ALL WORK AND NO PLAY MAKES JACK A DULL BOY: UNDERSTANDING AND PREVENTING CATASTROPHIC STRATEGY COLLAPSE IN RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

During post-training of large language models (LLMs) with Reinforcement Learning with Verifiable Rewards (RLVR), GRPO-style algorithms can exhibit severe late-stage collapse. Prompt-based probing reveals that this is not benign strategic pruning, but a harmful contraction of effective strategy capacity that makes distinct reasoning strategies increasingly inaccessible. To characterize this phenomenon, we define strategies through trajectory-level policy-update interactions and develop a unified theoretical framework combining optimization dynamics and information theory. We prove that major RLVR objectives progressively concentrate probability mass onto a single strategy, while sustaining nontrivial task accuracy requires a minimum strategy capacity. The conflict between these two results provides a mechanistic explanation for catastrophic collapse. We further derive the Mirrored Entanglement Index (MEI) as a lightweight online warning signal. To prevent collapse, we propose **Mesh Learning**, which exposes multiple reasoning strategies and prevents any single strategy from dominating optimization. Across AIME26, AIME25, MATH-500, GPQA, and LiveCodeBench, Mesh Learning consistently outperforms strong baselines across Qwen and Phi model families, with gains of up to 13.4 pp and 11.5 pp, respectively. These results establish strategy preservation as a key principle for stable RLVR. Code is available at Anonymous GitHub.

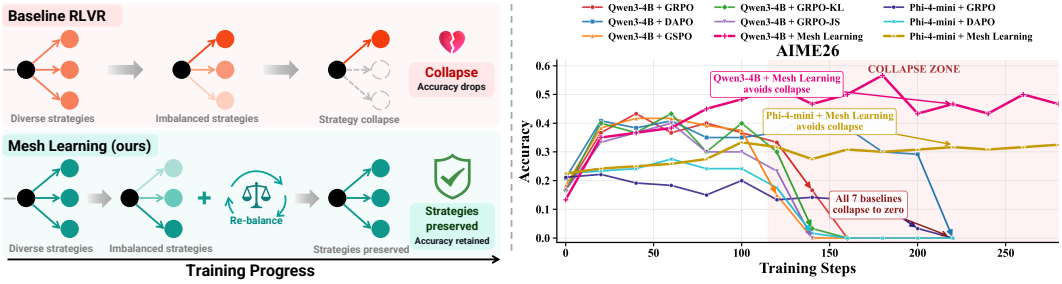


Figure 1: **Left:** Baseline RLVR progressively collapses to a single strategy, while MESH LEARNING preserves strategy diversity through rebalancing. **Right:** On AIME26, all baselines eventually collapse, whereas MESH LEARNING remains stable and accurate across two model families.

1 INTRODUCTION

In recent years, Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful paradigm for improving the complex reasoning capabilities of Large Language Models (LLMs), particularly in mathematics and coding (Guo et al., 2025; Chen et al., 2025; Xu et al., 2026a; Team et al., 2026; Zeng et al., 2026; Xu et al., 2026b). Driven by verifiable outcome supervision and scalable critic-free policy optimization, RLVR has been rapidly adopted in both academia and industry, giving rise to a range of optimization frameworks, including GRPO (Guo et al., 2025), DAPO (Yu

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2025), and GSPO (Zheng et al., 2025), as well as regularization schemes based on KL divergence (Shao et al., 2024) and JS divergence (Li et al., 2026). Together, these developments have established RLVR as a core post-training paradigm for modern reasoning models.

Despite its effectiveness in improving reasoning performance, RLVR can exhibit severe late-stage instability. Prior studies have reported similar collapse behaviors (Guo et al., 2025; Zheng et al., 2025; Gao et al., 2025; Li et al., 2026), and we observe the same phenomenon systematically in our experiments. As shown in Fig. 1, accuracy drops abruptly after a period of seemingly successful optimization across two model families, Qwen (Yang et al., 2025) and Phi (Abdin et al., 2025), and five representative training recipes, including GRPO, DAPO, GSPO, GRPO+KL, and GRPO+JS. We call this abrupt degradation **collapse**. Its recurrence across prior studies, model families, algorithms, and regularization schemes suggests a shared limitation of current RLVR optimization, yet its underlying mechanism remains poorly understood.

To characterize this behavior across RLVR algorithms, we develop a unified theory grounded in the optimization dynamics of policy updates. Since individual trajectories rarely recur exactly and are too fine-grained to describe reasoning behavior, we introduce *strategies* as groups of trajectories that evolve coherently under parameter updates, and derive their distribution updates from the underlying trajectory-level updates. Within this framework, we explain collapse by contradiction through two central theorems: **1)** If RLVR objectives (GRPO, DAPO, GSPO) persist in a non-collapse regime, the optimization inherently drives the policy toward a single dominant strategy with high probability; **2)** achieving nontrivial task accuracy requires the model to retain a minimum strategy capacity. Together, these results reveal a fundamental conflict: RLVR optimization progressively contracts strategy capacity, while successful reasoning requires sufficient capacity to remain available. This contradiction provides a mechanistic explanation for the catastrophic collapse observed in practice.

This mechanistic view also suggests what to monitor before accuracy collapses: the latent concentration of strategy dynamics. Since this transition is difficult to observe from standard training signals, we derive the **Mirrored Entanglement Index (MEI)** from the geometry of RLVR updates as a lightweight online proxy for strategy collapse, requiring no auxiliary training and little computational overhead. Calibrating MEI on pre-RLVR models yields a 3σ alarm threshold for detecting departure from the non-collapsed regime before the accuracy cliff. Across model families, RLVR algorithms, and regularization schemes, MEI rises above the threshold ahead of the eventual accuracy cliff, consistently tracking the theoretically predicted transition. Independent strategy-level analysis further confirms that this transition corresponds to the loss of distinct reasoning strategies, supporting strategy collapse as a shared mechanism underlying late-stage RLVR failure.

Building on the theoretical characterization and online measurability of strategy collapse, we propose **Mesh Learning**, a strategy-preserving RLVR framework that prevents individual strategies from dominating optimization and preserves multiple viable reasoning strategies throughout training. It consists of two complementary components: **1) Coach Prompting (CP)**, which augments each query with strategy prefixes that elicit distinct solution approaches and ensure broad strategy exposure; and **2) Strategy-Balancing Regularization**, which balances the learning progress across strategies to prevent any single strategy from overwhelming the others. In this way, CP constructs a diverse strategy mesh, while the regularizer preserves it throughout optimization. Across AIME26, AIME25, MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023), GPQA (Rein et al., 2024), and LiveCodeBench (v6) (Jain et al., 2024), Mesh Learning consistently outperforms strong RLVR and divergence-regularized baselines across both Qwen and Phi families, with gains of up to 13.4 and 11.5 percentage points (pp), respectively. These results highlight strategy preservation as a key principle for stable and effective RLVR optimization. In summary, our contributions are as follows:

❶ We establish a unified theory of RLVR dynamics, showing that inherent competition drives strategy concentration and explaining collapse through its conflict with the minimum strategy capacity required for successful reasoning. The analysis also yields MEI as a lightweight online indicator.

❷ We propose **Mesh Learning**, combining Coach Prompting and Strategy-Balancing Regularization to expose diverse reasoning strategies and prevent any strategy from dominating optimization.

❸ We provide extensive empirical validation across RLVR recipes and model families on AIME25/26, MATH-500, GPQA-Diamond, and LiveCodeBench, confirming the predicted collapse dynamics and the effectiveness of Mesh Learning, with gains of up to 13.4 pp on Qwen models and 11.5 pp on Phi models.

2 FROM TRAJECTORY TO STRATEGY

Since the exact recurrence of an individual trajectory is exceedingly unlikely, analyzing isolated trajectories provides limited insight into training dynamics. To study these dynamics more effectively, we introduce the notion of a ***strategy***, which captures a group of trajectories that respond coherently to local policy updates, i.e., they tend to be reinforced or suppressed together.

2.1 THE CONCEPT OF STRATEGY

Consider a prompt Q and a policy π_θ parameterized by θ. For a query Q, let its admissible trajectory space be $\mathcal{T}_Q = \{\tau_1, \tau_2, \ldots\}$. We first introduce the **Fisher Score** of a trajectory τ_i: $U_i = \nabla_\theta \log \pi_\theta(\tau_i \mid Q)$. Then consider an infinitesimal update induced by another trajectory τ_j: $\Delta\theta = \eta A_j U_j$, where η is the learning rate and A_j denotes the scalar reward signal associated with τ_j. By first-order approximation, the log-likelihood of τ_i changes by $\Delta \log \pi_\theta(\tau_i \mid Q) \approx \eta A_j U_i^\top U_j$.

Hence, we define the **trajectory coupling kernel** $K_{ij} \triangleq U_i^\top U_j$, which characterizes the first-order interaction between trajectories. For a positive update on τ_j, $K_{ij} > 0$ reinforces τ_i, whereas $K_{ij} < 0$ suppresses it, and $K_{ij} \approx 0$ indicates that the two trajectories are locally decoupled.

Intuitively, trajectories belonging to the same strategy should not only be positively coupled, but also respond similarly to updates induced by other trajectories. This suggests a natural partitioning based on pairwise similarity $\sup_k |K_{ik} - K_{jk}| \le \epsilon$. However, pairwise similarity alone is not transitive and therefore does not directly induce a valid partition. To make this intuition precise, we define **strategy** in Def. 1 through the neighborhoods of the examined trajectories.

Definition 1 (Strategy)**.**
Given a prompt Q, policy π_θ, and examined trajectories $\{\tau_1, \ldots, \tau_n\}$, let $K_{ij} = U_i^\top U_j$. For threshold ϵ, define the closed neighborhood of τ_i as

$$\mathcal{N}_i(\epsilon) := \{\tau_i\} \cup \left\{ \tau_j \mid \tau_j \neq \tau_i,\ K_{ij} > 0,\ \sup_k |K_{ik} - K_{jk}| \le \epsilon \cdot \sup_{a,b} |K_{ab}| \right\}. \quad (1)$$

Here, $K_{ij} > 0$ requires τ_i and τ_j to be positively coupled, while $\sup_k |K_{ik} - K_{jk}| \le \epsilon \cdot \sup_{a,b} |K_{ab}|$ further requires them to have similar coupling profiles with the examined trajectories. We say τ_i and τ_j are dynamically equivalent, denoted by $\tau_i \sim_\epsilon \tau_j$, if

$$\tau_i \sim_\epsilon \tau_j \iff \mathcal{N}_i(\epsilon) = \mathcal{N}_j(\epsilon). \quad (2)$$

*A **Strategy** is an equivalence class induced by $\sim_\epsilon$, with ϵ defined as the **sensitivity**.*

2.2 THE SEMANTIC PROXY OF STRATEGY

Def. 1 is deliberately optimizer-centric, grouping trajectories solely by their interactions under policy updates without semantic assumptions. We next investigate whether these strategies align with human-recognizable solution methods, enabling both interpretation and explicit control.

Using correct trajectories generated by Qwen-2.5-7B-Instruct on MATH-500, we categorize their solution methods with DeepSeek-V4-Pro and compare the coupling kernel K_{ij}. With $\epsilon = 0.05$, the resulting strategy partition empirically aligns with distinct solution methods. Fig. 2 (a) illustrates that the same problem can admit semantically distinct strategies, while Fig. 2 (b) shows at scale that within-method trajectory pairs are substantially more positively coupled than cross-method pairs, which are weakly or negatively coupled. This correspondence supports the semantic interpretability of our dynamically defined **strategies** and allows solution-method prefixes to serve as practical proxies for latent strategies. We use these prefixes for strategy elicitation in Sec. 3.2 and strategy-conditioned training in Sec. 4.1.

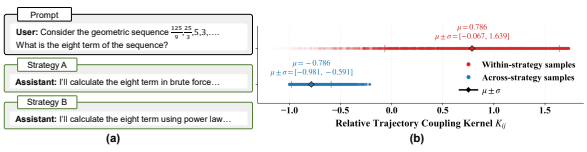


Figure 2: Semantic interpretability of strategies. (a) Distinct solution methods for the same problem. (b) On MATH-500, within-method trajectory pairs are substantially more coupled than cross-method pairs, supporting Def. 1.

3 THEORETICAL ANALYSIS

In this section, we prove the RLVR collapse mechanism by contradiction: **Theorem 1** shows that if the non-collapse regime persists, GRPO, DAPO, and GSPO reduce strategy capacity; **Theorem 2** yields a minimum strategy capacity $|\mathcal{S}_{\epsilon,k}|$ for nontrivial task accuracy. These consequences eventually conflict, causing an accuracy cliff. Then, we derive a **lightweight online metric** that alerts collapse via a 3σ threshold from non-collapsed training. Finally, **Theorem 3** builds an effective strategy-level RL system, explains regularization failures, and motivates our prevention in the next section. The main notation used throughout our theoretical analysis is summarized in Table 1.

Table 1: Core notation used in Sec. 3.

Symbol	Meaning	Symbol	Meaning
$Q, \mathcal{D}, \mathcal{D}', \mathcal{H}$	Query; test/training set; distribution	$S^{(\epsilon)}_{\alpha,Q}$	Strategy α in Q's step-0 ϵ-partition
$k, \pi_k, \pi_{old,k}$	Step; step-k current/old policy	$\nu^{(\epsilon)}_{\alpha,Q,k}$	Probability mass of $S^{(\epsilon)}_{\alpha,Q}$ under π_k
$Y(Q), \widehat{Y}(\tau)$	ground truth/ prediction of trajectory τ	$P_{acc,k}$	Model's step-k test set accuracy
$\epsilon, m_{\epsilon,Q}$	Sensitivity in Def. 1; step-0 strategy amount for Q under sensitivity ϵ	$\lvert\mathcal{S}_{\epsilon,k}\rvert$	Model's total number of usable strategies on $\mathcal{D}$ at step k

3.1 THE MAIN THEOREM OF STRATEGY COLLAPSE

If RLVR remains in **non-collapse** regime, where test accuracy stays nontrivial, it induces competition among strategies. We show that this competition inherently concentrates their probability mass.

Theorem 1 (The Main Theorem of Strategy Collapse)**.**
Suppose $\mathcal{D}, \mathcal{D}' \overset{\text{i.i.d.}}{\sim} \mathcal{H}$, with $|\mathcal{D}|, |\mathcal{D}'| \gg 1$, and GRPO, GSPO, or DAPO is optimized on $\mathcal{D}'$ by gradient descent with step-wise old-policy refresh $\pi_{\text{old},k} = \pi_{k-1}$. For $Q \in \mathcal{D}$, let η, $|\tau|$, and ϵ_{clip} denote the learning rate, trajectory length, and clipping ratio. Define $T_c \triangleq \lfloor \eta^{-1} \rfloor$. If $P_{acc,k} \in (0,1)$, for all $k \le T_c$, and $\eta \ll \epsilon_{\text{clip}} < 1 \ll |\tau| \ll \eta^{-1} \ll |\tau|^3$, then

$$\forall \epsilon > 0, \epsilon \ll 1, \zeta, \delta \in (\eta, 1), \zeta^2 \gg \eta \quad \exists K \in \mathbb{Z}^+, K \ll |\tau| < T_c,$$
$$\forall k \in \mathbb{Z}^+, K < k \le T_c, \quad \mathbb{P}\left(\max_{1 \le \alpha \le m_{\epsilon,Q}} \nu^{(\epsilon)}_{\alpha,Q,k} \ge 1 - \zeta \right) \ge 1 - \delta. \tag{3}$$

The probability is taken over the stochasticity of RLVR training on $\mathcal{D}'$.

Theorem 1 shows that, after burn-in time K, RLVR concentrates at least $1-\zeta$ of the strategy mass for Q on a single strategy throughout $K < k \le T_c$ with probability at least $1-\delta$, thereby **reducing effective strategy capacity**. The scale condition $\eta \ll \epsilon_{\text{clip}} < 1 \ll |\tau| \ll \eta^{-1} \ll |\tau|^3$ matches standard RLVR configurations and experiments (Sec. 5.1); the proof is in Appendix A.2. Since directly tracking $\nu^{(\epsilon)}_{\alpha,Q,k}$ is impractical, Sec. 3.3 derives MEI as a lightweight observable.

3.2 POWER LAW FOR STRATEGY CAPACITY LOWER BOUND

Assuming the non-collapse regime persists, **Theorem 1** implies concentration. We further derive a lower bound on usable strategy capacity for nontrivial accuracy, thus contradicting this assumption.

Theorem 2 (Power Law for Strategy Capacity Lower Bound)**.**
For $\mathcal{D}, \mathcal{D}' \overset{\text{i.i.d.}}{\sim} \mathcal{H}$, $|\mathcal{D}| = N$. Suppose $\mathbb{P}_{Q\sim\mathcal{D}}\left[\exists Q' \ne Q, \exists \tau \in S^{(\epsilon)}_{\cdot,Q} : \widehat{Y}(\tau) = Y(Q')\right] \le \gamma \ll 1$, requiring a diverse answer space. Fix $\rho, \epsilon \in (0,1)$ such that $\min_{Q\in\mathcal{D},\alpha} \nu^{(\epsilon)}_{\alpha,Q,0} \ge \rho$, (initially non-collapsed). Let $m_\epsilon \triangleq \min_{Q\in\mathcal{D}} m_{\epsilon,Q}$, then for $0 < k \ll |\tau|$ under RLVR optimization on $\mathcal{D}'$:

$$|\mathcal{S}_{\epsilon,k}| = \Omega_{\rho,\gamma}\left(m_\epsilon N^{P_{acc,k}-\gamma}\right). \tag{4}$$

Theorem 2 shows higher accuracy requires a **larger usable strategy capacity**. With **Theorem 1**, this explains accuracy collapse: RLVR contracts strategy capacity until violating the bound, making

performance unsustainable. Collapsing m_ϵ strategies to one removes entropy $\Delta H_{\text{req}} \approx \log m_\epsilon$, which exceeds budget $\Delta H_{\text{ad}} = o(\log m_\epsilon)$, thus breaking the non-collapse regime. As Theorem 1 holds with high probability, occasional non-collapse runs are expected. See details in Appendix A.3.

