmax logit vector produced by the base model π_{base} at generation step t for rollout k.

The resulting token-level entropy values $\{H_t^{(k)}\}$, obtained via temperature-scaled softmax, are pooled across all rollouts and token positions generated for question q. We then define a global entropy threshold using a percentile-based criterion.

$$h_\rho = \mathrm{Q}_\rho\big(\{H_{t'}^{(k')}\}_{k',t'}\big), \; \mathcal{F}(q) = \{(k,t) \mid H_t^{(k)} \geq h_\rho\} \tag{24}$$

Here, $\mathrm{Q}_\rho(\cdot)$ denotes the $(1-\rho)$-quantile operator applied to the empirical distribution of token-level entropy values $\{H_{t'}^{(k')}\}_{k',t'}$. The threshold h_ρ therefore selects the top-ρ fraction of highest-entropy token positions, and the resulting set $\mathcal{F}(q)$ identifies *forking tokens* (empirically, $\rho \approx 20\%$).

C. Collapse-Aware KL Regularization

We introduce a **selective** KL penalty that targets the most collapse-critical positions among entropy-intensive tokens.

1) Collapse pressure score: For a rollout generated by the behavior policy π_{old}, consider the sampled token $y_t^{(k)}$ at step t of rollout k. We define the **collapse pressure** at this step as:

$$s_{k,t} = \big(X_{k,t} - \bar{X}_G\big)\big(Y_{k,t} - \bar{Y}_G\big) \tag{25}$$

where $X_{k,t} := g_{k,t}\, r_{k,t}\, \hat{A}_{k,t}$ and $Y_{k,t} := \pi_\theta\big(y_t^{(k)} \mid x, y_{<t}^{(k)}\big)\big(\log \pi_\theta\big(y_t^{(k)} \mid x, y_{<t}^{(k)}\big) - \mu_{k,t}\big)$ with $\mu_{k,t} := \sum_v \pi_\theta(v \mid x, y_{<t}^{(k)}) \log \pi_\theta(v \mid x, y_{<t}^{(k)})$, and $\bar{X}_G, \bar{Y}_G$ are the prompt-group means defined in Eq. (20). The collapse pressure $s_{k,t}$ is computed online during RL training for each generated token.

2) Top-q% Collapse-Pressure Selection: Given the entropy threshold h_ρ, we focus on high-entropy token positions and compute a collapse-pressure threshold c_q corresponding to the top-q percentile, from which we define a binary mask $m_{k,t}$:

$$c_q = \mathrm{Q}_q\Big(\{\, s_{k,t} \;\mid\; H_t^{(k)} \geq h_\rho \,\}\Big) \tag{26}$$

$$m_{k,t} = \mathbb{I}\Big[\big(H_t^{(k)} \geq h_\rho\big) \;\wedge\; \big(s_{k,t} \geq c_q\big) \;\wedge\; \big(\hat{A}_{k,t} > 0\big)\Big] \tag{27}$$

where q is a small hyperparameter (typically $q \in [10\%, 15\%]$).

3) Token-wise KL penalty on selected positions: To prevent entropy collapse at the most critical decision points, we impose a token-wise KL regularization selectively applied to tokens exhibiting both high entropy and high collapse pressure. We define the KL divergence between the next-token distributions:

$$D_{k,t} = \mathrm{D}_{\mathrm{KL}}\Big(\pi_\theta(\cdot \mid x, y_{<t}^{(k)}) \;\Big\|\; \pi_{\text{old}}(\cdot \mid x, y_{<t}^{(k)})\Big) \tag{28}$$

We then incorporate this selective constraint directly into the GRPO (Group Relative Policy Optimization) objective [14]:

$$\max_\theta \; J_{\text{GRPO}}(\theta) - \frac{\beta}{\sum_k T_k} \sum_k \sum_{t=1}^{T_k} m_{k,t}\, D_{k,t} \tag{29}$$

where the KL penalty term is applied to tokens with $m_{k,t} = 1$.

IV. Experimental Setup

A. Training and Evaluation Setup.

All RLVR and RLIF experiments use the Open-R1 framework and are trained on the MATH training set (7,500 problems [15]) with Qwen2.5 Base models (1.5B, 3B, 7B). Each update samples 8 candidate solutions for 128 problems.