For empirical validation, we use a conservative estimate $\widehat{m} = 4$, obtained by asking DeepSeek-V4-Pro to enumerate distinct solution methods for 1,000 MATH problems and taking the minimum across prompts. This deliberately avoids overestimating m, ensuring that the empirical test does not rely on an inflated estimate of task strategy diversity. We then evaluate the resulting bound on models from 1.7B to 72B parameters. As shown in Fig. 3, all tested models satisfy the predicted lower bound on strategy capacity.

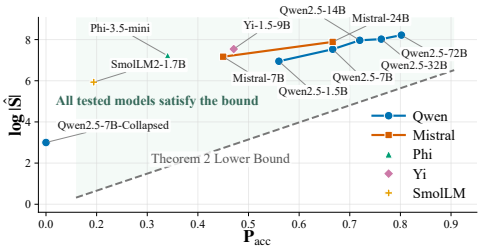


Figure 3: Empirical examination of Theorem 2. All tested models satisfy the lower bound.

3.3 How to Measure and Forecast Collapse

The preceding analysis identifies strategy concentration as the latent transition underlying accuracy collapse, but directly tracking strategy partitions or parameter-space gradients during training is impractical. We therefore seek a lightweight observable of this transition. In Appendix B, we prove that as collapse approaches, both the parameter-space log-gradients $g_i = \nabla_\theta \log \pi_\theta(\tau_i \mid Q)$ and the logit gradients $v_i = \nabla_z \log \pi(\tau_i \mid Q)$ contract, i.e., $\|g_i - g_j\| \to 0$ and $\|v_i - v_j\| \to 0$. To reduce memory overhead, we construct our monitor from v_i, which can be obtained directly from rollout logits. This yields a lightweight online observable, the **Mirrored Entanglement Index (MEI)**:

Definition 2 (Mirrored Entanglement Index (MEI)). *MEI* $= \| \sum_{i=1}^{G} v_i \|^2 / \sum_{i=1}^{G} \|v_i\|^2$.

Equivalently, $\text{MEI} = G - \sum_{i<j} \|v_i - v_j\|^2 / \sum_i \|v_i\|^2 = 1 + 2 \sum_{i<j} v_i^\top v_j / \sum_i \|v_i\|^2$. The first form shows that MEI increases as $\|v_i - v_j\| \to 0$, while the second gives a baseline of 1 for weakly coupled trajectories. Appendix B establishes that increasing MEI reflects v-concentration, which implies g-concentration and departure from the non-collapsed regime. Calibrating on pre-RLVR Qwen2.5-7B-Instruct and Qwen3-4B across 3 datasets yields the 3σ threshold $\text{MEI} = 1.013$.

Fig. 4 illustrates this behavior in GRPO training. MEI crosses the 3σ threshold early in training, before its final surge and the subsequent accuracy collapse. This provides a lightweight online signal for detecting departure from the non-collapsed regime, without explicitly tracking parameter-space gradients during optimization.

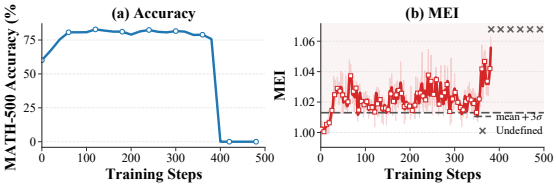


Figure 4: MEI as an alert signal of collapse.

3.4 Why Divergences Fail

The preceding results identify collapse as intrinsic to pure RLVR. Can a fixed divergence penalty prevent it? **Theorem 3** says no. Using a **derived strategy-level RL system**, we show sufficiently large reward gaps can drive KL- or JS-regularized optima arbitrarily close to a single strategy.

Theorem 3 (Failure of Divergence Regularization)**.** *Under **Theorem 1**, within step $k \ll |\tau|$, RLVR admits a static strategy-level optimization $\nu^D \triangleq \arg\max_\nu \{\mathbb{E}_\nu[R] - \beta D\}$, where $R(S)$ is the effective static reward, $\beta > 0$ the regularization strength, and D be KL or JS divergence with full-support ν^{ref}. Let $\Delta_R = R_{(1)} - R_{(2)}$ be the best-to-second-best reward gap, then:*

$$\forall \epsilon > 0, \exists C_D < \infty \text{ such that } \Delta_R/\beta \geq C_D \implies \max_i \nu_i^D \geq 1 - \epsilon. \tag{5}$$

Theorem 3 shows that if Δ_R/β exceeds a finite threshold C_D, the optimal distribution assigns at least $1 - \epsilon$ mass to one best strategy for any $\epsilon > 0$. Thus, fixed divergence penalties cannot universally prevent strategy concentration when reward dominates regularization. See details in Appendix C.1.

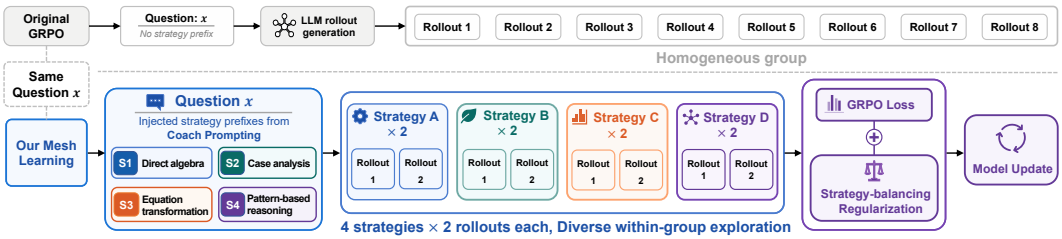


Figure 5: Mesh Learning versus standard GRPO. Mesh Learning elicits multiple strategies and balances their relative growth during optimization.

4 Methodology: Mesh Learning

Sec. 3 shows that preventing RLVR collapse requires preserving sufficient strategy capacity. This creates two challenges: training rollouts must elicit distinct strategies, and no single strategy should dominate optimization. We therefore propose **Mesh Learning** with two components: **1) Coach Prompting**, which elicits multiple reasoning strategies for each query, and **2) Strategy-Balancing Regularization**, which prevents any strategy from dominating optimization. Under the conditions analyzed in Appendix C.2, we further prove that the resulting dynamics converge to equilibrium retaining nonzero probability mass for every strategy. The pipeline of Mesh Learning is shown in Fig. 5, it replaces homogeneous GRPO rollouts with multi-strategy exploration and jointly optimizes GRPO with Strategy-Balancing Regularization.

4.1 Strategy Elicitation: Coach Prompting

Sec. 2.2 empirically shows that trajectories using the same human-recognizable solution method exhibit similar coupling patterns, motivating solution-method prefixes as practical proxies for the latent strategies in Def. 1. We therefore introduce ***Coach Prompting*** **(CP)** to expose multiple strategies for each training query. For each query Q, an offline Coach LLM generates m semantically distinct strategy prefixes, each specifying a candidate method, optionally with a brief description, but excluding calculations, intermediate solutions, and final answers. Each prefix conditions a separate rollout group, ensuring explicit exploration of all m strategies rather than relying on spontaneous policy discovery. These prefixes are generated once and fixed throughout training, incurring no online Coach cost.

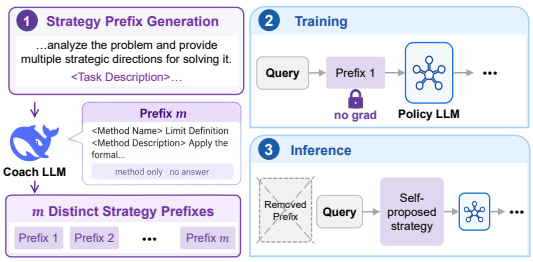


Figure 6: Coach Prompting. An offline Coach LLM generates fixed strategy prefixes for training; at inference, the policy proposes its own strategy and follows the same verification procedure.

During training, each prefix is treated as a candidate rather than a fixed instruction: the policy assesses its applicability, underlying mechanism, and alternatives before answering. At inference, no pre-generated prefix is provided; the policy proposes a candidate strategy itself and follows the same verification procedure, requiring no Coach LLM at test time. Coach-provided tokens are masked from the training loss. Full prompts and implementation details are provided in Appendix D.1.

4.2 Strategy-Balancing Regularization

We seek to prevent strategy elimination without hindering their joint improvement. For strategy S_i, let $\mathcal{C}_i \subset S_i$ denote its correct trajectories, and define $z_i = \frac{1}{|\mathcal{C}_i|} \sum_{\tau \in \mathcal{C}_i} \log \pi_\theta(\tau \mid Q)$ as a differentiable surrogate for its current support under the policy. Let $\mathbf{z} = (z_1, \ldots, z_m)$ and let $\mathcal{U}_m$ denote the uniform distribution over the m strategies. Directly minimizing $D_{\mathrm{KL}}(\mathcal{U}_m \| \mathrm{softmax}(\mathbf{z}))$ is biased by trajectory length, since longer trajectories accumulate more negative log-probability. Conversely, length normalization $z_i/|\tau_i|$ admits a padding shortcut, as high-probability filler tokens can increase

the normalized score without actually strengthening the underlying strategy. We therefore measure each strategy's relative growth against a fixed pre-RLVR reference policy π_{ref}.

Definition 3 (Strategy-Balancing Regularization)**.** *Let* $z_i^{\text{ref}} = \frac{1}{|\mathcal{C}_i|}\sum_{\tau\in\mathcal{C}_i}\log\pi_{ref}(\tau \mid Q)$ *denote the strategy scores under* π_{ref}*. We define*

$$\mathcal{J}_S = D_{\text{KL}}\left(\mathcal{U}_m \| \text{softmax}(\mathbf{z} - \mathbf{z}^{\text{ref}})\right), \tag{6}$$

and objective $\mathcal{J} = \mathcal{J}_{\text{RLVR}} - \mu\mathcal{J}_S$*. The three forms of* $\mathcal{J}_{\text{RLVR}}$ *are provided in Appendix A.1.*

Let $\mathbf{p} = \text{softmax}(\mathbf{z} - \mathbf{z}^{\text{ref}})$. Then $\nabla_{\mathbf{z}}\mathcal{J}_S = \mathbf{p} - \frac{1}{m}\mathbf{1}$, which suppresses relatively fast-growing strategies while reinforcing lagging ones. Meanwhile, $\mathcal{J}_S(\mathbf{z} + c\mathbf{1}) = \mathcal{J}_S(\mathbf{z})$, so the regularizer controls only relative strategy growth without opposing their common improvement. In Appendix C.2, we further prove that the resulting dynamics converge to existing equilibrium where every strategy retains nonzero probability, while preserving within-strategy trajectory competition to first order.

5 Experiments

5.1 Experimental Setup

Models and Data. We run full experiments on Qwen3-4B and Qwen2.5-7B-Instruct, with additional cross-family experiments on Phi-4-mini-reasoning. Qwen3-4B and Phi-4-mini-reasoning use thinking mode, while Qwen2.5-7B-Instruct uses standard instruction-tuned generation. For MATH-500, we train on the 7,500-problem Hendrycks-MATH training split (Hendrycks et al., 2021); for AIME25/26 and GPQA-Diamond, we use 7,000 DAPOMath (Yu et al., 2025) and 4,500 Wild-Sci (Liu et al., 2026) problems, respectively; and for LiveCodeBench, we use 4,500 Skywork-OR1-RL-Data problems (He et al., 2025). For Coach Prompting, each query is augmented with $m = 4$ distinct strategy prefixes generated offline by DeepSeek-V4-Pro and fixed throughout training.

Training Configuration. Mesh Learning is implemented on top of GRPO using VeRL (Sheng et al., 2024), with the clip ratio set to 0.2. For DAPO and GSPO, we follow the clip ratio configurations recommended in their original papers, while keeping all other hyperparameters identical across baselines. We use 16 rollouts per query and set the Strategy-Balancing Regularization weight to $\mu = 10^{-5}$. Remaining configurations and hyperparameters are provided in Appendix D.2.

Evaluation and Baselines. We evaluate on AIME25/26, MATH-500 (Hendrycks et al., 2021), GPQA-Diamond (Rein et al., 2024), and LiveCodeBench (v6) (Jain et al., 2024), covering mathematical, scientific, and code reasoning. We compare Mesh Learning with GRPO, DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025), and GRPO with KL or JS regularization (Li et al., 2026).

5.2 Main Results

Overall Performance. Table 2 summarizes results across two models and four benchmarks. We highlight two observations: **1) Mesh Learning consistently improves reasoning performance.** With the default $m = 4$, Mesh Learning achieves the best performance across all eight model–benchmark settings. Compared with the strongest RLVR baseline in each setting, Mesh Learning improves Qwen2.5-7B-Instruct by 1.6–3.4 pp and Qwen3-4B by 2.1–13.4 pp. The largest gain occurs on AIME26 with Qwen3-4B, improving accuracy from 43.3% to 56.7%. **2) We observe consistent gains across tasks and model scales.** Mesh Learning improves challenging competition-level benchmarks (AIME25/26) and broader mathematical and scientific reasoning benchmarks (MATH-500 and GPQA-Diamond). These gains hold across Qwen2.5-7B-Instruct and Qwen3-4B, indicating that the benefit is not specific to a single model or task. Trajectory-level examples are provided in Appendix E, while ablations on Coach Prompting and strategy number m are presented in Sec. 5.3.

Training Dynamics. Fig. 7 shows that several RLVR baselines improve but later suffer abrupt accuracy collapse, accompanied by a sharp rise in MEI beyond the 3σ threshold, while Mesh Learning keeps MEI low and consistently attains higher accuracy. GPQA is an exception: four choices imply $\gamma \approx 1$, violating Theorem 2's condition, so strategy concentration need not induce collapse.

Capability Frontier. To examine whether these gains persist under larger inference budgets, we report Pass@k on AIME25 and AIME26 for $k \in \{2, 8, 32, 128\}$. As shown in Fig. 8, Mesh Learning

Table 2: Main results and controlled ablations (Mean@4 Accuracy %). CP variants isolate the effect of Coach Prompting, while $m = 3$ and $m = 4$ compare different strategy capacities.

Method	Qwen-2.5-7B-Instruct				Qwen-3-4B			
	AIME26	AIME25	MATH-500	GPQA	AIME26	AIME25	MATH-500	GPQA
Standard RLVR Training								
Baseline	3.3	5.0	64.8	13.7	13.3	21.6	62.5	13.7
GRPO	9.2	13.3	69.6	33.4	43.3	40.0	81.6	38.9
DAPO	8.3	6.7	70.5	34.1	43.3	40.8	84.3	44.5
GSPO	9.2	5.8	71.4	34.3	41.6	38.3	79.8	45.8
GRPO + KL	9.2	13.3	71.8	31.6	43.3	40.0	82.2	42.3
GRPO + JS	10.0	6.4	71.4	28.5	40.0	43.3	81.8	44.8
Controlled with Coach Prompting (CP)								
GRPO + CP	10.0	13.3	69.6	31.7	43.3	40.0	81.6	37.8
DAPO + CP	7.5	5.8	69.7	26.3	36.6	33.3	84.5	42.3
GSPO + CP	6.7	8.3	68.5	32.1	41.7	33.3	85.3	43.7
GRPO + KL + CP	4.4	8.3	71.8	31.9	40.4	38.3	78.5	33.2
GRPO + JS + CP	10.0	7.8	72.0	30.7	40.0	41.4	80.4	30.7
Mesh Learning ($m = 3$)	10.3	16.3	72.4	36.7	52.5	43.7	84.8	44.6
Mesh Learning ($m = 4$)	**13.3**	**16.7**	**73.4**	**36.7**	**56.7**	**48.3**	**86.8**	**47.9**

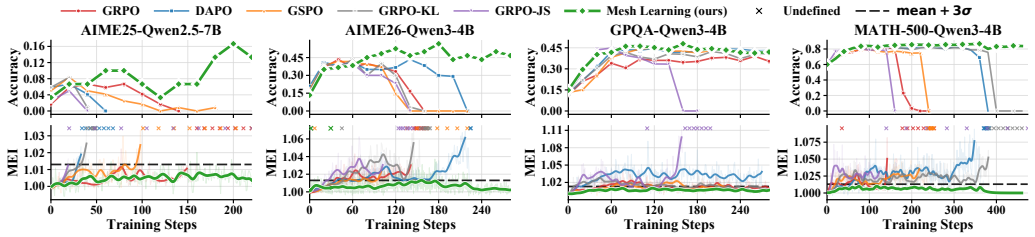


Figure 7: Training dynamics of accuracy (top) and MEI (bottom). Baselines can exhibit late-stage collapse with rising MEI, while Mesh Learning maintains stable accuracy and low MEI.

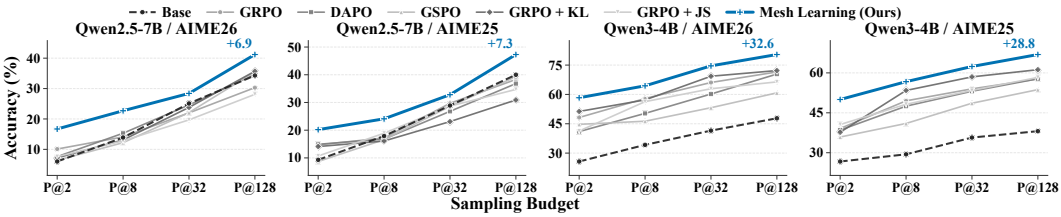


Figure 8: Pass@k performance on AIME25 and AIME26. Across increasing inference budgets, Mesh Learning consistently outperforms all baselines on both models.

consistently outperforms all baselines across both models, with its advantage remaining substantial as k increases. In particular, the gains at Pass@128 reach 6.9 and 7.3 pp for Qwen2.5-7B and 32.6 and 28.8 pp for Qwen3-4B on AIME26 and AIME25, respectively. These results indicate that Mesh Learning improves not only single-budget accuracy but also the attainable performance under substantially larger sampling budgets.

Cross Family Generalization. Beyond Qwen3-4B and Qwen2.5-7B-Instruct, we further evaluate Mesh Learning on Phi-4-mini-reasoning to examine transfer across model families. As shown in Table 3, Mesh Learning consistently outperforms both GRPO and DAPO across all four benchmarks, achieving gains over the

Table 3: Cross-family generalization on Phi-4-mini-reasoning (Mean@4 Accuracy %).

Method	AIME26	AIME25	MATH-500	GPQA
GRPO	22.3	24.1	79.9	18.4
DAPO	25.0	23.4	77.3	23.0
Mesh Learning	**33.3**	**30.8**	**83.2**	**34.5**

stronger baseline of 8.3, 6.7, 3.3, and 11.5 pp on AIME26, AIME25, MATH-500, and GPQA-Diamond, respectively. These results provide additional evidence that the effectiveness of Mesh Learning is not specific to the Qwen family.

Generalization to Coding Tasks. To evaluate generalization beyond math reasoning, we extend experiments to LiveCodeBench v6. Table 4 shows that Mesh Learning reaches 56.4% on Qwen3-4B and 36.3% on Phi-4-mini-reasoning, outperforming DAPO by 4.2 pp and 3.6 pp, suggesting that its benefits transfer to coding tasks.

Table 4: LiveCodeBench v6 (Mean@4 Accuracy %)

Model	Baseline	GRPO	DAPO	Mesh Learning
Qwen-3-4B	30.5	48.4	52.2	**56.4**
Phi-4-mini-reasoning	24.8	31.5	32.7	**36.3**

5.3 Ablation Studies

Ablation for Coach-Prompting Priors. Since Coach Prompting provides extra strategy information, we compare each RLVR baseline with a CP-only counterpart using the same prefixes but without Strategy-Balancing Regularization. As shown in Table 2, CP alone brings little or no improvement, while full Mesh Learning consistently performs better. This indicates that the gains do not come from stronger strategy priors alone, but from preserving these strategies during optimization.

Number of Strategies. As shown in Table 2, increasing m from 3 to 4 leads to only small differences overall, showing that Mesh Learning is not sensitive to modest changes in m. However, we found that an overly small m limits the number of strategies exposed during training and therefore restricts the strategy capacity that can be preserved. We use $m = 4$ by default.

6 Related Work

RLVR optimization. Following GRPO (Shao et al., 2024), substantial work has improved RLVR through better sampling, clipping, importance weighting, and optimization objectives, including DAPO (Yu et al., 2025), CISPO (Chen et al., 2025), GSPO (Zheng et al., 2025), SAPO (Gao et al., 2025), Dr. GRPO (Liu et al., 2025), VAPO (Yue et al., 2025), REINFORCE++ (Hu et al., 2025), GMPO (Zhao et al., 2026) and FAPO (Ding et al., 2026). These methods primarily improve RLVR's optimization behavior, whereas we focus on the mechanism underlying its late-stage collapse.

Understanding and Preventing Collapse. Recent work has documented collapse during prolonged RLVR training. Liang et al. (2026) show that Pass@1 improvements can coincide with reduced generation diversity and degraded high-k Pass@k, while Nguyen et al. (2025) attribute reasoning-boundary shrinkage to negative interference and winner-take-all dynamics. Yuan et al. (2026) link diversity collapse to continued optimization on already-saturated problems, and M-GRPO (Bai et al., 2025) studies abrupt performance collapse under long-horizon training. A common remedy is to constrain policy drift with fixed-reference KL regularization (Shao et al., 2024), while alternative divergences such as forward KL and JS have been shown to better preserve solution diversity (Li et al., 2026). In contrast, we provide a strategy-level theoretical account: starting from trajectory-level update interactions, we derive RLVR's tendency toward strategy concentration and a complementary lower bound on the strategy capacity required for nontrivial task performance. Their tension directly explains catastrophic collapse and motivates Mesh Learning.

7 Conclusion

This work characterizes, anticipates, and mitigates catastrophic collapse in RLVR. We develop a strategy-level theory showing that RLVR progressively concentrates strategy mass, while successful reasoning requires sufficient strategy capacity; their conflict results collapse. This analysis further yields MEI as a lightweight online warning signal. We propose **Mesh Learning**, combining Coach Prompting and Strategy-Balancing Regularization to expose and preserve diverse strategies during optimization. Experiments across AIME26, AIME25, MATH-500, GPQA, and LiveCodeBench consistently demonstrate its effectiveness across Qwen and Phi model families. In conclusion, **All Work And No Play Makes Jack A Dull Boy**. RLVR may trade away strategy diversity for continued gains in reward, eventually exhausting the capacity required for further meaningful optimization. Our results highlight strategy preservation as a key principle for stable and effective RLVR.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, code implementation, and the refinement and correctness checking of theorem proofs. Generative AI tools were not used to independently write the manuscript from scratch or to generate citations or bibliographic references.

All AI-assisted code was reviewed, verified, and tested by the authors. All theorem statements and proofs, including those for which generative AI tools were used as an auxiliary correctness check, were independently examined and verified by the authors. No citations were generated by generative AI tools, and all references included in the paper were manually checked by the authors to confirm their authenticity and relevance.

We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, proofs, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code and scripts are available in our anonymous GitHub repository at Anonymous GitHub. All datasets required for training, including those processed with our Coach Prompting procedure, are provided in the supplementary materials as data.zip, except for the coding dataset, which is too large to include in the submission.

REFERENCES

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

Bizhe Bai, Hongming Wu, Peng Ye, and Tao Chen. M-grpo: Stabilizing self-supervised reinforcement learning for large language models with momentum-anchored policy optimization. *arXiv preprint arXiv:2512.13070*, 2025.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Yuyang Ding, Chi Zhang, Juntao Li, Haibin Lin, Xin Liu, and Min Zhang. Fapo: Flawed-aware policy optimization for efficient and reliable reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 157026–157050, 2026.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

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

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems*, 2021.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, et al. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. In *International Conference on Learning Representations*, volume 2026, pp. 64742–64768, 2026.

Xiao Liang, Zhong-Zhi Li, Yeyun Gong, yelong shen, Yingnian Wu, Zhijiang Guo, and Weizhu Chen. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 79443–79491, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/805cd624d38e39ca061e8fd431e20825-Paper-Conference.pdf.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Tengxiao Liu, Deepak Nathani, Zekun Li, Kevin Yang, and William Yang Wang. Wildsci: Advancing scientific reasoning from in-the-wild literature. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11677–11695, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Phuc Minh Nguyen, Chinh D La, Duy MH Nguyen, Nitesh V Chawla, Binh T Nguyen, and Khoa D Doan. The reasoning boundary paradox: How reinforcement learning constrains language models. *arXiv preprint arXiv:2510.02230*, 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.

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

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.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Tianshi Xu, Yuteng Chen, and Meng Li. Cleaner: Self-purified trajectories boost agentic reinforcement learning. *arXiv preprint arXiv:2601.15141*, 2026b.

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

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

Suqin Yuan, Jinkun Chen, Jiyang Zheng, Muyang Li, Lei Feng, Dadong Wang, Tao Xiang, Tongliang Liu, and Bo An. Understanding diversity collapse in rlvr via the lens of overtraining. *arXiv preprint arXiv:2606.15455*, 2026.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, xun wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 41354–41374, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/44a1f18afd6d5cc34d7e5c3d8a80f63b-Paper-Conference.pdf.

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

A Appendix A: Math Proof On The Collapse Of RLVR

A.1 Definitions and Mathematical Framework

A.1.1 Notations

Back To formalize collapse in Group Relative Policy Optimization (GRPO), we use the following notation:

- π: model policy; θ: model parameters.
- $\mathcal{D}$: test dataset with prompts $Q \in \mathcal{D}$.
- τ: generated trajectory.
- $\mathcal{T}_Q$: trajectory space conditioned on prompt Q.
- $\mathcal{T}$: global trajectory space.

A.1.2 Mathematical Expressions of Three RLVR Objectives

Let $\{\tau_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot \mid Q)$ denote a response group and A_i the group-relative advantage of τ_i. Following the response-level GRPO formulation in DeepSeek-R1 (Guo et al., 2025), define

$$\rho_i(\theta) \triangleq \frac{\pi_\theta(\tau_i \mid Q)}{\pi_{\text{old}}(\tau_i \mid Q)}. \tag{7}$$

We display only the policy-gradient objectives; KL regularization is analyzed separately.

For GRPO,

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\epsilon_{clip}, 1+\epsilon_{clip})\, A_i)\,. \tag{8}$$

For DAPO, define the token-level ratio

$$r_{i,t}(\theta) \triangleq \frac{\pi_\theta(\tau_{i,t} \mid Q, \tau_{i,<t})}{\pi_{\text{old}}(\tau_{i,t} \mid Q, \tau_{i,<t})}, \tag{9}$$

with objective

$$\begin{aligned} \mathcal{J}_{\text{DAPO}}(\theta) = \frac{1}{\sum_{i=1}^{G}|\tau_i|}\sum_{i=1}^{G}\sum_{t=1}^{|\tau_i|} \min(r_{i,t}(\theta)A_i, \text{clip}(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\, A_i)\,, \\ \text{s.t.} \quad 0 < \sum_{i=1}^{G} \mathbb{I}\{\tau_i \text{ is correct}\} < G. \end{aligned} \tag{10}$$

For GSPO, define the length-normalized sequence ratio

$$s_i(\theta) \triangleq \rho_i(\theta)^{1/|\tau_i|} = \exp\left(\frac{1}{|\tau_i|}\log \rho_i(\theta)\right), \tag{11}$$

and

$$\mathcal{J}_{\text{GSPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \min(s_i(\theta)A_i, \text{clip}(s_i(\theta), 1-\epsilon_{clip}, 1+\epsilon_{clip})\, A_i)\,. \tag{12}$$

A.1.3 Measures on Strategy Space

At optimization step k, the conditional probability mass of strategy $S_{\alpha,Q}$ under prompt Q is

$$\nu^{(\epsilon)}_{\alpha,Q,k} \triangleq \sum_{\tau \in S^{(\epsilon)}_{\alpha,Q}} \pi_k(\tau \mid Q). \tag{13}$$

A.2 PROOF FOR THEOREM 1

Back **Proof Sketch for Theorem 1.** RLVR acts on the model through the following chain: it directly updates policy parameters, which gradually changes trajectory probabilities and in turn affects strategy distribution. So the core idea of our proof is **simple**: To study how strategies compete and evolve, we lift the parameter update signals ALL the way up to the strategy distribution, deriving an evolution dynamics on the strategy simplex. The proof is completed through 1 preparation step and 5 main steps, organized as below:

Step 0: Dealing with the Clipping Term: We make one simple reduction to handle clipping. For GRPO and GSPO, define a binary indicator m_i, where $m_i = 0$ if trajectory τ_i is clipped and 1 otherwise. We then absorb clipping into the effective advantage,

$$A_i^{\text{eff}} = m_i A_i, \tag{14}$$

so their gradients reduce to $G^{-1}\sum_i \rho_i A_i^{\text{eff}}$ and $G^{-1}\sum_i s_i A_i^{\text{eff}}$, respectively. For DAPO, the same reduction is applied token-wise using indicators $m_{i,t}$, yielding the corresponding masked token-level gradient.

Step 1: The Dynamics of Trajectory Probabilities: Consider step k, we derive that trajectory dynamics share the same form under the three surrogates in first order:

$$\Delta\pi_k(\tau_i|Q) = \eta\pi_k(\tau_i|Q)\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}_{j,k}^{\text{eff}} r_{j,k} + \xi_{i,k}^{(\text{batch})}. \tag{15}$$

The noise has two origins, $\xi_{i,k}^{(\text{batch})} = \xi_{i,k}^{(\text{sample})} + \xi_{i,k}^{(\text{advantage})}$, satisfying $\mathbb{E}[\xi_k^{(batch)}|\mathcal{F}_k] = 0$. $r_{i,k}$ is the correction factor, which is 1 for GRPO, $|\tau_j|^{-1}\rho_{j,k}^{1/|\tau_j|-1}$ for GSPO, and $\rho_{j,k}^{-1}|\tau|^{-1}$ for DAPO in first order. (Appendix A.2.1)

Step 2: Trajectory Evolution within a Strategy: We then derive the evolution of trajectories inside a strategy, finding it in the form of a **replicator dynamics** in first order:

$$\Delta q_{i|\alpha,Q,k} = \eta \cdot q_{i|\alpha,Q,k} \cdot (f_{i,k} - f_\alpha) + \widetilde{\xi}_{i,k}^{(batch)} + \widetilde{\xi}_{i,k}^{(data)} \tag{16}$$

Where $f_{i,k} = \frac{1}{|\mathcal{D}'|}\sum_{Q'}\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}_{j,k}^{\text{eff}} r_{j,k}$ is the fitness for trajectory τ_i, and $q_{i|\alpha,Q,k} = \frac{\pi_k(\tau_i|Q)}{\nu_{\alpha,Q,k}}$ trajectory distribution within $S_{\alpha,Q}^{(\epsilon)}$. (Appendix A.2.2)

Step 3: Adiabatic Approximation: Before analyzing the strategy dynamics, we establish the scale separation. The variables involved are π, ν, q, A, K, r. We define their relative changing rate as $\omega_\pi, \omega_\nu, \omega_q, \omega_r, \omega_K, \omega_A$. Using equation 16, we prove that in condition that $\eta \ll \epsilon_{clip} < 1 \ll |\tau| \ll \eta^{-1}$, and within time window $k \ll |\tau|$:

$$\omega_q, \omega_K, \omega_r, \omega_A \ll \omega_\nu \sim \omega_\pi \tag{17}$$

Hence, the quantities q, A, K, r can be frozen in the time window when deriving the strategy dynamics, so **we can focus attention on the strategy evolution**. (Appendix A.2.3)

Step 4: Hitting time analysis: We first associate each strategy $S_{\alpha,Q}$ with its effective strategy-level fitness $a_{\alpha,Q}$. By Lemma 11, distinct strategies have almost surely different fitness. We then construct a potential function $\Phi_k(\nu_{\alpha,Q,k}, a_{\alpha,Q})$, whose drift is **almost surely** a positive definite quadric form. This makes Φ_k a submartingale, and using Azuma-Hoeffding inequality, we prove that in condition $\eta^{-1} \ll |\tau|^3$:

$$\mathbb{P}\left(\exists k < K : \max_\alpha \nu_{\alpha,Q,k} \geq 1 - \zeta\right) \geq 1 - \frac{\delta}{2} \tag{18}$$

Further analyze shows that hitting time $K \ll \mathcal{O}(|\tau|)$, meaning the hitting happens within the adiabatic zone, thereby validating the consistency of our analysis. (Appendix A.2.4)

Step 5: Preservation after the adiabatic window: Step 4 shows that the process **can reach** a concentrated state. We next fix the dominant strategy o and show, using Azuma–Hoeffding again, that it remains dominant throughout the remaining time window:

$$\mathbb{P}\left(\exists k \in \mathbb{Z}^+,\ K < k \leq T_c : \nu_{o,Q,k} < 1 - \zeta \,\middle|\, \nu_{o,Q,K} \geq 1 - \mathcal{O}(\eta)\right) \leq \delta/2 \tag{19}$$

Combining it with the result in Step 4, we can derive the convergence expression in the main text. (Appendix A.2.5) **This completes the proof.** ■

We will start from the simplest case, which only considers influence exerted by the group-relative reward in RLVR algorithms. We emphasize that divergence regularization is not neglected; rather, its role and mechanism are discussed and analyzed separately in Theorem 3. Our main focus on group-

relative reward in A.2 is to set up a relatively clear and analytic framework, in order to facilitate further discussions and effective theory development.

A.2.1 Step 1: Deriving the Dynamics of Trajectory Probabilities:

Back At optimization step k, let $\mathcal{F}_k$ denote the training history before the current rollout, write $\pi_k \triangleq \pi_{\theta_k}$, and let $\pi_{\text{old},k}$ be the behavior policy used to sample the current group. Define

$$\Delta\pi_k(\tau_i|Q) \triangleq \pi_{\theta_{k+1}}(\tau_i|Q) - \pi_{\theta_k}(\tau_i|Q), \qquad U_{i,k}(Q) \triangleq \nabla_\theta \log \pi_\theta(\tau_i|Q)\big|_{\theta=\theta_k}. \tag{20}$$

The first-order probability increment used in this step is

$$\Delta\pi_k(\tau_i|Q) = \nabla_\theta \pi_\theta(\tau_i|Q)^\top\big|_{\theta=\theta_k} \cdot \Delta\theta_k. \tag{21}$$

Using the log-derivative identity, this becomes

$$\Delta\pi_k(\tau_i|Q) = \pi_k(\tau_i|Q) \cdot U_{i,k}(Q)^\top \cdot \Delta\theta_k. \tag{22}$$

We first give the GRPO derivation. Let $B_k = \{\tau_j\}_{j=1}^G$ be the response group sampled from $\pi_{\text{old},k}(\cdot|Q')$. Away from clipping boundaries, decompose each effective group-relative advantage as

$$A^{\text{eff}}_{j,k} = \bar{A}^{\text{eff}}_{j,k} + \varepsilon^{(\text{advantage})}_{j,k}, \qquad \bar{A}^{\text{eff}}_{j,k} \triangleq \mathbb{E}\left[A^{\text{eff}}_{j,k} \mid \tau_j, Q', \mathcal{F}_k\right], \qquad \mathbb{E}\left[\varepsilon^{(\text{advantage})}_{j,k} \mid \tau_j, Q', \mathcal{F}_k\right] = 0. \tag{23}$$

Note that group-wise centering does not imply $\bar{A}^{\text{eff}}_{j,k} = 0$ once we condition on the focal trajectory τ_j. When the clipped branch is active, the surrogate is locally flat and hence $A^{\text{eff}}_{j,k} = 0$. Away from clipping, writing

$$A_{j,k} = r(\tau_j) - \frac{1}{G}\sum_{\ell=1}^{G} r(\tau_\ell),$$

Calculating its expectation, we have

$$\begin{aligned}\bar{A}^{\text{eff}}_{j,k} &= \mathbb{E}[A_{j,k} \mid \tau_j, Q', \mathcal{F}_k] \\ &= \frac{G-1}{G}\left(r(\tau_j) - \mathbb{E}_{\tau\sim\pi_{\text{old},k}(\cdot|Q')}[r(\tau) \mid Q']\right),\end{aligned} \tag{24}$$

which is generally nonzero.