Both paradigms share identical hyperparameters for a fair comparison and differ only in reward construction. RLVR uses verifiable rewards based on answer correctness, comparing the model's final prediction (from the `\boxed{Your Answer}` field) with the ground truth. RLIF instead employs *self-certainty* [7] as an internal feedback signal, reflecting the model's confidence in its reasoning without external labels.

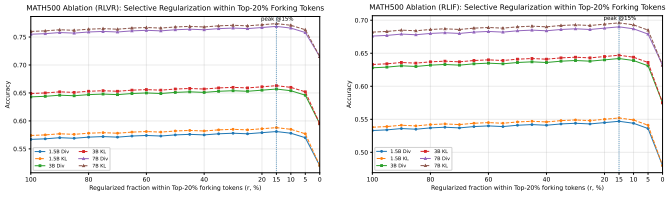
We evaluate on four held-out mathematical reasoning benchmarks: **GSM8K** [16], **MATH500** [17], **AMC** [18], and **AIME2024** [18]. For each problem, we generate 16 independent responses under stochastic decoding with temperature $T = 1.0$ and report average exact-match accuracy (Acc@16).

B. Ablation Studies

1) Forking-Token Definition Ablation: We conduct an ablation study to examine the role of forking tokens in preserving reasoning diversity under entropy-control regularization. This analysis spans model scales, benchmarks, and both RLVR and RLIF settings, while varying the entropy percentile threshold used to define the forking-token set, thereby assessing which proportion of high-entropy tokens yields optimal performance.

We compare two regularization terms adopted from prior work. The first is a **diversity-aware regularizer** (**Div**) [13]:

$$\mathcal{J}_{\text{Div}}(\pi_\theta) = -\sum_{t=1}^{T} \frac{\pi_\theta(y_t \mid y_{<t})}{\pi_{\text{old}}(y_t \mid y_{<t})} \log \pi_\theta(y_t \mid y_{<t}) \tag{30}$$



(a) RLVR (MATH500) (b) RLIF (MATH500)

Fig. 2: MATH500 ablation on selective Div/KL entropy control within the top-20% forking tokens (ranked by SPINE).

Applying Div to the **top 100%** of forking tokens is mathematically equivalent to the original DA-PO objective from prior work, and thus provides a suitable baseline for comparison.

The second is a **KL-based entropy-control regularizer** (**KL**) between the rollout policy and the current policy [12]:

$$\mathcal{J}_{\mathrm{KL}}(\pi_\theta) = -\sum_{t=1}^{T} \mathrm{D}_{\mathrm{KL}}\Big(\pi_{\mathrm{old}}(y_t \mid y_{<t}) \,\Big\|\, \pi_\theta(y_t \mid y_{<t})\Big) \quad (31)$$

We sweep the entropy percentile threshold h_p (Eq. 24) used to define the forking tokens and apply each regularizer only to tokens with entropy above h_p, where $p \in \{0, 10, \ldots, 100\}$. At $p = 0\%$, no entropy regularization is applied, making the setup equivalent to the native training objectives of each paradigm: standard GRPO for RLVR and the INTUITOR objective [7], which uses self-certainty as internal feedback for RLIF.

Table I reveals a consistent non-monotonic trend across all benchmarks and model sizes, peaking when forking tokens are defined as the top 20% highest-entropy tokens. Although slight gains occasionally appear at 10% or 30%, they are not consistent across tasks or scales. Overall, allocating entropy regularization to **20%** yields the most robust trade-off between exploration preservation and confident convergence under both RLVR and RLIF. Notably, the proposed structurally guided allocation also improves over two established methods. Compared to DA-PO, which applies entropy regularization uniformly across reasoning steps (implemented in our table as RLVR with Top-100% Div), selectively constraining only structurally critical tokens yields superior performance. Likewise, relative to INTUITOR, whose training objective permits entropy to decay without structural constraints (corresponding to RLIF with Top-0%), preserving uncertainty at key decision points leads to more effective reasoning outcomes. Across these settings, the KL regularizer consistently outperforms Div, further supporting the effectiveness of divergence-based entropy control when regularization is applied selectively to structurally critical tokens. These results indicate that optimal performance arises not from uniformly enforcing or fully relaxing entropy control, but from allocating it according to structural importance within the reasoning process.