Here the randomness caused by forming a finite group and estimating its group-relative advantage is included in $\varepsilon^{(\text{advantage})}_{j,k}$. With $\rho_{j,k} = \pi_k(\tau_j|Q')/\pi_{\text{old},k}(\tau_j|Q')$, its stochastic parameter update is

$$\Delta\theta_k = \eta \cdot \frac{1}{G}\sum_{j=1}^{G} \rho_{j,k} \cdot A^{\text{eff}}_{j,k} \cdot U_{j,k}(Q'). \tag{25}$$

Define its conditional population drift and the two parameter-space noise terms by

$$\bar{g}^{\text{GRPO}}_k(Q') \triangleq \sum_j \pi_{\text{old},k}(\tau_j|Q')\rho_{j,k}\bar{A}^{\text{eff}}_{j,k}U_{j,k}(Q') = \sum_j \pi_k(\tau_j|Q')\bar{A}^{\text{eff}}_{j,k}U_{j,k}(Q'), \tag{26}$$

$$\epsilon^{(\text{sample})} \triangleq \frac{1}{G}\sum_{j=1}^{G} \rho_{j,k}\bar{A}^{\text{eff}}_{j,k}U_{j,k}(Q') - \bar{g}^{\text{GRPO}}_k(Q'), \tag{27}$$

$$\epsilon^{(\text{advantage})} \triangleq \frac{1}{G}\sum_{j=1}^{G} \rho_{j,k}\varepsilon^{(\text{advantage})}_{j,k}U_{j,k}(Q'). \tag{28}$$

By construction,

$$\mathbb{E}\left[\epsilon^{(\text{sample})} \mid Q', \mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left[\epsilon^{(\text{advantage})} \mid Q', \mathcal{F}_k\right] = 0, \tag{29}$$

so that

$$\Delta\theta_k = \eta\bar{g}^{\text{GRPO}}_k(Q') + \eta\epsilon^{(\text{sample})} + \eta\epsilon^{(\text{advantage})}. \tag{30}$$

Introduce the step-indexed trajectory coupling kernel

$$K_{ij,k}(Q,Q') \triangleq U_{i,k}(Q)^\top U_{j,k}(Q'). \tag{31}$$

Substitution into the probability increment gives

$$\Delta\pi_k(\tau_i|Q) = \eta\pi_k(\tau_i|Q)\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}^{\text{eff}}_{j,k} + \xi^{(\text{batch})}_{i,k}, \tag{32}$$

where

$$\begin{aligned}
\xi^{(\text{batch})}_{i,k} &\triangleq \xi^{(\text{sample})}_{i,k} + \xi^{(\text{advantage})}_{i,k},\\
\xi^{(\text{sample})}_{i,k} &\triangleq \eta\pi_k(\tau_i|Q)U_{i,k}(Q)^\top \epsilon^{(\text{sample})},\\
\xi^{(\text{advantage})}_{i,k} &\triangleq \eta\pi_k(\tau_i|Q)U_{i,k}(Q)^\top \epsilon^{(\text{advantage})}.
\end{aligned} \tag{33}$$

Both components, and hence $\xi^{(\text{batch})}_{i,k}$, have zero conditional mean given $(Q',\mathcal{F}_k)$; independence between the two components is not required.

We then give the GSPO derivation, differentiating $s_{j,k} = \rho_{j,k}^{1/|\tau_j|}$ gives

$$\nabla_\theta s_{j,k} = \frac{s_{j,k}}{|\tau_j|}U_{j,k}(Q') = \rho_{j,k}\left(\frac{1}{|\tau_j|}\rho_{j,k}^{\frac{1}{|\tau_j|}-1}\right)U_{j,k}(Q'). \tag{34}$$

Thus GRPO and GSPO share the response-level form

$$\Delta\pi_k(\tau_i|Q) = \eta\pi_k(\tau_i|Q)\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}^{\text{eff}}_{j,k}r_{j,k} + \xi^{(\text{batch})}_{i,k}, \tag{35}$$

with

$$r^{\text{GRPO}}_{j,k} = 1, \qquad r^{\text{GSPO}}_{j,k} = \frac{1}{|\tau_j|}\rho_{j,k}^{\frac{1}{|\tau_j|}-1}. \tag{36}$$

For GSPO, the same sample/advantage decomposition above is understood with every sampled response contribution multiplied by $r^{\text{GSPO}}_{j,k}$.

We finally give the DAPO derivation, define the token score and token-level importance ratio by

$$u_{j,t,k}(Q') \triangleq \nabla_\theta \log \pi_\theta(\tau_{j,t}|Q',\tau_{j,<t})\big|_{\theta=\theta_k}, \qquad \rho_{j,t,k} \triangleq \frac{\pi_k(\tau_{j,t}|Q',\tau_{j,<t})}{\pi_{\text{old},k}(\tau_{j,t}|Q',\tau_{j,<t})}. \tag{37}$$

For fixed focal trajectory τ_i and prompts (Q,Q'), define the DAPO correction matrix $r^{DAPO}_{i,j,k}$ through a orthogonal condition:

$$U_{i,k}(Q)^\top\left(U_{j,k}(Q')r^{\text{DAPO}}_{i,j,k} - \frac{G}{\rho_{j,k}\sum_{j'=1}^{G}|\tau_{j'}|}\sum_{t=1}^{|\tau_j|}m_{j,t,k}\rho_{j,t,k}u_{j,t,k}(Q')\right) \triangleq 0 \tag{38}$$

We now prove that all $r^{DAPO}_{i,j,k}$ share a central value $r^{DAPO}_{j,k}$.

Lemma 1 (The scalar equivalence of DAPO correction matrix)**.**
For the element in DAPO correction matrix $r^{DAPO}_{i,j,k}$, there exists a scalar $r^{DAPO}_{j,k}$ so that:

$$r^{DAPO}_{i,j,k} = r^{DAPO}_{j,k} + \mathcal{O}(\eta) \tag{39}$$

Proof

Proof. Recall that $r^{\text{DAPO}}_{j,k}$ is defined, for each fixed focal trajectory τ_i, through

$$U_{i,k}(Q)^\top\left(U_{j,k}(Q')r^{\text{DAPO}}_{j,k} - \frac{G}{\rho_{j,k}Z_k}\sum_{t=1}^{|\tau_j|}m_{j,t,k}\rho_{j,t,k}u_{j,t,k}(Q')\right) = 0. \tag{40}$$

Since

$$K_{ij,k}(Q,Q') = U_{i,k}(Q)^\top U_{j,k}(Q') = \sum_{t=1}^{|\tau_j|} \kappa_{ij,t,k}(Q,Q'), \tag{41}$$

we obtain, whenever $K_{ij,k}(Q,Q') \neq 0$,

$$r_{i,j,k}^{\mathrm{DAPO}} = \frac{G}{\rho_{j,k} Z_k} \frac{\sum_{t=1}^{|\tau_j|} m_{j,t,k}\rho_{j,t,k}\kappa_{ij,t,k}(Q,Q')}{K_{ij,k}(Q,Q')}. \tag{42}$$

Adding and subtracting $\sum_t \kappa_{ij,t,k}(Q,Q') = K_{ij,k}(Q,Q')$ in the numerator gives the exact decomposition

$$r_{i,j,k}^{\mathrm{DAPO}} = \frac{G}{\rho_{j,k} Z_k} \left[1 + \frac{\sum_{t=1}^{|\tau_j|} \big(m_{j,t,k}\rho_{j,t,k} - 1\big)\kappa_{ij,t,k}(Q,Q')}{K_{ij,k}(Q,Q')}\right]. \tag{43}$$

The token-normalized DAPO update satisfies

$$\begin{aligned} \|\Delta\theta_k\| &\le \frac{\eta}{Z_k} \sum_{j=1}^{G} |A_{j,k}| \sum_{t=1}^{|\tau_j|} m_{j,t,k}\rho_{j,t,k}\|u_{j,t,k}(Q')\| \\ &= O(\eta), \end{aligned} \tag{44}$$

where bounded advantages, token-level ratios, and token-level gradients have been used. Consequently, a first-order expansion of the token-level importance ratio gives

$$\begin{aligned} \log \rho_{j,t,k} &= u_{j,t,k}(Q')^\top \Delta\theta_k + O(\|\Delta\theta_k\|^2) \\ &= O(\eta), \end{aligned} \tag{45}$$

and hence

$$\rho_{j,t,k} = 1 + O(\eta). \tag{46}$$

Because $\eta \ll \epsilon_{\mathrm{clip}}$, the local update does not cross a clipping boundary. Therefore,

$$m_{j,t,k} = 1, \qquad m_{j,t,k}\rho_{j,t,k} - 1 = O(\eta). \tag{47}$$

This implies that

$$\begin{aligned} &\left|\frac{\sum_{t=1}^{|\tau_j|} \big(m_{j,t,k}\rho_{j,t,k} - 1\big)\kappa_{ij,t,k}(Q,Q')}{K_{ij,k}(Q,Q')}\right| \\ &\le \max_{1\le t\le|\tau_j|} |m_{j,t,k}\rho_{j,t,k} - 1| \, \frac{\sum_{t=1}^{|\tau_j|} |\kappa_{ij,t,k}(Q,Q')|}{|K_{ij,k}(Q,Q')|} \\ &= O(\eta), \end{aligned} \tag{48}$$

uniformly over i. Substitution into Eq. equation 43 yields

$$r_{i,j,k}^{\mathrm{DAPO}} = \frac{G}{\rho_{j,k} Z_k}\big(1 + O(\eta)\big) \tag{49}$$

uniformly over the focal trajectory.
Define $r_{j,k}^{DAPO} = \frac{G}{\rho_{j,k} Z_k}$, we thus have:

$$\boxed{r_{i,j,k}^{DAPO} = r_{j,k}^{DAPO} + \mathcal{O}(\eta)} \tag{50}$$

thus completing the proof. □

Indeed,

$$\begin{aligned}\Delta\pi_k^{\mathrm{DAPO}}(\tau_i|Q) &= \frac{\eta\pi_k(\tau_i|Q)}{\sum_{j'=1}^{G}|\tau_{j'}|}\sum_{j=1}^{G}A_{j,k}\sum_{t=1}^{|\tau_j|}m_{j,t,k}\rho_{j,t,k}U_{i,k}(Q)^\top u_{j,t,k}(Q') \\ &= \eta\pi_k(\tau_i|Q)\frac{1}{G}\sum_{j=1}^{G}\rho_{j,k}A_{j,k}K_{ij,k}(Q,Q')r_{j,k}^{\mathrm{DAPO}}+\mathcal{O}(\eta^2).\end{aligned} \tag{51}$$

Since we study this dynamics within fixed horizon $T_c \sim \eta^{-1}$, the higher order residue terms can be neglected. Separating $A_{j,k}$ into its conditional mean and zero-mean residual, and separating the finite-group average from its conditional expectation, gives the same population form

$$\Delta\pi_k(\tau_i|Q) = \eta\pi_k(\tau_i|Q)\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}_{j,k}^{\mathrm{eff}}r_{j,k}^{\mathrm{DAPO}}+\xi_{i,k}^{(\mathrm{batch})}, \tag{52}$$

where $\xi_{i,k}^{(\mathrm{batch})} = \xi_{i,k}^{(\mathrm{sample})} + \xi_{i,k}^{(\mathrm{advantage})}$ and both components have zero conditional mean given $(Q', \mathcal{F}_k)$. For DAPO in this unified expression, $\bar{A}_{j,k}^{\mathrm{eff}}$ denotes the conditional mean of $A_{j,k}$ itself, because the token-level clipping masks have already been absorbed into $r_{j,k}^{\mathrm{DAPO}}$.

A.2.2 Step 2: Deriving the Dynamics of Trajectory Dominance within a Strategy:

Back Currently, we haven't applied any analysis on **Strategies**, so a wise choice is to first focus on the trajectory evolution within a **specific strategy**. To begin with, let's recall the partitioning method proposed in the Preliminary part. Throughout this proof, $S = S_{\alpha,Q}$ denotes a fixed equivalence class of the strategy partition induced by the initial parameters θ_0, as specified in Theorem 1; hence its membership does not vary with the optimization step k.

Based on this partitioning method, we will give a second glance at the derived equation at the end of the previous subsection:

$$\Delta\pi_k(\tau_i|Q) = \eta\cdot\pi_k(\tau_i|Q)\cdot\sum_j \pi_k(\tau_j|Q')\cdot K_{ij,k}(Q,Q')\cdot\bar{A}_{j,k}^{\mathrm{eff}}\cdot r_{j,k}+\xi_{i,k}^{(batch)} \tag{53}$$

The updated prompt Q' may change over time, fluctuating the dynamics above. In order to remove this fluctuation, we define:

$$f_{i,k} = \frac{1}{|\mathcal{D}'|}\sum_{Q'}\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}_{j,k}^{\mathrm{eff}}r_{j,k} \tag{54}$$

and define:

$$\xi_{i,k}^{(data)} = \sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}_{j,k}^{\mathrm{eff}}r_{j,k} - \frac{1}{|\mathcal{D}'|}\sum_{Q'}\sum_j \pi_k(\tau_j|Q')K_{ij,k}(Q,Q')\bar{A}_{j,k}^{\mathrm{eff}}r_{j,k} \tag{55}$$

The dynamics can be re-written as:

$$\Delta\pi_k(\tau_i|Q) = \eta\cdot\pi_k(\tau_i|Q)\cdot f_{i,k} + \eta\cdot\pi_k(\tau_i|Q)\cdot\xi_{i,k}^{(data)} + \xi_{i,k}^{(batch)} \tag{56}$$

Based on this, we introduce the relative proportion of a specific trajectory τ_i in strategy S:

$$q_{i|\alpha,Q,k} \triangleq \frac{\pi_k(\tau_i|Q)}{\nu_{\alpha,Q,k}} \tag{57}$$

Consider the upgrades in $q_{i|\alpha,Q,k}$ driven by trajectory upgrades, according to chain-rule, we can have:

$$\Delta q_{i|\alpha,Q,k} = \frac{\Delta\pi_k(\tau_i|Q)\cdot\nu_{\alpha,Q,k} - \Delta\nu_{\alpha,Q,k}\cdot\pi_k(\tau_i|Q)}{\nu_{\alpha,Q,k}^2} \tag{58}$$

in which,

$$\Delta\nu_{\alpha,Q,k} = \sum_{\tau_\ell\in S_{\alpha,Q}}\Delta\pi_k(\tau_\ell|Q). \tag{59}$$

For simplicity, define the global-fitness average within S by

$$f_\alpha \triangleq \sum_{\tau_\ell \in S} q_{\ell|\alpha,Q,k} \cdot f_{\ell,k}. \tag{60}$$

The data- and batch-noise terms induced in the normalized proportions are, respectively,

$$\widetilde{\xi}^{(data)}_{i,k} \triangleq \eta \cdot q_{i|\alpha,Q,k} \left(\xi^{(data)}_{i,k} - \sum_{\tau_\ell \in S} q_{\ell|\alpha,Q,k} \xi^{(data)}_{\ell,k} \right), \tag{61}$$

and

$$\widetilde{\xi}^{(batch)}_{i,k} \triangleq \frac{1}{\nu_{\alpha,Q,k}} \left(\xi^{(batch)}_{i,k} - q_{i|\alpha,Q,k} \sum_{\tau_\ell \in S} \xi^{(batch)}_{\ell,k} \right). \tag{62}$$

By construction, both normalized noise terms preserve the within-strategy simplex:

$$\sum_{\tau_i \in S} \widetilde{\xi}^{(data)}_{i,k} = 0, \qquad \sum_{\tau_i \in S} \widetilde{\xi}^{(batch)}_{i,k} = 0. \tag{63}$$

Under the same conditional-unbiasedness assumptions as in Step 1, including the uniform draw of Q' from $\mathcal{D}'$, both normalized noise terms also have zero conditional mean given $\mathcal{F}_k$.

$$\Delta q_{i|\alpha,Q,k} = \eta \cdot q_{i|\alpha,Q,k}(f_{i,k} - f_\alpha) + \widetilde{\xi}^{(data)}_{i,k} + \widetilde{\xi}^{(batch)}_{i,k} \tag{64}$$

This equation has the form of stochastic ***replicator-like dynamics***.

Remark 1. *The evolution advantage $f_{i,k}$ has become less direct after such complicated calculation, so it is necessary to make some explanations on its property, and what does it mean for the total training process:*

To begin with, if $f_{i,k} > f_\alpha$, the trajectory τ_i receives a positive incentive, meaning the model is actively adopting such expressions, and the higher the $f_{i,k}$, the more the model is willing to learn such expressions.

*On the contrary, if $f_{i,k} < f_\alpha$, the trajectory τ_i receives a negative incentive, meaning the model is actively abandoning such expressions, which in our bodytext is named as **expression pruning**.*

*Another thing to mention is that, the calculation of $f_{i,k}$ is not restricted to the current prompt; rather, it is computed on the whole dataset. This could lead to a seemingly weird but actually reasonable conclusion: for those **less general yet correct** expressions/strategies, they could receive a **negative incentive**, making the model less likely to study such expressions.*

Consequently, there exists a trade off: for a given expression/strategy, if it is both correct, useful and generalized (e.g. brute computation), the model would be very much willing to learn it; on the contrary, if a given expression/strategy is correct, but hard and field-specific. The model would tend to give up on this expression/strategy, and in turn pay more attention to more friendly expressions or strategies.

A.2.3 Step 3: Adiabatic Approximation

Back In this section, we revisit the expression of the evolution advantage f_i proposed in **Step 2**. Fix an optimization step k and a strategy $S = S_{\alpha,Q}$. The key observation is that the replicator dynamics only depends on the relative fitness $f_i - f_\alpha$: large common components of f_i are canceled by the baseline f_α, leaving a residual that can evolve on a slower scale. This motivates the adiabatic separation developed next. To make this structure explicit, we decompose the trajectory probability as

$$\pi_k(\tau_i|Q) = \nu_{\alpha,Q,k} q_{i|\alpha,Q,k}, \qquad \tau_i \in S_{\alpha,Q}, \tag{65}$$

where $\nu_{\alpha,Q,k}$ denotes the total probability mass of strategy $S_{\alpha,Q}$, and

$$q_{i|\alpha,Q,k} \triangleq \frac{\pi_k(\tau_i|Q)}{\nu_{\alpha,Q,k}} \tag{66}$$

denotes the corresponding in-strategy distribution.