2) Sparse Entropy Control within Forking Tokens: Unlike the previous ablation where p defines the forking-token set itself, here the forking set is fixed to the top-20% highest-entropy tokens. We conduct an ablation study on MATH500 by varying the fraction r of tokens within this set that receive entropy regularization, ranked by SPINE collapse pressure, to examine whether a sweet spot also exists within the forking-token set. When $r = 100\%$, all forking tokens are regularized; when $r = 0\%$, none are. Across model sizes under both RLVR and RLIF, performance shows a non-monotonic trend, peaking at approximately **15%** of the forking-token set. These results show that better performance is achieved by allowing most forking tokens to remain unconstrained while regulating only a very small, structurally critical subset (≈2–3% of all tokens). Similar patterns are observed on GSM8K, AMC, and AIME.

3) Where to Regularize under a Fixed Budget: We evaluate whether, under a fixed KL regularization budget, reasoning performance is determined not by the number of regularized tokens, but by **which token positions are selected**. To ensure a strictly fair comparison, all variants apply KL regularization to the same *proportion* of tokens as selected by SPINE.

a) Token selection strategies: We compare eight token-selection strategies (see Table II). Specifically, **SPINE** selects tokens at the intersection of forking tokens and high collapse-pressure tokens; **Rand-All** selects tokens uniformly at random from the entire sequence; **Rand-PA** restricts random selection to positive-advantage tokens; **Rand-F** samples randomly from the forking-token set; **Rand-F-PA** further restricts this to positive-advantage forking tokens; **Rand-NF** samples from non-forking positions; **KL-Cov** selects tokens according to covariance-aware entropy control, following the approach proposed in [12]; and **Top-CP** applies regularization to the top tokens ranked by collapse pressure across the full sequence.

b) Entropy collapse metric: To quantify how much reasoning diversity is lost during training, we additionally report *collapse_degree*, a scalar proxy measure of entropy collapse. Specifically, collapse_degree captures the relative reduction in the fraction of high-entropy forking tokens after RL training.

c) Results: Despite regularizing the same proportion of tokens (on average) across all variants, we observe substantial differences in both reasoning performance and entropy dynamics. This shows that the entropy–performance trade-off is governed not by the *amount* of entropy preserved, but by its *structural placement*. Regularizing collapse-critical tokens yields consistent gains across model sizes and tasks, whereas constraining structurally irrelevant tokens provides limited benefit; in particular, Rand-NF, which regularizes only non-forking positions, performs worst, often falling below even Rand-All. Notably, lower *collapse_degree* does not necessarily imply better performance—several strategies achieve comparable collapse degree but differ markedly in accuracy—indicating that indiscriminate entropy control retains uninformative rather than reasoning-relevant diversity. Crucially, even when KL regularization is applied *locally* at the same rate, its impact on preventing entropy collapse differs *globally* depending on where it is placed. Selectively suppressing updates at tokens with high collapse contribution enables local interventions to propagate along the sequence, mitigating global entropy collapse while preserving exploration and improving performance under an identical budget. Moreover, RLIF variants exhibit systematically higher collapse degrees than RLVR across all

TABLE I: Ablation on the entropy percentile threshold used to define the forking-token set for entropy regularization (Top-p%).