Accordingly, instead of directly proving that f_i itself evolves slowly, we establish a separation of time scales between the strategy-level mass $\nu_{\alpha,Q,k}$ and the microscopic quantities governing f_i. Specifically, we will show that $K_{ij}(Q,Q')$, A_i, r_i, and the in-strategy distribution $q_{i|\alpha,Q,k}$ all evolve much more slowly than the strategy-level dynamics.

Denote $\omega_K \triangleq |\frac{\Delta K_{ij}(Q,Q')}{K_{ij}(Q,Q')}|$, $\omega_A \triangleq |\frac{\Delta \bar{A}_i^{eff}}{\bar{A}_i^{eff}}|$, $\omega_r \triangleq |\frac{\Delta r_i}{r_i}|$, and

$$\omega_q \triangleq \left(\sum_{\tau_i \in S_{\alpha,Q}} q_{i|\alpha,Q,k} \left| \frac{\Delta q_{i|\alpha,Q,k}}{q_{i|\alpha,Q,k}} \right|^2 \right)^{1/2}, \tag{67}$$

We further note that the trajectory-level and strategy-level dynamics evolve on the same leading-order time scale. From the trajectory dynamics,

$$\omega_\pi \triangleq \left| \frac{\Delta \pi_i}{\pi_i} \right| \sim |f_i|, \tag{68}$$

while the strategy-level dynamics derived above gives

$$\omega_\nu \triangleq \left| \frac{\Delta \nu_{\alpha,Q,k}}{\nu_{\alpha,Q,k}} \right| \sim |f_S|. \tag{69}$$

For any $\tau_i \in S$, the geometric coherence of trajectories within the same strategy implies

$$|f_i - f_S| = O(\epsilon B). \tag{70}$$

Hence we may write

$$f_i = f_S + \delta_i, \qquad |\delta_i| = O(\epsilon B). \tag{71}$$

Away from turning points where $|f_S|$ vanishes, with $|f_S| = O(B)$, we obtain

$$\frac{\omega\pi}{\omega_\nu} = \frac{|f_i|}{|f_S|} \quad = \left| 1 + \frac{\delta_i}{f_S} \right| = 1 + O(\epsilon). \tag{72}$$

Therefore,

$$\boxed{\omega_\pi \sim \omega_\nu}. \tag{73}$$

Thus, any quantity evolving much more slowly than $\pi_k(\tau_i|Q)$ also evolves much more slowly than the strategy mass $\nu_{\alpha,Q,k}$, up to the same local adiabatic approximation.

To prove that ω_K, ω_A, ω_q, and ω_r are all slow variables relative to the strategy-level evolution, we first identify that $\omega_K \ll \omega_\pi$, $\omega_A \ll \omega_\pi$, $\omega_r \ll \omega_\pi$. We will start from proving $\omega_K \ll \omega_\pi$:

Lemma 2 (The Bounds of log-prob gradients)**.**
Fix a finite training window $k \in [k_1, k_2]$. Then

$$\exists m, M > 0, \quad m|\tau| \le \|U_{i,k}\| \le M|\tau|, \quad \textit{a.s.} \tag{74}$$

throughout this window.

Proof

To establish the relative rates of change between strategy probabilities and the advantage landscape, we first analyze the magnitude of the score function (log-gradient), $\nabla_\theta \log \pi_\theta(\tau|Q)$, with respect to the trajectory length $|\tau|$. We consider one extreme scenario to bound its scaling behavior.

Upper Bound

In the presence of strong dependencies across the sequence, each conditional log-probability gradient may be bounded by a constant M such that $\|\nabla_\theta \log \pi_\theta(t_i|\text{prefix})\| \le M$. Under

these conditions, the gradients accumulate linearly along the trajectory:

$$\|U_{i,k}\| \le \sum_{i=1}^{|\tau|} \|\nabla_\theta \log \pi_\theta(t_i|\text{prefix})\| \le |\tau| \cdot M \tag{75}$$

Therefore, the upper bound for the log-gradient magnitude scales as:

$$\|\nabla_\theta \log \pi_\theta(\tau|Q)\| = \mathcal{O}(|\tau|) \tag{76}$$

Lower Bound
We first exclude the degenerate case $U_{i,k} = 0$. For any fixed trajectory, $U_{i,k}$ is a smooth score function of the model parameters. Unless this score function vanishes identically, its zero level set is a Lebesgue-null set under generic continuous model parameters. Hence,

$$U_{i,k} \neq 0 \qquad \text{a.s.} \tag{77}$$

For the finite training window $k \in [k_1, k_2]$, a finite union of such null events remains null. Moreover, since the trajectory space is countable, we may choose one trajectory from each nonempty strategy, and the union of the corresponding null events is still null. Therefore, almost surely, every strategy contains a trajectory satisfying $U_{i,k} \neq 0$ throughout the whole window.
Because the training window is finite and $|\tau| > 0$, the finite sequence $\{\|U_{i,k}\|/|\tau|\}_{k=k_1}^{k_2}$ consists entirely of strictly positive terms almost surely. It therefore attains an almost-surely positive minimum. Define

$$m \triangleq \min_{k\in[k_1,k_2]} \frac{\|U_{i,k}\|}{|\tau|} > 0 \qquad \text{a.s.} \tag{78}$$

By construction, this gives the uniform lower bound

$$\|U_{i,k}\| \ge m|\tau|, \qquad k \in [k_1, k_2]. \tag{79}$$

We next propagate this lower bound to the remaining trajectories in the same strategy. By the definition of dynamic equivalence,

$$|K_{ii} - K_{ji}| \le \epsilon \sup_{a,b} |K_{ab}|, \qquad |K_{ij} - K_{jj}| \le \epsilon \sup_{a,b} |K_{ab}|. \tag{80}$$

Therefore,

$$\begin{aligned} \|U_{i,k} - U_{j,k}\|^2 &= K_{ii} + K_{jj} - 2K_{ij} \\ &\le 2\epsilon \sup_{a,b} |K_{ab}| \le 2\epsilon M^2 |\tau|^2, \end{aligned} \tag{81}$$

and hence

$$\|U_{j,k}\| \ge \|U_{i,k}\| - \|U_{i,k} - U_{j,k}\| \ge \left(m - \sqrt{2\epsilon} M\right) |\tau|. \tag{82}$$

For sufficiently small fixed ϵ, the coefficient remains strictly positive. Together with the upper bound $\|U_{i,k}\| \le M|\tau|$, we obtain

$$\boxed{\|U_{i,k}\| = \Theta(|\tau|) \qquad \text{a.s.}} \tag{83}$$

throughout the finite training window.
Thus, the score-function magnitude has linear scale in trajectory length almost surely, which yields the scaling required for the time-scale separation in Proposition 3.

This is a **local-in-time** scaling statement, rather than a uniform claim over the entire training process. The constants m, M may depend on the chosen training window, but are independent of $|\tau|$. The upper bound follows from bounded token-level score norms, while the lower bound only requires that their accumulated score retains a non-vanishing projection along some direction, excluding asymptotically exact cancellation. Thus, $\|U_{i,k}\| = \Theta(|\tau|)$ describes scaling with trajectory length within a fixed window, not monotonic growth over training.

Based on this complexity analysis, we propose the following lemma regarding the relative rates of change, under the condition of $|\tau| \gg 1$:

Lemma 3 (Adiabatic Property of Kernel)**.**
Let ω_π denote the relative rate of change in trajectory probability, and ω_K denote the relative rate of change in the advantage kernel matrix elements K_{ij}. Under standard training regimes, we have:

$$\frac{\omega_K}{\omega_\pi} \ll 1 \tag{84}$$

Proof

Proof. For an autoregressive trajectory τ_i of length $|\tau|$,

$$U_i = \sum_{t=1}^{|\tau|} \nabla_\theta \log \pi_\theta(\tau_{i,t} \mid Q, \tau_{i,<t}). \tag{85}$$

Assuming uniformly bounded token-level gradients and Hessians,

$$\|U_i\| = O(|\tau|), \qquad \|\nabla_\theta U_i\|_{\mathrm{op}} = O(|\tau|). \tag{86}$$

Let $\Delta\theta$ denote one optimization update. For trajectories actively affected by the update, we assume that the score does not become asymptotically orthogonal to the update direction, so that

$$|U_i^\top \Delta\theta| = \Theta(|\tau|\|\Delta\theta\|). \tag{87}$$

Using a first-order Taylor expansion,

$$\Delta \log \pi_\theta(\tau_i|Q) = U_i^\top \Delta\theta + O(|\tau|\|\Delta\theta\|^2). \tag{88}$$

Under the usual small-update condition $|\tau|\|\Delta\theta\| \ll 1$, this gives

$$\frac{\Delta\pi_\theta(\tau_i|Q)}{\pi_\theta(\tau_i|Q)} = \exp(\Delta \log \pi_\theta(\tau_i|Q)) - 1 = \Delta \log \pi_\theta(\tau_i|Q) + O\Big((\Delta \log \pi_\theta(\tau_i|Q))^2\Big). \tag{89}$$

Therefore,

$$\boxed{\omega_{\pi,k} = \Theta(|\tau|\|\Delta\theta\|).} \tag{90}$$

We next consider the kernel

$$K_{ij} = U_i^\top U_j. \tag{91}$$

Since

$$\Delta U_i = O(|\tau|\|\Delta\theta\|), \tag{92}$$

its one-step variation satisfies

$$\begin{aligned} \Delta K_{ij} &= (\Delta U_i)^\top U_j + U_i^\top (\Delta U_j) + (\Delta U_i)^\top (\Delta U_j) \\ &= O(|\tau|^2 \|\Delta\theta\|). \end{aligned} \tag{93}$$

For active kernel entries whose magnitude is of the natural scale $|K_{ij}| = \Omega(|\tau|^2)$, we therefore obtain

$$\boxed{\omega_{K,k} = \left|\frac{\Delta K_{ij}}{K_{ij}}\right| = O(\|\Delta\theta\|).} \tag{94}$$

Combining the two estimates yields

$$\boxed{\frac{\omega_{K,k}}{\omega_{\pi,k}} = O\left(\frac{1}{|\tau|}\right).} \tag{95}$$

Hence, for long reasoning trajectories, the kernel evolves on a much slower time scale than the trajectory probabilities and can be treated as approximately frozen during strategy redistribution. □

Lemma 3 holds under the requirement $\eta \ll 1 \ll |\tau|$. We now proceeds to prove that under the requirement of $\eta|\tau| \ll 1$, $\omega_r \ll \omega_\pi$:

Lemma 4 (Adiabatic Property of Correction Factor)**.**
The relative changing rate of the correction factor satisfies:

$$\frac{\omega_r}{\omega_\pi} \ll 1 \tag{96}$$

Proof

From previous analysis, we already obtained the evolution dynamics of $\pi_\theta(\tau_i|Q)$:

$$\Delta\pi_\theta(\tau_i|Q) = \eta \cdot \pi_\theta(\tau_i|Q) \cdot f_i + \xi_i^{(batch)} + \xi_i^{(data)} \tag{97}$$

This means that we can directly obtain the relative changing rate ω_π:

$$\omega_\pi \sim \eta \cdot f_i \sim \mathcal{O}(\eta) \tag{98}$$

Because r_i is a continuous function of ρ_i, and $\omega_\rho = \Delta \log \rho_i$, we consider $\Delta \log \rho_i$. To begin with, we first consider the dynamical property of ρ_i:

$$\log \rho_i = \Delta \log \pi_\theta(\tau|Q) = \eta \nabla_\theta \log \pi_\theta(\tau|Q)^\top \cdot \Delta\theta \tag{99}$$

Therefore, doing the derivatives on the equation again, we have:

$$\Delta \log \rho_i = \Delta\theta^\top \nabla_\theta^2 \log \pi_\theta(\tau_i|Q) \Delta\theta \tag{100}$$

According to previous Lemma 2, $\nabla_\theta^2 \log \pi_\theta(\tau_i|Q) \sim \mathcal{O}(|\tau|)$, $\Delta\theta \sim \mathcal{O}(\eta)$, therefore:

$$\Delta \log \rho_i \sim \mathcal{O}(\eta^2|\tau|) \tag{101}$$

Therefore, divide ω_r by ω_π, we have:

$$\frac{\omega_r}{\omega\pi} \sim \mathcal{O}(\eta|\tau|) \tag{102}$$

When the experiment configuration satisfies: $\eta|\tau| \ll 1$, we have:

$$\frac{\omega_\rho}{\omega_\pi} \ll 1 \tag{103}$$

Since $r = f(\rho_i)$, we have: $\omega_r = C \cdot \omega_\rho$, in which C is a near constant value compared with $|\tau|$, we have:

$$\frac{\omega_r}{\omega_\pi} \ll 1 \tag{104}$$

Why the correction factor varies at second order. Because the old policy is refreshed at every step, $\pi_{\text{old},k} = \pi_{k-1}$, we have

$$\rho_{i,k} = \frac{\pi_k(\tau_i \mid Q)}{\pi_{k-1}(\tau_i \mid Q)}, \qquad \log \rho_{i,k} = \log \pi_k(\tau_i \mid Q) - \log \pi_{k-1}(\tau_i \mid Q). \tag{105}$$

Hence $\log \rho_{i,k}$ is already the first discrete difference of the log-policy, and its temporal variation is the second difference:

$$\Delta \log \rho_{i,k} = \log \pi_{k+1}(\tau_i \mid Q) - 2 \log \pi_k(\tau_i \mid Q) + \log \pi_{k-1}(\tau_i \mid Q). \tag{106}$$

Since each parameter update is $\mathcal{O}(\eta)$ and the score field varies smoothly across consecutive steps, the common first-order $\mathcal{O}(\eta)$ increment cancels, leaving

$$\Delta \log \rho_{i,k} = \mathcal{O}(\eta^2|\tau|). \tag{107}$$

Thus the correction factor evolves on a slower scale than the policy itself.

The previous lemma works for GRPO ($\omega_r = 0$) and GSPO ($\omega_r \ll \omega_\pi$). However, directly analyzing ω_r in DAPO optimization is too hard, as it contains too many non-convex surrogates. We instead seek another approach to validate the time-scale separation in DAPO:

Lemma 5 (Adiabatic Property of the DAPO Effective Correction)**.**
For DAPO, define the mixed token-level kernel

$$\kappa_{ij,t,k}(Q,Q') \triangleq U_{i,k}(Q)^\top u_{j,t,k}(Q'), \tag{108}$$

and the effective correction term

$$\mathcal{R}^{\mathrm{DAPO}}_{ij,k} \triangleq \frac{G}{Z_k}\sum_{t=1}^{|\tau_j|} m_{j,t,k}\rho_{j,t,k}\kappa_{ij,t,k}, \qquad Z_k \triangleq \sum_{\ell=1}^{G}|\tau_\ell|. \tag{109}$$

Within an optimization block with fixed clipping active set, assuming bounded token-level gradients and Hessians and excluding asymptotically degenerate cancellation among active token contributions, the relative changing rate of $\mathcal{R}^{\mathrm{DAPO}}_{ij,k}$ satisfies

$$\boxed{\frac{\omega_{\mathcal{R}}}{\omega_\pi} = O\left(\frac{1}{|\tau|}\right) \ll 1}. \tag{110}$$

Moreover, the DAPO population term admits the exact identity

$$\pi_k(\tau_j|Q')K_{ij,k}(Q,Q')r^{\mathrm{DAPO}}_{j,k} = \pi_{\mathrm{old},k}(\tau_j|Q')\mathcal{R}^{\mathrm{DAPO}}_{ij,k}. \tag{111}$$

Hence, although the raw factor $r^{\mathrm{DAPO}}_{j,k}$ need not itself be adiabatic, the effective DAPO correction entering the population dynamics evolves on a much slower time scale than the trajectory probability.

Proof

Proof. From the definition of the DAPO correction factor,

$$r^{\mathrm{DAPO}}_{j,k} = \frac{G}{\rho_{j,k}Z_k}\frac{\sum_{t=1}^{|\tau_j|} m_{j,t,k}\rho_{j,t,k}U_{i,k}(Q)^\top u_{j,t,k}(Q')}{K_{ij,k}(Q,Q')}. \tag{112}$$

Multiplying both sides by $\pi_k(\tau_j|Q')K_{ij,k}(Q,Q')$ gives

$$\begin{aligned}&\pi_k(\tau_j|Q')K_{ij,k}(Q,Q')r^{\mathrm{DAPO}}_{j,k}\\ &= \frac{G}{Z_k}\frac{\pi_k(\tau_j|Q')}{\rho_{j,k}}\sum_{t=1}^{|\tau_j|} m_{j,t,k}\rho_{j,t,k}\kappa_{ij,t,k}(Q,Q').\end{aligned} \tag{113}$$

Since the sequence-level importance ratio satisfies

$$\rho_{j,k} = \frac{\pi_k(\tau_j|Q')}{\pi_{\mathrm{old},k}(\tau_j|Q')}, \tag{114}$$

we obtain the exact cancellation

$$\boxed{\pi_k(\tau_j|Q')K_{ij,k}(Q,Q')r^{\mathrm{DAPO}}_{j,k} = \pi_{\mathrm{old},k}(\tau_j|Q')\mathcal{R}^{\mathrm{DAPO}}_{ij,k}}. \tag{115}$$

Therefore, it suffices to study the time scale of $\mathcal{R}^{\mathrm{DAPO}}_{ij,k}$.
We first consider the token-level importance ratio

$$\rho_{j,t,k} = \frac{\pi_{\theta_k}(\tau_{j,t} \mid Q', \tau_{j,<t})}{\pi_{\mathrm{old},k}(\tau_{j,t} \mid Q', \tau_{j,<t})}. \tag{116}$$

Within one optimization block, the denominator is fixed. Hence, using a first-order Taylor expansion,

$$\begin{aligned}\Delta\log\rho_{j,t,k} &= u_{j,t,k}(Q')^\top\Delta\theta + O(\|\Delta\theta\|^2)\\ &= O(\|\Delta\theta\|),\end{aligned} \tag{117}$$

where bounded token-level gradients have been used. Therefore,

$$\frac{\Delta\rho_{j,t,k}}{\rho_{j,t,k}} = O(\|\Delta\theta\|). \tag{118}$$

We next consider the mixed token-level kernel

$$\kappa_{ij,t,k} = U_{i,k}^{\top} u_{j,t,k}. \tag{119}$$

By the bounded-gradient and bounded-Hessian assumptions used in Lemma 2,

$$\|U_{i,k}\| = O(|\tau|), \qquad \|\Delta U_{i,k}\| = O(|\tau|\|\Delta\theta\|), \tag{120}$$

while for a single token,

$$\|u_{j,t,k}\| = O(1), \qquad \|\Delta u_{j,t,k}\| = O(\|\Delta\theta\|). \tag{121}$$

Consequently,

$$\begin{aligned} \Delta\kappa_{ij,t,k} &= (\Delta U_{i,k})^{\top} u_{j,t,k} + U_{i,k}^{\top}(\Delta u_{j,t,k}) + O(\|\Delta\theta\|^2) \\ &= O(|\tau|\|\Delta\theta\|). \end{aligned} \tag{122}$$

For active mixed-kernel entries of their natural scale $|\kappa_{ij,t,k}| = \Omega(|\tau|)$, this yields

$$\frac{\Delta\kappa_{ij,t,k}}{\kappa_{ij,t,k}} = O(\|\Delta\theta\|). \tag{123}$$

Now define each active token contribution by

$$x_{j,t,k} \triangleq m_{j,t,k}\rho_{j,t,k}\kappa_{ij,t,k}. \tag{124}$$

On any interval where the clipping active set is fixed, $\Delta m_{j,t,k} = 0$. Hence,

$$\begin{aligned} \Delta x_{j,t,k} &= m_{j,t,k}\left(\kappa_{ij,t,k}\Delta\rho_{j,t,k} + \rho_{j,t,k}\Delta\kappa_{ij,t,k}\right) + O(\|\Delta\theta\|^2) \\ &= x_{j,t,k} O(\|\Delta\theta\|). \end{aligned} \tag{125}$$

Since

$$\mathcal{R}_{ij,k}^{\mathrm{DAPO}} = \frac{G}{Z_k}\sum_t x_{j,t,k}, \tag{126}$$

and G/Z_k is fixed inside the optimization block, under the non-degenerate-cancellation condition

$$\left|\sum_t x_{j,t,k}\right| = \Omega\left(\sum_t |x_{j,t,k}|\right), \tag{127}$$

we obtain

$$\boxed{\omega_{\mathcal{R}} = \left|\frac{\Delta\mathcal{R}_{ij,k}^{\mathrm{DAPO}}}{\mathcal{R}_{ij,k}^{\mathrm{DAPO}}}\right| = O(\|\Delta\theta\|)}. \tag{128}$$

From Lemma 2, the trajectory probability evolves at the rate

$$\omega_{\pi} = \Theta(|\tau|\|\Delta\theta\|). \tag{129}$$

Combining the two estimates therefore gives

$$\boxed{\frac{\omega_{\mathcal{R}}}{\omega_{\pi}} = O\left(\frac{1}{|\tau|}\right) \ll 1}. \tag{130}$$

Thus, after the exact cancellation of the sequence-level importance ratio, the remaining DAPO correction consists only of token-level importance ratios and mixed token-level kernels, both of which evolve adiabatically relative to the trajectory probability. □

Although DAPO admits a scalarized correction at first order, the temporal variation of its higher-order residual may not be benignly controlled. We therefore analyze the exact correction entering the population dynamics, shown in the lemma.

We then prove that: under requirements $\eta \ll \epsilon < 1, \eta \cdot |\tau| \ll 1$, the following lemma stands:

Lemma 6 (Adiabatic Property of Clipped Reward Signal)**.**
The relative changing rate of the clipping boundary satisfies:

$$\frac{\omega_A}{\omega_\pi} \ll 1 \tag{131}$$

Proof

To prove this, we first consider the clipping effect, for trajectories that can produce normal gradients, they satisfy:

$$1 - \epsilon_{low} \leq \rho_i \leq 1 + \epsilon_{high} \tag{132}$$

Or equivalently, to first order approximation:

$$-\epsilon_{low} \leq \log \rho_i \leq \epsilon_{high} \tag{133}$$

Given that ϵ is still an observable amount compared with 1, therefore:

$$\eta|\tau| \ll \min\{\epsilon_{low}, \epsilon_{high}\} \tag{134}$$

Therefore, we claim that:

$$\Delta \log \rho_i \ll \min\{\epsilon_{low}, \epsilon_{high}\} \tag{135}$$

This means that, during the update step, only a $\mathcal{O}(\eta^2|\tau|)$ proportion of trajectories will undergo the clipping transition; therefore, the relative change rate of the reward signal A_i follows:

$$\omega_A = \left|\frac{\Delta A_i}{A_i}\right| < \mathcal{O}(\eta^2|\tau|) \cdot \left|\frac{A_{\text{clipped}}}{A_{\min}}\right| = \mathcal{O}(\eta^2|\tau|). \tag{136}$$

Therefore, we can still conclude:

$$\frac{\omega_A}{\omega_\pi} \sim \mathcal{O}(\eta|\tau|) \ll 1 \tag{137}$$

After proving the four lemmas above, we can already control the last three terms $\omega_K, \omega_A, \omega_r$. We now turn to the last term q. To analyze it, we decompose each trajectory probability into its strategy-level mass and its in-strategy distribution, and show that the latter evolves on a much slower time scale due to the geometric coherence of trajectories within the same **Strategy**. This establishes the adiabaticity of the in-strategy distribution and allows us to locally freeze it in the subsequent derivation.

Lemma 7 (Adiabatic Property of Strategies)**.** *For the fixed strategy $S = S_{\alpha,Q}$, let its relative changing rate be $\omega_\nu \triangleq \frac{\Delta \nu_{\alpha,Q,k}}{\nu_{\alpha,Q,k}}$. Then:*

$$\frac{\omega_\nu}{\omega_q} \gg 1 \tag{138}$$

To begin with, we first prove a useful lemma regarding f_i:

Lemma 8 (f_i is unbiased)**.**
For the f_i definition given, we have:

$$\sum_i \pi_\theta(\tau_i|Q) \cdot f_i = 0 \tag{139}$$

Proof

By definition, we can expand the summation above:

$$\sum_i \pi_\theta(\tau_i|Q) \cdot f_i = \sum_{i,j} \pi_\theta(\tau_i|Q) \cdot K_{ij}(Q, Q') \cdot \pi_\theta(\tau_j|Q') \cdot A_j \cdot r_j \tag{140}$$

Denote:

$$U_i(Q) \triangleq \nabla_\theta \log(\tau_i|Q) \tag{141}$$

then $K_{ij}(Q, Q') = U_i(Q)^\top \cdot U_j(Q)$, also it's obvious that:

$$\sum_i \pi_\theta(\tau_i|Q) \cdot U_i(Q) = 0 \tag{142}$$

Therefore:

$$\sum_i \pi_\theta(\tau_i|Q) \cdot f_i = \sum_j \big(\sum_i \pi_\theta(\tau_i|Q) \cdot U_i(Q)^\top \big) \cdot U_j(Q') \cdot \pi_\theta(\tau_j|Q') \cdot A_j \cdot r_j = 0 \tag{143}$$

Therefore we conclude:

$$\sum_i \pi_\theta(\tau_i|Q) \cdot f_i = 0 \tag{144}$$

Using this beautiful lemma, we can rewrite the dynamics into an anti-symmetric form:

$$\Delta\pi_k(\tau_i|Q) = \eta \cdot \pi_k(\tau_i|Q) \cdot f_{i,k} + \xi_{i,k} = \eta \cdot \pi_k(\tau_i|Q) \cdot \sum_j \pi_k(\tau_j|Q) \cdot f_{i,k} + \xi_{i,k} \tag{145}$$

Using $\sum_j \pi_k(\tau_j|Q) \cdot f_{j,k} = 0$, we have:

$$\Delta\pi_k(\tau_i|Q) = \eta \cdot \sum_{i,j} \pi_k(\tau_i|Q) \cdot \pi_k(\tau_j|Q) \cdot (f_{i,k} - f_{j,k}) + \xi_{i,k} \tag{146}$$

Let $J_{ij,k} \triangleq \pi_k(\tau_i|Q) \cdot \pi_k(\tau_j|Q) \cdot (f_{i,k} - f_{j,k})$, we have:

$$\boxed{J_{ij,k} = -J_{ji,k}} \tag{147}$$

For the fixed strategy $S = S_{\alpha,Q}$, expand $\Delta\nu_{\alpha,Q,k}$:

$$\Delta\nu_{\alpha,Q,k} = \sum_{\tau_i \in S_{\alpha,Q}} \Delta\pi_k(\tau_i|Q) \tag{148}$$

Substitute $\Delta\pi_k(\tau_i|Q)$ with the dynamics derived before, we have:

$$\Delta\nu_{\alpha,Q,k} = \eta \cdot \sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j} J_{ij,k} + \xi_{\alpha,Q,k} \tag{149}$$

Then split the right hand side into an intra-strategy term and inter-strategy term:

$$\Delta\nu_{\alpha,Q,k} = \eta \cdot \sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j \in S_{\alpha,Q}} J_{ij,k} + \eta \cdot \sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j \notin S_{\alpha,Q}} J_{ij,k} + \xi_{\alpha,Q,k} \tag{150}$$

Because of the anti-symmetric property of $J_{ij,k}$, the first term will vanish:

$$\sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j \in S_{\alpha,Q}} J_{ij,k} = 0 \tag{151}$$

Therefore, we have:

$$\Delta\nu_{\alpha,Q,k} = \eta \cdot \sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j \notin S_{\alpha,Q}} J_{ij,k} + \xi_{\alpha,Q,k} \tag{152}$$

We now show that the same anti-symmetric flow structure is exactly preserved after coarse-graining trajectories into strategies. For any two strategies $S = S_{\alpha,Q}$ and $T = S_{\beta,Q}$, define the strategy-level flow as

$$J_{\alpha\beta,Q,k} \triangleq \sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j \in S_{\beta,Q}} J_{ij,k}. \tag{153}$$

Expanding $J_{ij,k}$ gives

$$\begin{aligned} J_{\alpha\beta,Q,k} &= \sum_{\tau_i \in S_{\alpha,Q}} \sum_{\tau_j \in S_{\beta,Q}} \pi_k(\tau_i|Q)\pi_k(\tau_j|Q)(f_{i,k} - f_{j,k}) \\ &= \left(\sum_{\tau_i \in S_{\alpha,Q}} \pi_k(\tau_i|Q) f_{i,k} \right) \left(\sum_{\tau_j \in S_{\beta,Q}} \pi_k(\tau_j|Q) \right) \\ &\quad - \left(\sum_{\tau_i \in S_{\alpha,Q}} \pi_k(\tau_i|Q) \right) \left(\sum_{\tau_j \in S_{\beta,Q}} \pi_k(\tau_j|Q) f_{j,k} \right). \end{aligned} \tag{154}$$

Recall that

$$\nu_{\alpha,Q,k} = \sum_{\tau_i \in S_{\alpha,Q}} \pi_k(\tau_i|Q), \tag{155}$$

and define the strategy-level fitness as

$$f_{\alpha,Q,k} \triangleq \frac{1}{\nu_{\alpha,Q,k}} \sum_{\tau_i \in S_{\alpha,Q}} \pi_k(\tau_i|Q) f_{i,k}. \tag{156}$$

Therefore,

$$\sum_{\tau_i \in S_{\alpha,Q}} \pi_k(\tau_i|Q) f_{i,k} = \nu_{\alpha,Q,k} f_{\alpha,Q,k}, \tag{157}$$

and similarly for $S_{\beta,Q}$. Hence,

$$\begin{aligned} J_{\alpha\beta,Q,k} &= \nu_{\alpha,Q,k} f_{\alpha,Q,k} \nu_{\beta,Q,k} - \nu_{\alpha,Q,k} \nu_{\beta,Q,k} f_{\beta,Q,k} \\ &= \boxed{\nu_{\alpha,Q,k} \nu_{\beta,Q,k} (f_{\alpha,Q,k} - f_{\beta,Q,k})}. \end{aligned} \tag{158}$$

Thus, the trajectory-level anti-symmetric flow induces an exactly analogous strategy-level flow. In particular,

$$\boxed{J_{\alpha\beta,Q,k} = -J_{\beta\alpha,Q,k}}. \tag{159}$$

Therefore, the strategy dynamics can be written as

$$\Delta\nu_{\alpha,Q,k} = \eta \cdot \sum_{\beta \neq \alpha} \nu_{\alpha,Q,k} \nu_{\beta,Q,k} (f_{\alpha,Q,k} - f_{\beta,Q,k}) + \xi_{\alpha,Q,k}. \tag{160}$$

Since the strategy partition covers the trajectory space, the $\beta = \alpha$ term is identically zero, and we may equivalently write

$$\Delta\nu_{\alpha,Q,k} = \eta \cdot \nu_{\alpha,Q,k} \sum_{\beta} \nu_{\beta,Q,k} (f_{\alpha,Q,k} - f_{\beta,Q,k}) + \xi_{\alpha,Q,k}. \tag{161}$$

Expanding the summation,

$$\Delta\nu_{\alpha,Q,k} = \eta \cdot \nu_{\alpha,Q,k} \left[f_{\alpha,Q,k} \sum_{\beta} \nu_{\beta,Q,k} - \sum_{\beta} \nu_{\beta,Q,k} f_{\beta,Q,k} \right] + \xi_{\alpha,Q,k}. \tag{162}$$

Because

$$\sum_{\beta} \nu_{\beta,Q,k} = 1, \tag{163}$$

and

$$\begin{aligned} \sum_{\beta} \nu_{\beta,Q,k} f_{\beta,Q,k} &= \sum_{\beta} \sum_{\tau_i \in S_{\beta,Q}} \pi_k(\tau_i|Q) f_{i,k} \\ &= \sum_{i} \pi_k(\tau_i|Q) f_{i,k} = 0, \end{aligned} \tag{164}$$

we finally obtain

$$\boxed{\Delta\nu_{\alpha,Q,k} = \eta \cdot \nu_{\alpha,Q,k} f_{\alpha,Q,k} + \xi_{\alpha,Q,k}}. \tag{165}$$

Then

$$\boxed{\Delta\nu_{\alpha,Q,k} \sim \nu_{\alpha,Q,k} f_{\alpha,Q,k}}. \tag{166}$$

Therefore, the relative time scale of the strategy mass is

$$\boxed{\omega_\nu \triangleq \frac{|\Delta\nu_{\alpha,Q,k}|}{\nu_{\alpha,Q,k}} \sim |f_{\alpha,Q,k}|}. \tag{167}$$

We now derive the dynamics within the strategy. Recall that

$$q_{i|\alpha,Q,k} = \frac{\pi_k(\tau_i|Q)}{\nu_{\alpha,Q,k}}, \tag{168}$$

we have

$$\Delta q_{i|\alpha,Q,k} \sim q_{i|\alpha,Q,k}(f_i - f_{\alpha,Q,k}). \tag{169}$$

Thus,

$$\boxed{\frac{\Delta q_{i|\alpha,Q,k}}{q_{i|\alpha,Q,k}} \sim f_i - f_{\alpha,Q,k}}. \tag{170}$$

Unlike $\nu_{\alpha,Q,k}$, whose evolution is driven by the absolute strategy-level fitness $f_{\alpha,Q,k}$, the intra-strategy redistribution is driven only by fitness fluctuations within the strategy.