Model	Ratio	RLVR								RLIF							
		GSM8K		MATH500		AMC		AIME		GSM8K		MATH500		AMC		AIME	
		Div	KL	Div	KL	Div	KL	Div	KL	Div	KL	Div	KL	Div	KL	Div	KL
1.5B	Top 100%	0.716	0.721	0.533	0.538	0.129	0.131	0.000	0.000	0.676	0.680	0.495	0.499	0.060	0.063	0.000	0.000
	Top 90%	0.722	0.727	0.539	0.544	0.136	0.139	0.000	0.015	0.683	0.687	0.502	0.506	0.067	0.070	0.000	0.000
	Top 80%	0.729	0.734	0.546	0.551	0.142	0.145	0.000	0.015	0.690	0.694	0.509	0.513	0.073	0.076	0.000	0.000
	Top 70%	0.735	0.740	0.552	0.557	0.148	0.151	0.015	0.015	0.697	0.701	0.516	0.520	0.080	0.083	0.000	0.015
	Top 60%	0.741	0.746	0.558	0.564	0.153	0.156	0.015	0.015	0.703	0.708	0.522	0.527	0.085	0.089	0.000	0.015
	Top 50%	0.746	0.752	0.562	0.568	0.157	0.160	0.015	0.026	0.709	0.714	0.528	0.533	0.090	0.094	0.015	0.015
	Top 40%	0.749	0.755	0.564	0.570	0.158	0.161	0.015	0.026	0.711	0.716	0.529	0.534	0.091	0.095	0.015	0.015
	Top 30%	0.753	0.759	0.564	0.571	0.158	**0.166**	**0.033**	0.026	0.713	0.719	0.530	**0.539**	0.091	0.095	**0.026**	**0.026**
	Top 20%	**0.756**	0.762	**0.567**	**0.574**	**0.161**	0.165	0.031	**0.033**	**0.716**	**0.722**	**0.533**	0.538	**0.094**	**0.098**	0.015	0.022
	Top 10%	0.752	**0.763**	0.564	0.571	0.158	0.162	0.015	0.026	0.712	0.719	0.530	0.535	0.091	0.095	0.015	0.015
	Top 0%	0.702	0.702	0.520	0.520	0.120	0.120	0.000	0.000	0.655	0.655	0.480	0.480	0.055	0.055	0.000	0.000
3B	Top 100%	0.793	0.797	0.603	0.608	0.249	0.252	0.015	0.015	0.764	0.767	0.580	0.583	0.123	0.126	0.015	0.015
	Top 90%	0.801	0.806	0.611	0.616	0.257	0.260	0.033	0.033	0.772	0.776	0.588	0.592	0.131	0.135	0.033	0.033
	Top 80%	0.807	0.812	0.617	0.622	0.263	0.266	0.033	0.033	0.779	0.783	0.595	0.599	0.138	0.141	0.033	0.033
	Top 70%	0.814	0.819	0.624	0.630	0.270	0.274	0.054	0.054	0.787	0.791	0.603	0.607	0.145	0.149	0.033	0.033
	Top 60%	0.819	0.824	0.629	0.635	0.275	0.278	0.054	0.054	0.794	0.798	0.610	0.615	0.150	0.154	0.033	0.033
	Top 50%	0.824	0.829	0.634	0.640	0.281	0.285	0.054	0.054	0.799	0.804	0.616	0.621	0.156	0.160	0.033	0.054
	Top 40%	0.826	0.832	0.636	0.642	0.283	0.287	0.054	0.054	0.803	0.808	0.620	0.625	0.160	0.164	0.033	0.054
	Top 30%	0.830	0.836	0.640	0.646	**0.291**	0.291	0.054	0.061	0.807	0.812	0.625	0.630	0.162	**0.170**	0.033	0.054
	Top 20%	**0.833**	**0.839**	**0.643**	0.649	0.290	**0.294**	0.061	0.061	**0.810**	0.815	**0.628**	**0.633**	**0.165**	0.169	0.054	0.054
	Top 10%	0.830	0.836	0.640	**0.650**	0.287	0.291	**0.065**	**0.072**	0.807	**0.816**	0.625	0.630	0.162	0.166	**0.057**	**0.061**
	Top 0%	0.780	0.780	0.595	0.595	0.245	0.245	0.015	0.015	0.760	0.760	0.575	0.575	0.120	0.120	0.015	0.015
7B	Top 100%	0.827	0.831	0.716	0.721	0.553	0.558	0.125	0.125	0.809	0.813	0.633	0.637	0.259	0.263	0.033	0.054
	Top 90%	0.835	0.840	0.724	0.729	0.561	0.566	0.136	0.136	0.817	0.822	0.641	0.646	0.267	0.271	0.054	0.061
	Top 80%	0.842	0.847	0.731	0.736	0.568	0.573	0.135	0.138	0.825	0.829	0.649	0.653	0.275	0.279	0.061	0.061
	Top 70%	0.849	0.854	0.738	0.743	0.575	0.580	0.141	0.142	0.831	0.836	0.656	0.661	0.282	0.286	0.061	0.072
	Top 60%	0.853	0.858	0.742	0.747	0.580	0.585	0.141	0.144	0.837	0.842	0.661	0.666	0.287	0.291	0.072	0.072
	Top 50%	0.858	0.863	0.747	0.752	0.585	0.590	0.145	0.147	0.842	0.847	0.666	0.671	0.292	0.296	0.082	0.086
	Top 40%	0.860	0.865	0.749	0.754	0.587	0.592	0.147	0.147	0.844	0.849	0.669	0.674	0.294	0.299	0.082	0.086
	Top 30%	**0.866**	0.868	0.752	0.757	0.590	0.600	0.154	0.156	0.848	0.851	0.673	**0.683**	0.299	0.301	0.082	**0.092**
	Top 20%	0.865	**0.871**	**0.755**	**0.760**	**0.593**	**0.604**	**0.162**	**0.165**	0.848	**0.854**	**0.676**	0.682	**0.299**	**0.304**	**0.086**	**0.092**
	Top 10%	0.862	0.868	0.752	0.757	0.590	0.601	0.160	**0.165**	**0.849**	0.851	0.673	0.679	0.296	0.301	0.082	0.086
	Top 0%	0.825	0.825	0.715	0.715	0.550	0.550	0.125	0.125	0.808	0.808	0.632	0.632	0.258	0.258	0.054	0.054

strategies, reflecting additional entropy-reducing pressure from confidence-based intrinsic objectives, which makes structurally targeted regularization even more critical under RLIF.

C. SPINE Integration with Existing RLVR and RLIF Methods

1) RLVR Results: Table III reports RLVR results on Qwen2.5 Base models. **Baseline** corresponds to standard GRPO with verifiable rewards. **w/ Forking RL**, **w/ Div**, and **w/ KL-Cov** are implemented *identically* to the original methods proposed in their respective prior works [10], [12], [13]. Specifically, **w/ Forking RL** restricts policy-gradient updates to high-entropy forking-token positions while masking gradients elsewhere, **w/ Div** applies the diversity-aware regularization, and **w/ KL-Cov** follows the covariance-based entropy-control objective. **w/ Forking RL + SPINE** extends the forking-token RLVR setting by preventing entropy collapse only at a small subset of collapse-critical tokens within the forking-token set. Similarly, **w/ Div + SPINE** applies the same diversity regularizer *only* to the SPINE-selected tokens, while **KL-Cov + SPINE** further refines token selection by applying the covariance-based collapse pressure score (Eq. (25)) to identify collapse-critical positions within the forking-token set.

Across all model sizes, adding **SPINE** consistently improves each baseline. We further examine the **KL-Cov + SPINE** case, which achieves the highest overall performance. The results indicate that SPINE more precisely identifies collapse-critical tokens than KL-Cov, enabling more effective mitigation of entropy collapse and improved RLVR accuracy.

2) RLIF Results: Table IV reports results under the RLIF setting on Qwen2.5 Base models. All methods rely exclusively on intrinsic feedback signals following prior RLIF approaches, without access to external verifiable rewards. **w/ Self-Certainty**, **w/ Tok-Entropy**, and **w/ Traj-Entropy** are implemented *identically* to the original formulations proposed in prior work [7], [9], [19]. Specifically, **Self-Certainty** uses the KL divergence between the model's next token distribution and a uniform distribution over the vocabulary as a trajectory-level confidence signal; **Token-Level Entropy** applies a dense intrinsic reward that penalizes entropy at each decoding step; and **Trajectory-Level Entropy** aggregates token entropies into a single sequence-level confidence objective. Their SPINE variants — **Self-Certainty + SPINE**, **Tok-Entropy + SPINE**, and **Traj-Entropy + SPINE** — retain the same intrinsic reward formulations but apply entropy-preserving KL regularization only to the SPINE-selected subset of collapse-critical tokens. This enables structurally targeted mitigation of entropy collapse while leaving the original RLIF objectives unchanged.

Across model sizes, intrinsic-reward baselines outperform the base models but yield only limited improvements on more challenging benchmarks, indicating that entropy-reducing intrinsic objectives can suppress exploratory behavior and thereby constrain the model's reasoning capabilities. In contrast, augmenting each intrinsic-reward objective with **SPINE** consistently provides additional gains by locally preventing entropy collapse at structurally critical tokens, which in turn sustains more effective global exploration during RL training.

TABLE II: Ablation on token selection under a fixed KL-regularization budget. All methods apply KL regularization to the same proportion of tokens. RLVR uses GRPO with verifiable rewards, while RLIF uses self-certainty as internal feedback.