To quantify this difference, define the RMS relative intra-strategy speed

$$\omega_{q,S}^2 \triangleq \sum_{i\in S} q_{i|\alpha,Q,k} \left(\frac{\Delta q_{i|\alpha,Q,k}}{q_{i|\alpha,Q,k}}\right)^2. \tag{171}$$

Using the dynamics above,

$$\omega_{q,S}^2 = \sum_{i\in S} q_{i|\alpha,Q,k}(f_i - f_{\alpha,Q,k})^2. \tag{172}$$

Define the within-strategy fitness fluctuation

$$\sigma_S^2 \triangleq \sum_{i\in S} q_{i|\alpha,Q,k}(f_i - f_{\alpha,Q,k})^2. \tag{173}$$

Then

$$\boxed{\omega_{q,S} = \sigma_S}. \tag{174}$$

Combining the two exact identities gives

$$\boxed{\frac{\omega_{q,S}}{\omega_\nu} = \frac{\sigma_S}{|f_{\alpha,Q,k}|}}. \tag{175}$$

It remains to bound σ_S. Recall that the trajectory-level fitness takes the form

$$f_i(Q) = \frac{1}{|\mathcal{D}'|} \sum_{Q'} \sum_j \pi_j(Q') K_{ij}(Q, Q') a_j(Q'), \tag{176}$$

where

$$a_j(Q') \triangleq \bar{A}_j^{\text{eff}}(Q') r_j(Q'). \tag{177}$$

For two trajectories $i, \ell \in S_{\alpha,Q}$,

$$f_i - f_\ell = \frac{1}{|\mathcal{D}'|} \sum_{Q',j} \pi_j(Q') \left[K_{ij}(Q, Q') - K_{\ell j}(Q, Q')\right] a_j(Q'). \tag{178}$$

Before proceeding, we prove a useful lemma:

Lemma 9 (Expanded $\epsilon-$similarity of strategies)**.**
For $\tau_i, \tau_l \in S_{\alpha,Q}$, for any other trajectory τ_k generated under Q', we still have:

$$\sup_{Q',j} \|K_{ij}(Q, Q') - K_{\ell j}(Q, Q')\| \le \mathcal{O}(\epsilon) \tag{179}$$

Proof

Proof. Note that the trajectories generated given Q is nearly ∞, so obviously, for trajectory τ_s, the fisher score $U_s(Q)$ should be desne within the vector space. Therefore, we can find one trajectory s satisfying

$$\|U_s(Q) - U_j(Q')\| \le \epsilon \tag{180}$$

Therefore:

$$\|K_{ij}(Q, Q') - K_{\ell j}(Q, Q')\| = \|(U_i(Q) - U_\ell(Q))U_j(Q')\| \tag{181}$$

Using Cauchy inequality:

$$\begin{aligned}\|K_{ij}(Q,Q') - K_{\ell j}(Q,Q')\| \le \|K_{is}(Q,Q) - K_{\ell s}(Q,Q)\| \\ + \|(U_i(Q) - U_\ell(Q))(U_j(Q') - U_s(Q))\|\end{aligned} \tag{182}$$

Using Definition 1, we have:

$$\sup_{Q',j} \|K_{ij}(Q,Q') - K_{\ell j}(Q,Q')\| \le \mathcal{O}(\epsilon) \tag{183}$$

This complets the lemma. □

Then

$$|f_i - f_\ell| \le \frac{\epsilon}{|\mathcal{D}'|} \sum_{Q',j} \pi_k(\tau_j|Q')|a_j(Q')|. \tag{184}$$

Define

$$B \triangleq \frac{1}{|\mathcal{D}'|} \sum_{Q',j} \pi_k(\tau_j|Q')|a_j(Q')|. \tag{185}$$

Therefore,

$$\boxed{|f_i - f_\ell| \le \epsilon B}. \tag{186}$$

Hence the fitness variation within a strategy is uniformly of order $O(\epsilon B)$, which implies

$$\boxed{\sigma_S = O(\epsilon B)}. \tag{187}$$

Away from turning points where the strategy-level mean fitness vanishes, assume

$$|f_{\alpha,Q,k}| = O(B). \tag{188}$$

We then obtain

$$\begin{aligned}\frac{\omega_{q,S}}{\omega_\nu} &= \frac{\sigma_S}{|f_{\alpha,Q,k}|} \\ &= O(\epsilon) \ll 1.\end{aligned} \tag{189}$$

Therefore,

$$\boxed{\omega_q \ll \omega_\nu}. \tag{190}$$

Thus, strategy selection through $\nu_{\alpha,Q,k}$ occurs on the fast time scale governed by the strategy-level mean fitness, whereas redistribution among trajectories within the same strategy is driven only by the much smaller intra-strategy fitness fluctuations. Consequently, $q_{i|\alpha,Q,k}$ can be treated as adiabatically frozen over the local time scale on which $\nu_{\alpha,Q,k}$ evolves.

Having established the preceding adiabatic estimates, the quantities q, A, K, r remain effectively frozen over time zone $k \ll |\tau|$. **Hence, we can restrict attention to the strategy-level evolution.**

A.2.4 Step 4: Hitting time analysis

Back We prove that the strategy dynamics induced by GRPO/GSPO/DAPO admits a monotonic potential function, which inevitably drives the strategy distribution toward concentration under mild generic conditions.

Fix an optimization step k. For each query Q, write its active strategy partition as $\{S_{\alpha,Q}\}_{\alpha=1}^{m_Q}$ and define

$$\nu_{\alpha,Q,k} \triangleq \sum_{\tau_i \in S_{\alpha,Q}} \pi_k(\tau_i \mid Q), \qquad q_{i|\alpha,Q,k} \triangleq \frac{\pi_k(\tau_i \mid Q)}{\nu_{\alpha,Q,k}}, \quad \tau_i \in S_{\alpha,Q}. \tag{191}$$

Thus $\sum_{i \in S_{\alpha,Q}} q_{i|\alpha,Q,k} = 1$. The fitness of $S_{\alpha,Q}$ is the conditional trajectory average

$$f_{\alpha,Q,k} = \sum_{\tau_i \in S_{\alpha,Q}} q_{i|\alpha,Q,k} f_{i,k}(Q). \tag{192}$$

According to the trajectory-level evolution derived in Step 2,

$$f_{i,k}(Q) = \frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{j} \pi_k(\tau_j \mid Q') K_{ij,k}(Q, Q') a_{j,k}(Q'), \qquad a_{j,k}(Q') \triangleq \bar{A}^{\text{eff}}_{j,k} r_{j,k}. \tag{193}$$

Therefore,

$$\boxed{f_{\alpha,Q,k} = \frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{j} \pi_k(\tau_j \mid Q') a_{j,k}(Q') \sum_{i \in S_{\alpha,Q}} q_{i|\alpha,Q,k} K_{ij,k}(Q, Q').} \tag{194}$$

This indicates that the evolution of one strategy is determined by the gradient coupling with trajectories from all training problems.

For a strategy $S_{\beta,Q'}$, define its conditional mean effective coefficient and its mean score by

$$a_{\beta,Q',k} \triangleq \sum_{j \in S_{\beta,Q'}} q_{j|\beta,Q',k} a_{j,k}(Q'), \qquad \overline{U}_{\beta,Q',k} \triangleq \sum_{j \in S_{\beta,Q'}} q_{j|\beta,Q',k} U_{j,k}(Q'). \tag{195}$$

Using $K_{ij,k}(Q, Q') = U_{i,k}(Q)^\top U_{j,k}(Q')$, the strategy fitness can then be written as

$$f_{\alpha,Q,k} = \frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta=1}^{m_{Q'}} \nu_{\beta,Q',k} a_{\beta,Q',k} \overline{U}^\top_{\alpha,Q,k} \overline{U}_{\beta,Q',k} + R^{(a)}_{\alpha,Q,k}, \tag{196}$$

where $R^{(a)}_{\alpha,Q,k}$ is the within-strategy coefficient factorization error quantified below. The strategy evolution is

$$\Delta \nu_{\alpha,Q,k} = \eta \nu_{\alpha,Q,k} f_{\alpha,Q,k} + \xi_{\alpha,Q,k}. \tag{197}$$

Before proceeding, we combine the adiabatic results in **Step 3** to show that $a_{\beta,Q',k}$ is also slow relative to the strategy mass in GRPO, GSPO, and DAPO.

Lemma 10 (The adiabatic property of $a_{\beta,Q',k}$)**.** *Let ω_a be the relative changing rate of $a_{\beta,Q',k}$ and retain ω_ν for that of $\nu_{\beta,Q',k}$. Then*

$$\frac{\omega_a}{\omega_\nu} \ll 1 \tag{198}$$

Proof

Recall that

$$a_{\beta,Q',k} = \sum_{j \in S_{\beta,Q'}} q_{j|\beta,Q',k} \bar{A}^{\text{eff}}_{j,k} r_{j,k}. \tag{199}$$

In GRPO, GSPO scenarios, we have already obtained:

$$\omega_q \ll \omega_\nu \tag{200}$$

$$\omega_r \ll \omega_\nu \tag{201}$$

$$\omega_U \sim \omega_K \ll \omega_\nu \tag{202}$$

$$\omega_A \ll \omega_\nu \tag{203}$$

Consequently, ω_a is controlled by $\omega_q + \omega_A + \omega_r$ and hence $\omega_a \ll \omega_\nu$ for GRPO and GSPO. For DAPO scenario, the previous analysis cannot be fully repeated, the difficulty lies in ω_r. Instead we seek another approach by $\omega_{\mathcal{R}}$. First we recall the definition of $\mathcal{R}_{ij}$:

$$\mathcal{R}^{\text{DAPO}}_{ij} \triangleq \frac{G}{Z} \sum_t m_{j,t} \rho_{j,t} U_i^\top u_{j,t} = \rho_j K_{ij}(Q, Q') r^{DAPO}_{ij} \tag{204}$$

Note that $\omega_K \ll \omega_\nu$, $\omega_\rho \ll \omega_\nu$ and $\omega_{\mathcal{R}} \ll \omega_\nu$, we have:

$$\omega^{DAPO}_r = (\omega_{\mathcal{R}} - \omega_K - \omega_\rho) \sim \omega_{\mathcal{R}} \ll \omega_\nu \tag{205}$$

Then following the standard procedure before, we also have:

$$\omega_a^{DAPO} \ll \omega_\nu \tag{206}$$

This lemma allows $a_{\beta,Q',k}$ to be frozen on the local validity window. Below, $a_{\beta,Q'}$ denotes this frozen value; all remaining k-dependence is carried by the strategy masses and mean scores. To facilitate our following analysis, we prove another generic property of $a_{\beta,Q'}$ and $a_{\alpha,Q}$:

Lemma 11 (Impossibility of exact equivalence)**.** *Let $\{a_{\alpha,Q}\}_{\alpha=1}^{m_{\epsilon,Q}}$ be the fitness set of strategies partitioned on Q. The fitness satisfies a generic property:*

$$\mathbb{P}(a_{\alpha,Q} = a_{\beta,Q}) = 0, \qquad \forall\, \alpha \neq \beta. \tag{207}$$

Proof

Proof. The trajectory-level fitness $a_{j,Q} = \bar{A}_{j,k}^{\text{eff}} r_{j,k}$ is piecewise continuous in the underlying importance ratio $\rho_{j,k}$. Clipping may create an atom at zero for an individual trajectory, but the strategy fitness

$$a_{\alpha,Q} = \sum_{j \in S_{\alpha,Q}} q_{j|\alpha,Q}\, a_{j,Q} \tag{208}$$

averages over a countably infinite family of trajectories. Except for the degenerate event that all effective contributions are clipped or structurally cancel, a strategy therefore contains an active trajectory with positive weight whose fitness varies continuously.
For two distinct strategies $\alpha \neq \beta$, suppose that $S_{\alpha,Q}$ contains an active trajectory $j^\star$ with $q_{j^\star|\alpha,Q} > 0$. Fix all other trajectory contributions and denote the varying contribution of $j^\star$ by x. Then the two strategy fitnesses can be written as

$$a_{\alpha,Q} = c + q_{j^\star|\alpha,Q} x, \qquad a_{\beta,Q} = d, \tag{209}$$

where c and d are fixed by the remaining trajectories. Therefore, $a_{\alpha,Q} = a_{\beta,Q}$ requires

$$x = \frac{d - c}{q_{j^\star|\alpha,Q}}, \tag{210}$$

i.e., one exact value of a continuously varying quantity. Such an event has probability **zero**. Thus, unless all active contributions vanish, cancel identically, or the two strategies are related by an exact structural symmetry,

$$\mathbb{P}(a_{\alpha,Q} = a_{\beta,Q}) = 0, \qquad \forall\, \alpha \neq \beta. \tag{211}$$

□

We then construct the potential function. With the coefficients frozen as above, define

$$\Phi_k \triangleq \frac{1}{|\mathcal{D}|} \sum_{Q \in \mathcal{D}} \sum_{\alpha=1}^{m_Q} a_{\alpha,Q} \nu_{\alpha,Q,k}. \tag{212}$$

Also define the two strategy-level empirical vectors

$$\begin{aligned} \widehat{g}_{\mathcal{D},k} &\triangleq \frac{1}{|\mathcal{D}|} \sum_{Q \in \mathcal{D}} \sum_{\alpha=1}^{m_Q} a_{\alpha,Q} \nu_{\alpha,Q,k} \overline{U}_{\alpha,Q,k}, \\ \widehat{g}_{\mathcal{D}',k} &\triangleq \frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta=1}^{m_{Q'}} a_{\beta,Q'} \nu_{\beta,Q',k} \overline{U}_{\beta,Q',k}. \end{aligned} \tag{213}$$

Substituting the strategy fitness into Eq. equation 212 gives

$$\Delta\Phi_k = \eta D_k + \Xi_k, \qquad D_k \triangleq \langle \widehat{g}_{\mathcal{D},k}, \widehat{g}_{\mathcal{D}',k} \rangle + R_k^{(a)}, \tag{214}$$

where

$$R_k^{(a)} \triangleq \frac{1}{|\mathcal{D}|} \sum_{Q \in \mathcal{D}} \sum_{\alpha=1}^{m_Q} a_{\alpha,Q} \nu_{\alpha,Q,k} R_{\alpha,Q,k}^{(a)},$$
$$\Xi_k \triangleq \frac{1}{|\mathcal{D}|} \sum_{Q \in \mathcal{D}} \sum_{\alpha=1}^{m_Q} a_{\alpha,Q} \xi_{\alpha,Q,k}. \tag{215}$$

Thus D_k is the normalized deterministic drift, $R_k^{(a)}$ is the error caused by replacing the trajectory coefficients inside a strategy by their conditional mean, and Ξ_k is the martingale noise.

Lemma 12 (Strategy-factorization error)**.** *On the frozen window, suppose $|a_{i,k}(Q)| \le A$ and $|K_{ij,k}(Q,Q')| \le K_{\max}$. Apply Lemma 9, so that trajectories $j, j_\beta \in S_{\beta,Q'}$ satisfy*

$$\sup_{Q,i} \left| K_{ij,k}(Q,Q') - K_{ij_\beta,k}(Q,Q') \right| \le \mathcal{O}(\epsilon). \tag{216}$$

Then

$$b_{a,k} \triangleq |R_k^{(a)}| \le 2A^2 \mathcal{O}(\epsilon) = \mathcal{O}(\epsilon). \tag{217}$$

Proof

Proof. For $j \in S_{\beta,Q'}$, write $a_{j,k}(Q') = a_{\beta,Q'} + d_{j,Q',k}$. By the definition of the conditional mean, $\sum_j q_{j|\beta,Q',k} d_{j,Q',k} = 0$. Hence, for every opposing trajectory $i \in S_{\alpha,Q}$,

$$\begin{aligned} &\left| \sum_{j \in S_{\beta,Q'}} q_{j|\beta,Q',k} d_{j,Q',k} K_{ij,k}(Q,Q') \right| \\ &= \left| \sum_{j \in S_{\beta,Q'}} q_{j|\beta,Q',k} d_{j,Q',k} \left(K_{ij,k}(Q,Q') - K_{ij_\beta,k}(Q,Q') \right) \right| \\ &\le \epsilon K_{\max} \sum_{j \in S_{\beta,Q'}} q_{j|\beta,Q',k} |d_{j,Q',k}| \le 2A\mathcal{O}(\epsilon). \end{aligned} \tag{218}$$

Averaging over both datasets and using $|a_{\alpha,Q}| \le A$ proves Eq. equation 217. □

Since $\mathcal{D}$ and $\mathcal{D}'$ are sampled from the same task distribution, both vectors in Eq. equation 213 estimate $g_k \triangleq \mathbb{E}_{Q \sim \mathcal{H}} \left[\sum_{\alpha=1}^{m_Q} a_{\alpha,Q} \nu_{\alpha,Q,k} \overline{U}_{\alpha,Q,k} \right]$. Write $\delta_{\mathcal{D},k} \triangleq \widehat{g}_{\mathcal{D},k} - g_k, \qquad \delta_{\mathcal{D}',k} \triangleq \widehat{g}_{\mathcal{D}',k} - g_k.$

Lemma 13 (Finite-data error)**.** *Suppose that the query-wise vector inside Eq. equation A.2.4 has norm at most R on the frozen window. Conditional on the frozen state, suppose also that the queries entering the two empirical vectors are independent samples from $\mathcal{H}$; for a reused training dataset, it suffices instead to assume the corresponding uniform or stability-based generalization bound. Then*

$$\begin{aligned} b_{\text{data},k} &\triangleq \|g_k\| \left(\|\delta_{\mathcal{D},k}\| + \|\delta_{\mathcal{D}',k}\| \right) + \|\delta_{\mathcal{D},k}\| \|\delta_{\mathcal{D}',k}\| \\ &= \mathcal{O}_{\mathbb{P}}\left(|\mathcal{D}|^{-1/2} + |\mathcal{D}'|^{-1/2} \right). \end{aligned} \tag{219}$$

Uniformly over $0 \le k < T_c$, the same bound holds with the usual factor $\sqrt{\log(T_c/\rho)}$ at confidence $1 - \rho$.

Proof

Proof. Hilbert-space concentration gives, with probability at least $1 - \rho$,

$$\|\delta_{\mathcal{D},k}\| \le cR \sqrt{\frac{\log(4/\rho)}{|\mathcal{D}|}}, \qquad \|\delta_{\mathcal{D}',k}\| \le cR \sqrt{\frac{\log(4/\rho)}{|\mathcal{D}'|}} \tag{220}$$

for a universal constant c. Expanding the inner product and applying Cauchy–Schwarz yields

$$\begin{aligned}\langle \widehat{g}_{\mathcal{D},k}, \widehat{g}_{\mathcal{D}',k} \rangle &= \|g_k\|^2 + \langle g_k, \delta_{\mathcal{D},k} \rangle + \langle g_k, \delta_{\mathcal{D}',k} \rangle + \langle \delta_{\mathcal{D},k}, \delta_{\mathcal{D}',k} \rangle \\ &\geq \|g_k\|^2 - b_{\mathrm{data},k}.\end{aligned} \tag{221}$$

A union bound over the finite window proves the uniform statement. □

Lemma 14 (Almost sure positivity of the population drift)**.** *Fix k and the frozen coefficients. Suppose that, for at least one query of positive $\mathcal{H}$-mass, the active strategy coefficients are not all equal. Then, with respect to Lebesgue measure on the admissible strategy-score configurations satisfying the score identities,*

$$\boxed{\|g_k\|^2 > 0 \quad a.s.} \tag{222}$$

Consequently, the conclusion holds almost surely under any geometry law that is absolutely continuous with respect to this Lebesgue measure.

Proof

Proof. For each query, the score identity gives

$$\sum_{\alpha=1}^{m_Q} \nu_{\alpha,Q,k} \overline{U}_{\alpha,Q,k} = 0. \tag{223}$$

Consequently,

$$g_k = \mathbb{E}_{Q\sim\mathcal{H}} \left[\sum_{\alpha=1}^{m_Q-1} \nu_{\alpha,Q,k} \left(a_{\alpha,Q} - a_{m_Q,Q} \right) \overline{U}_{\alpha,Q,k} \right]. \tag{224}$$

Because the coefficients are not all equal, at least one coefficient in Eq. equation 224 is nonzero. Thus g_k is a nonzero linear function of the free strategy-score vectors. Its zero set is the kernel of a nonzero linear map, hence a proper linear subspace and therefore a Lebesgue-null set. Outside this set $g_k \neq 0$, which is equivalent to $\|g_k\|^2 > 0$. □

Before officially using Azuma-Hoeffding inequality, we first prove a lemma showing this inequality can still be applied to almost surely positive drifts.