Model	Method	RLVR GSM8K	RLVR MATH500	RLVR AMC	RLVR AIME	RLVR collapse_degree	RLIF GSM8K	RLIF MATH500	RLIF AMC	RLIF AIME	RLIF collapse_degree
1.5B	Rand-All	0.708	0.521	0.112	0.000	0.28	0.690	0.510	0.076	0.000	0.45
	Rand-PA	0.724	0.533	0.126	0.015	0.26	0.709	0.524	0.094	0.015	0.42
	Rand-F	0.722	0.535	0.126	0.015	0.23	0.712	0.522	0.094	0.015	0.42
	Rand-F-PA	0.737	0.560	0.139	0.033	0.21	0.714	0.532	0.102	0.026	**0.36**
	Rand-NF	0.688	0.500	0.096	0.000	0.33	0.656	0.481	0.055	0.000	0.48
	KL-Cov [12]	0.738	0.560	0.138	0.054	0.24	0.713	0.531	0.101	0.033	0.41
	Top-CP	0.752	0.564	0.162	**0.073**	0.21	0.710	0.536	0.103	0.047	0.41
	SPINE	**0.769**	**0.581**	**0.170**	0.072	**0.18**	**0.728**	**0.545**	**0.104**	**0.054**	0.37
3B	Rand-All	0.786	0.597	0.232	0.054	0.24	0.765	0.580	0.140	0.033	0.39
	Rand-PA	0.798	0.605	0.256	0.072	0.21	0.783	0.598	0.165	0.054	0.36
	Rand-F	0.800	0.610	0.261	0.071	0.17	0.788	0.596	0.165	0.062	0.35
	Rand-F-PA	0.813	0.624	0.270	0.086	0.15	0.802	0.618	0.174	0.072	0.32
	Rand-NF	0.767	0.576	0.215	0.012	0.28	0.761	0.566	0.121	0.009	0.40
	KL-Cov [12]	0.815	0.626	0.274	0.086	0.19	0.806	0.617	0.173	0.082	0.38
	Top-CP	0.835	0.639	0.292	0.090	0.15	**0.822**	0.628	0.172	0.082	0.33
	SPINE	**0.847**	**0.657**	**0.301**	**0.092**	**0.13**	0.820	**0.630**	**0.176**	**0.086**	**0.31**
7B	Rand-All	0.818	0.709	0.553	0.103	0.22	0.802	0.650	0.280	0.054	0.38
	Rand-PA	0.832	0.725	0.569	0.123	0.18	0.820	0.676	0.301	0.086	0.35
	Rand-F	0.829	0.723	0.575	0.123	0.16	0.818	0.676	0.301	0.086	0.33
	Rand-F-PA	0.846	0.747	0.591	0.160	0.13	0.839	0.690	0.312	0.092	**0.30**
	Rand-NF	0.800	0.690	0.527	0.095	0.25	0.789	0.633	0.259	0.033	0.39
	KL-Cov [12]	0.841	0.750	0.584	0.141	0.17	0.838	0.689	0.311	0.092	0.36
	Top-CP	**0.881**	0.755	0.614	0.171	0.14	0.848	0.689	0.313	0.099	0.32
	SPINE	0.878	**0.768**	**0.622**	**0.181**	**0.11**	**0.852**	**0.692**	**0.314**	**0.110**	**0.30**

TABLE III: RLVR results on Qwen2.5 Base models.

	GSM8K	MATH500	AMC	AIME
Qwen2.5-1.5B-Base				
Baseline (GRPO)	0.702	0.520	0.120	0.000
w/ Forking RL [10]	0.745	0.556	0.147	0.035
w/ Forking RL + SPINE	0.760	0.572	0.166	0.066
w/ Div [13]	0.716	0.533	0.129	0.000
w/ Div + SPINE	0.764	0.576	0.167	0.065
w/ KL-Cov [12]	0.738	0.560	0.138	0.054
w/ KL-Cov + SPINE	0.768	0.580	0.169	0.072
Qwen2.5-3B-Base				
Baseline (GRPO)	0.780	0.595	0.245	0.015
w/ Forking RL [10]	0.812	0.620	0.269	0.066
w/ Forking RL + SPINE	0.838	0.648	0.293	0.086
w/ Div [13]	0.793	0.603	0.249	0.015
w/ Div + SPINE	0.842	0.652	0.296	0.084
w/ KL-Cov [12]	0.815	0.626	0.274	0.086
w/ KL-Cov + SPINE	0.846	0.656	0.299	0.090
Qwen2.5-7B-Base				
Baseline (GRPO)	0.825	0.715	0.550	0.125
w/ Forking RL [10]	0.834	0.731	0.582	0.133
w/ Forking RL + SPINE	0.869	0.758	0.622	0.156
w/ Div [13]	0.827	0.716	0.553	0.125
w/ Div + SPINE	0.873	0.762	0.617	0.165
w/ KL-Cov [12]	0.841	0.750	0.584	0.141
w/ KL-Cov + SPINE	0.876	0.766	0.620	0.171

TABLE IV: RLIF results on Qwen2.5 Base models.

	GSM8K	MATH500	AMC	AIME
Qwen2.5-1.5B-Base				
Baseline	0.582	0.450	0.040	0.000
w/ self-certainty [7]	0.655	0.480	0.055	0.000
w/ self-certainty + SPINE	0.728	0.545	0.104	0.054
w/ tok-level entropy [9]	0.672	0.493	0.068	0.015
w/ tok-level entropy + SPINE	0.734	0.548	0.106	0.057
w/ traj-level entropy [19]	0.648	0.462	0.049	0.014
w/ traj-level entropy + SPINE	0.721	0.534	0.098	0.049
Qwen2.5-3B-Base				
Baseline	0.673	0.544	0.093	0.000
w/ self-certainty [7]	0.760	0.575	0.120	0.015
w/ self-certainty + SPINE	0.820	0.630	0.176	0.086
w/ tok-level entropy [9]	0.775	0.588	0.131	0.036
w/ tok-level entropy + SPINE	0.822	0.633	0.176	0.085
w/ traj-level entropy [19]	0.746	0.568	0.112	0.024
w/ traj-level entropy + SPINE	0.804	0.626	0.165	0.076
Qwen2.5-7B-Base				
Baseline	0.742	0.598	0.236	0.056
w/ self-certainty [7]	0.808	0.632	0.258	0.054
w/ self-certainty + SPINE	0.852	0.692	0.314	0.110
w/ tok-level entropy [9]	0.821	0.640	0.270	0.048
w/ tok-level entropy + SPINE	0.854	0.692	0.316	0.112
w/ traj-level entropy [19]	0.801	0.614	0.246	0.044
w/ traj-level entropy + SPINE	0.848	0.676	0.301	0.104

D. Test-Time Scaling Analysis

To analyze how SPINE influences reasoning behavior under different post-training paradigms, we examine Pass@k scaling curves (Figure 3). This evaluation characterizes how RL-based post-training shapes the model's reasoning distribution and affects test-time scaling behavior under stochastic sampling.

Under RLVR, SPINE consistently outperforms both the Base model and standard RLVR (GRPO). The largest gains appear in the small-k regime, indicating improved **sampling efficiency**: SPINE increases the likelihood of solution-relevant trajectories that already exist in the base model. This improvement arises from structurally targeted entropy preservation at high-impact forking tokens, which enables more effective distribution reshaping and higher early-sample success without sacrificing long-horizon reasoning coverage. As k grows, the performance gap between Base and SPINE narrows, suggesting that the underlying **reasoning support**—which problems are ultimately solvable—remains largely determined by the base model [20], [21]. By contrast, standard RLVR performs largely token-agnostic updates, allowing entropy to collapse

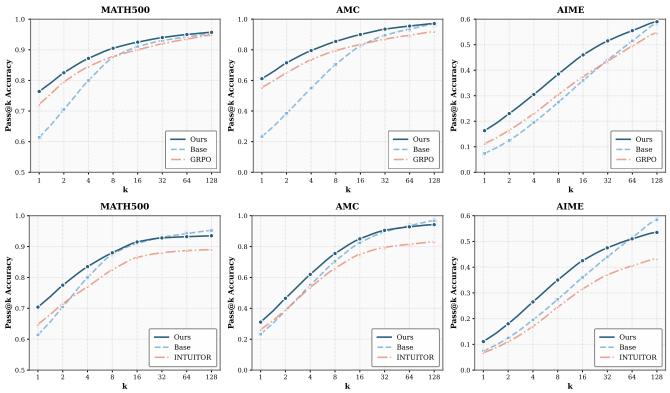


Fig. 3: Pass@k scaling under RLVR (top) and RLIF (bottom).

even at structurally critical branching tokens. This encourages premature convergence toward a limited set of reasoning paths.

The dynamics differ under RLIF. Objectives based on internal confidence inherently promote **entropy reduction**, pushing the policy toward sharper, lower-diversity distributions [8]. Standard RLIF methods such as INTUITOR therefore tend to concentrate probability mass on a limited set of high-certainty trajectories, reducing reasoning diversity and slowing performance growth as k increases. In several settings, the Base model surpasses INTUITOR at large k, indicating reduced reasoning coverage under RLIF. SPINE consistently improves over INTUITOR by counteracting collapse at structurally critical positions. However, because the dominant optimization pressure in RLIF still favors low-entropy regions, SPINE primarily **slows** rather than fully reverses diversity erosion. Thus, SPINE yields better scaling than standard RLIF, but the gains are smaller than those observed under RLVR, highlighting a fundamental difference between the two paradigms.

V. Conclusion

We showed that entropy collapse in RLVR and RLIF is not caused by uniform entropy decay, but by premature overconfidence at a small set of structurally critical forking tokens that determine reasoning branches. These findings identify entropy collapse as a shared structural failure mode of RL-based post-training. This local collapse propagates globally, suppressing exploration during RL training and ultimately constraining test-time scaling at convergence through redundant samples that waste additional compute. To address this, we proposed **SPINE**, which preserves entropy only at the most collapse-prone subset of forking tokens via selective KL penalty, covering $\sim$2–3% of all tokens. Across model sizes and benchmarks, SPINE consistently improves accuracy and Pass@k scaling under both RLVR and RLIF. Ablations show that performance depends more on *where* entropy is preserved than on how much, highlighting structural placement as the key factor.

Acknowledgment

This work was supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) under Grant KRIT-CT-23-021, funded by the Defense Acquisition Program Administration (DAPA).

References

[1] A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney *et al.*, "Openai o1 system card," *arXiv preprint arXiv:2412.16720*, 2024.

[2] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[3] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[4] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto, "s1: Simple test-time scaling," *arXiv preprint arXiv:2501.19393*, 2025.

[5] Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, and P. Liu, "Limo: Less is more for reasoning," *arXiv preprint arXiv:2502.03387*, 2025.

[6] F. Chen, A. Raventos, N. Cheng, S. Ganguli, and S. Druckmann, "Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning," *arXiv preprint arXiv:2502.07154*, 2025.

[7] X. Zhao, Z. Kang, A. Feng, S. Levine, and D. Song, "Learning to reason without external rewards," *arXiv preprint arXiv:2505.19590*, 2025.

[8] Y. Zhang, Z. Zhang, H. Guan, Y. Cheng, Y. Duan, C. Wang, Y. Wang, S. Zheng, and J. He, "No free lunch: Rethinking internal feedback for llm reasoning," *arXiv preprint arXiv:2506.17219*, 2025.

[9] M. Prabhudesai, L. Chen, A. Ippoliti, K. Fragkiadaki, H. Liu, and D. Pathak, "Maximizing confidence alone improves reasoning," *arXiv preprint arXiv:2505.22660*, 2025.

[10] S. Wang, L. Yu, C. Gao, C. Zheng, S. Liu, R. Lu, K. Dang, X. Chen, J. Yang, Z. Zhang *et al.*, "Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning," *arXiv preprint arXiv:2506.01939*, 2025.

[11] Q. Li, R. Xue, J. Wang, M. Zhou, Z. Li, X. Ji, Y. Wang, M. Liu, Z. Yang, M. Qiu *et al.*, "Cure: Critical-token-guided re-concatenation for entropy-collapse prevention," *arXiv preprint arXiv:2508.11016*, 2025.

[12] G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen *et al.*, "The entropy mechanism of reinforcement learning for reasoning language models," *arXiv preprint arXiv:2505.22617*, 2025.

[13] J. Yao, R. Cheng, X. Wu, J. Wu, and K. C. Tan, "Diversity-aware policy optimization for large language model reasoning," *arXiv preprint arXiv:2505.23433*, 2025.

[14] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[15] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt, "Measuring mathematical problem solving with the math dataset," *arXiv preprint arXiv:2103.03874*, 2021.

[16] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[17] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *The Twelfth International Conference on Learning Representations*, 2023.

[18] J. Li, E. Beeching, L. Tunstall, B. Lipkin, R. Soletskyi, S. Huang, K. Rasul, L. Yu, A. Q. Jiang, Z. Shen *et al.*, "Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions," *Hugging Face repository*, vol. 13, no. 9, p. 9, 2024.

[19] S. Agarwal, Z. Zhang, L. Yuan, J. Han, and H. Peng, "The unreasonable effectiveness of entropy minimization in llm reasoning," *arXiv preprint arXiv:2505.15134*, 2025.

[20] Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, S. Song, and G. Huang, "Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?" *arXiv preprint arXiv:2504.13837*, 2025.

[21] A. Karan and Y. Du, "Reasoning with sampling: Your base model is smarter than you think," *arXiv preprint arXiv:2510.14901*, 2025.