Lemma 15 (Azuma under almost-surely positive drift)**.**
Consider a process on the fixed finite horizon $0 \leq k < n \leq T_c$ satisfying $X_{k+1} - X_k = D_k + \xi_k$, where $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$, $|\xi_k| \leq b$, and $D_k > 0$ almost surely for every $k < T_c$. Then, outside a null event, the finite drift sequence admits a lower bound $c > 0$, and for every $t > 0$,

$$\mathbb{P}\left(X_n - X_0 \leq nc - t\right) \leq \mathbb{P}\left(\sum_{k=0}^{n-1} \xi_k \leq -t \right). \tag{225}$$

Consequently, the usual Azuma–Hoeffding bound applies on this full-measure event.

Proof

Proof. For every $k < T_c$, introduce the null event N_k, defined by

$$N_k = \{D_k = 0\}, \qquad \mathbb{P}(N_k) = 0. \tag{226}$$

Because the horizon is finite, define

$$N = \bigcup_{i=0}^{T_c-1} N_i, \qquad \mathbb{P}(N) = 0. \tag{227}$$

On N^c, every member of the finite sequence $D_0, \ldots, D_{T_c-1}$ is strictly positive. Hence its minimum exists and satisfies

$$c = \min_{0 \le k < T_c} D_k > 0. \tag{228}$$

Let $A = \{X_n - X_0 \le nc - t\}$. Since $\mathbb{P}(N) = 0$, the correct decomposition is

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap N) + \mathbb{P}(A \cap N^c) \\ &= \mathbb{P}(A \cap N^c). \end{aligned} \tag{229}$$

On $A \cap N^c$, we have

$$X_n - X_0 = \sum_{k=0}^{n-1} D_k + \sum_{k=0}^{n-1} \xi_k \ge nc + \sum_{k=0}^{n-1} \xi_k. \tag{230}$$

Therefore,

$$A \cap N^c \subseteq \left\{ \sum_{k=0}^{n-1} \xi_k \le -t \right\}, \tag{231}$$

which gives

$$\mathbb{P}\left(X_n - X_0 \le nc - t\right) \le \mathbb{P}\left(\sum_{k=0}^{n-1} \xi_k \le -t \right). \tag{232}$$

This proves that the null degeneracy set does not affect the Azuma–Hoeffding argument on the fixed finite horizon. □

Combining Eqs. equation 214, equation 217, and equation 221 gives the pointwise bound

$$\boxed{D_k \ge \|g_k\|^2 - b_{a,k} - b_{\text{data},k}.} \tag{233}$$

Fix a target query Q_0 and $\zeta > 0$, and let

$$\tau_\zeta \triangleq \inf \left\{ 0 \le k \le T_c : \max_{1 \le \alpha \le m_{Q_0}} \nu_{\alpha,Q_0,k} \ge 1 - \zeta \right\}, \tag{234}$$

with $\inf \varnothing = T_c + 1$. The pre-collapse indices are the finite set

$$\mathcal{I}_{\zeta,Q_0} \triangleq \{0 \le k < T_c : k < \tau_\zeta\}. \tag{235}$$

Outside the finite union of the null sets in Lemma 14, either this set is empty or

$$C' \triangleq \min_{k \in \mathcal{I}_{\zeta,Q_0}} \|g_k\|^2 > 0. \tag{236}$$

Therefore, reconsidering the drift term D we have:

$$D_k \ge C' - b_{data,k} - b_{a,k} = C' - \mathcal{O}(\epsilon) - \mathcal{O}(|D|^{-1/2} + |D'|^{-1/2}) \triangleq C > 0, a.s. \tag{237}$$

Fix any $\zeta > 0$, and freeze $a_{Q,\alpha}$ throughout the following analysis. Since $\{\nu_{\alpha,Q_0,k}\}_{\alpha=1}^{m_{Q_0}}$ lies in the probability simplex, the potential is uniformly bounded above by

$$\Phi_k \le \frac{1}{|\mathcal{D}|} \sum_{Q} \max_{\alpha \in Q} a_{Q,\alpha}. \tag{238}$$

We introduce the indicator

$$\xi_k = \mathbf{1}\left\{ \max_{\alpha} \nu_{\alpha,Q_0,k} < 1 - \zeta \right\}. \tag{239}$$

Thus, $\xi_k = 1$ exactly when the strategy distribution remains outside the ζ-collapse region.

By Lemma 14 and the finite-horizon argument above, the deterministic drift is almost surely strictly positive at every pre-collapse step in the closed region

$$\max_{\alpha} \nu_{\alpha,Q_0,k} \le 1 - \zeta. \tag{240}$$

Consequently, for this fixed ζ and fixed finite horizon, the realized finite drift sequence admits an almost-sure lower bound $C > 0$, so the deterministic drift is at least ηC whenever $\xi_k = 1$. Outside this region, the deterministic drift remains non-negative. Therefore,

$$\Delta\Phi_k \geq \eta C \xi_k + \Xi_k, \tag{241}$$

where

$$\mathbb{E}[\Xi_k \mid \mathcal{F}_k] = 0, \qquad |\Xi_k| \leq \eta b. \tag{242}$$

Summing Eq. equation 241 from $k = 0$ to $K - 1$ gives

$$\Phi_K - \Phi_0 \geq \eta C \sum_{k=0}^{K-1} \xi_k + \sum_{k=0}^{K-1} \Xi_k. \tag{243}$$

In particular, on the event that the dynamics remains outside the ζ-collapse region throughout the first K steps, we have $\xi_0 = \cdots = \xi_{K-1} = 1$, and hence

$$\Phi_K - \Phi_0 \geq K\eta C + \sum_{k=0}^{K-1} \Xi_k. \tag{244}$$

Combining this inequality with Eq. equation 238, such an event necessarily implies

$$\sum_{k=0}^{K-1} \Xi_k \leq \frac{1}{|\mathcal{D}|} \sum_{Q} \max_{\alpha \in Q} a_{Q,\alpha} - \Phi_0 - K\eta C. \tag{245}$$

If

$$K \geq \frac{2}{\eta C} \left(\frac{1}{|\mathcal{D}|} \sum_{Q} \max_{\alpha \in Q} a_{Q,\alpha} - \Phi_0 \right), \tag{246}$$

then Eq. equation 245 implies

$$\sum_{k=0}^{K-1} \Xi_k \leq -\frac{1}{2} K\eta C. \tag{247}$$

By the Azuma–Hoeffding inequality,

$$\begin{aligned} &\mathbb{P}\left(\xi_0 = \cdots = \xi_{K-1} = 1\right) \\ &\leq \mathbb{P}\left(\sum_{k=0}^{K-1} \Xi_k \leq -\frac{1}{2} K\eta C \right) \\ &\leq \exp\left(-\frac{\left(\frac{1}{2} K\eta C\right)^2}{2K\eta^2 b^2} \right) = \exp\left(-\frac{KC^2}{8b^2} \right). \end{aligned} \tag{248}$$

Therefore, if

$$\boxed{K \geq \max\left\{ \frac{2}{\eta C} \left(\frac{1}{|\mathcal{D}|} \sum_{Q} \max_{\alpha \in Q} a_{Q,\alpha} - \Phi_0 \right), \frac{8b^2}{C^2} \log\left(\frac{2}{\delta}\right) \right\},} \tag{249}$$

then

$$\boxed{\mathbb{P}\left(\exists k < K : \max_{\alpha} \nu_{\alpha, Q_0, k} \geq 1 - \zeta \right) \geq 1 - \frac{\delta}{2}.} \tag{250}$$

We further estimates the scale of K to show that it is much smaller than our focused validity horizon $T_c \sim \mathcal{O}(\frac{1}{\eta})$. Before that, we recall from **Step 3** that:

$$U_{j,k} = \Theta(|\tau|) \tag{251}$$

Recalling the definition of $\|g_k\|^2$, we found that $U_{j,k}$ appears here in a quadric form, therefore we have:

$$C \sim \|g_k\|^2 \sim \|U_{i,k}\|^2 = \Theta(|\tau|^2) \tag{252}$$

Note that $a_{Q,\alpha}$ is basically the clipped version of RLVR advantage, its scale is

$$a_{Q,\alpha} \sim \mathcal{O}(1) \tag{253}$$

Therefore, so long as $\eta^{-1} \ll |\tau|^3$, the first term can be estimated as:

$$\boxed{\frac{2}{\eta C}\left(\frac{1}{|\mathcal{D}|}\sum_{Q\in\mathcal{D}}\max_{S_\alpha\in Q} a_{Q,\alpha} - \Phi_0\right) \sim \mathcal{O}(\eta^{-1}\cdot|\tau|^{-2}) \ll \mathcal{O}(|\tau|) \ll \mathcal{O}(\eta^{-1}) \sim T_c} \tag{254}$$

Considering the second term, note that $\delta > \eta$, we have:

$$\frac{8b^2}{C^2}\log\left(\frac{2}{\delta}\right) \sim \mathcal{O}(\log(\eta^{-1})) \ll T_c \tag{255}$$

This result of the burn-in time $K \ll \mathcal{O}(|\tau|)$ manifests that the near-collapsed strategy distribution can be achieved during the adiabatic period, meaning that our analysis is self-consistent in the scale-separation operation. We further show that the near-collapse distribution can be preserved throughout the fixed horizon T_c, known as the **preservation property**.

A.2.5 Step 5: Preservation after the adiabatic window.

Back Apply the preceding hitting result, we set an extreme ζ' to $\mathcal{O}(\eta)$ and with failure probability $\delta/2$. For each query Q, let o be the strategy that carries the maximal mass at the end of this hitting stage, and keep this index fixed throughout the argument below. For a finite collection of queries, take K_Q to be the hitting time of each query Q, and let the maximum of these time as $K \triangleq \max_Q K_Q$. Absorbing the constant in the tightened tolerance into η, the hitting result gives

$$\mathbb{P}\left(\nu_{o,Q,K} \geq 1 - \mathcal{O}(\eta)\right) \geq 1 - \frac{\delta}{2}, \qquad \nu^\star_{Q,k} \triangleq 1 - \nu_{o,Q,k}. \tag{256}$$

It remains to show that the fixed strategy o cannot lose its dominance before the validity horizon T_c except on an event of probability at most $\delta/2$.

After the first adiabatic period $T_{ad} = \mathcal{O}(|\tau|)$, the strategy fitnesses can no longer be frozen. Retaining the notation of the preceding derivation, define

$$\Delta_{Q,k} \triangleq \frac{1}{1-\nu^\star_{Q,k}}\sum_{\alpha\neq o}\nu_{\alpha,Q,k}f_{\alpha,Q,k} - f_{o,Q,k}. \tag{257}$$

The exact strategy normalization identity gives

$$\sum_{\alpha\neq o}\nu_{\alpha,Q,k}f_{\alpha,Q,k} + (1-\nu^\star_{Q,k})f_{o,Q,k} = 0. \tag{258}$$

Consequently,

$$\Delta_{Q,k} = -2f_{o,Q,k}, \qquad \sum_{\alpha\neq o}\nu_{\alpha,Q,k}f_{\alpha,Q,k} = \frac{1}{2}(1-\nu^\star_{Q,k})\Delta_{Q,k}. \tag{259}$$

To bound this gap, recall the strategy-level score identity

$$(1-\nu^\star_{Q,k})\overline{U}_{o,Q,k} + \sum_{\alpha\neq o}\nu_{\alpha,Q,k}\overline{U}_{\alpha,Q,k} = 0. \tag{260}$$

Up to the first possible violation of the preservation event, the tightened tolerance gives $\nu^\star_{Q,k} \leq \zeta' = \mathcal{O}(\eta)$. Together with $\|\overline{U}_{\alpha,Q,k}\| = \mathcal{O}(|\tau|)$, this implies

$$\|\overline{U}_{o,Q,k}\| \leq \frac{\nu^\star_{Q,k}}{1-\nu^\star_{Q,k}}\mathcal{O}(|\tau|) = \mathcal{O}(\eta|\tau|). \tag{261}$$

Substituting this estimate into

$$f_{o,Q,k} = \frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta} \nu_{\beta,Q',k} a_{\beta,Q',k} \overline{U}_{o,Q,k}^{\top} \overline{U}_{\beta,Q',k} + R^{(a)}_{o,Q,k} \tag{262}$$

The rest part can also be estimated, first consider:

$$\sum_{\beta} \nu_{\beta,Q',k} \cdot \bar{U}_{\beta,Q',k} = 0 \tag{263}$$

Let the optima in Q' space be o', then:

$$\begin{aligned} &\frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta} \nu_{\beta,Q',k} a_{o',Q',k} \overline{U}_{\beta,Q',k} \\ &\quad = \frac{1}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta \neq o'} \nu_{\beta,Q',k} (a_{\beta,Q',k} - a_{o',Q',k}) \overline{U}_{\beta,Q',k} \\ &\quad \leq \mathcal{O}(\eta) \cdot \mathcal{O}(1) \cdot \mathcal{O}(|\tau|) = \mathcal{O}(\eta|\tau|) \end{aligned} \tag{264}$$

and absorbing the higher-order factorization remainder yields

$$|\Delta_{Q,k}| \leq \bar{\Delta}, \qquad \bar{\Delta} = \mathcal{O}(\eta^2 |\tau|^2) \ll 1. \tag{265}$$

We now establish the martingale bound directly for the dominant mass $\nu_{o,Q,k} = 1 - \nu^{\star}_{Q,k}$. Its evolution satisfies

$$\begin{aligned} \Delta\nu^{\star}_{Q,k} &= \eta \sum_{\alpha \neq o} \nu_{\alpha,Q,k} f_{\alpha,Q,k} + \Xi^{\star}_k = \frac{\eta}{2} \nu_{o,Q,k} \Delta_{Q,k} + \Xi^{\star}_k, \\ \nu_{o,Q,k+1} &= \left(1 - \frac{\eta}{2} \Delta_{Q,k}\right) \nu_{o,Q,k} - \Xi^{\star}_k. \end{aligned} \tag{266}$$

Because $\Delta_{Q,k} \leq \bar{\Delta}$,

$$\nu_{o,Q,k+1} \geq \left(1 - \frac{\eta}{2} \bar{\Delta}\right) \nu_{o,Q,k} - \Xi^{\star}_k. \tag{267}$$

Let

$$A_k = \left(1 - \frac{\eta}{2} \bar{\Delta}\right)^{k-K}, \qquad K \leq k \leq T_c. \tag{268}$$

The scale condition guarantees $1 - \eta\bar{\Delta}/2 > 0$. Iterating Eq. equation 267 gives

$$\frac{\nu_{o,Q,k}}{A_k} \geq \nu_{o,Q,K} - \sum_{i=K}^{k-1} \frac{\Xi^{\star}_i}{A_{i+1}}. \tag{269}$$

Moreover,

$$0 < a \triangleq A_{T_c} \leq A_k \leq 1, \qquad a \sim e^{-\bar{\Delta}/2} \approx 1. \tag{270}$$

Retain the existing martingale notation

$$M_k = \sum_{i=K}^{k-1} \frac{\Xi^{\star}_i}{A_{i+1}}. \tag{271}$$

Since $\mathbb{E}[\Xi^{\star}_i \mid \mathcal{F}_i] = 0$ and $|\Xi^{\star}_i| \leq b\eta$, its increments satisfy

$$\left| \frac{\Xi^{\star}_i}{A_{i+1}} \right| \leq \frac{b\eta}{a}. \tag{272}$$

On the event in Eq. equation 256, if $\nu_{o,Q,k} < 1 - \zeta$ for some $K < k \leq T_c$, then Eqs. equation 269 and equation 270 imply

$$M_k > 1 - \mathcal{O}(\eta) - \frac{1 - \zeta}{A_k} \geq \frac{a(1 - \mathcal{O}(\eta)) - (1 - \zeta)}{a}. \tag{273}$$

Under the stated scale condition, the tightened hitting tolerance is chosen so that $a(1-\mathcal{O}(\eta)) > 1-\zeta$. The maximal Azuma–Hoeffding inequality and $T_c - K \le \eta^{-1}$ therefore give

$$\begin{aligned}
&\mathbb{P}\left(\exists k \in \mathbb{Z}^+,\ K < k \le T_c : \nu_{o,Q,k} < 1-\zeta \,\middle|\, \nu_{o,Q,K} \ge 1-\mathcal{O}(\eta)\right) \\
&\quad \le \mathbb{P}\left(\max_{K<k\le T_c} M_k \ge \frac{a(1-\mathcal{O}(\eta)) - (1-\zeta)}{a}\right) \\
&\quad \le \exp\left(-\frac{\left(a(1-\mathcal{O}(\eta)) - (1-\zeta)\right)^2}{2b^2\eta}\right) = \mathcal{O}(\exp(-\eta^{-1})).
\end{aligned} \tag{274}$$

Considering that $\delta = \Omega(\eta)$, we have:

$$\frac{\delta}{2} = \Omega(\eta) \gg \mathcal{O}(\exp(-\eta^{-1})) \ge \mathbb{P}\left(\max_{K<k\le T_c} M_k \ge \frac{a(1-\mathcal{O}(\eta)) - (1-\zeta)}{a}\right) \tag{275}$$

Therefore,

$$\mathbb{P}\left(\forall k \in \mathbb{Z}^+,\ K < k \le T_c, \nu_{o,Q,k} \ge 1-\zeta \,\middle|\, \nu_{o,Q,K} \ge 1-\zeta'\right) \ge 1 - \frac{\delta}{2}. \tag{276}$$

Combining this bound with Eq. equation 256 gives

$$\begin{aligned}
&\mathbb{P}\left(\forall k \in \mathbb{Z}^+,\ K < k \le T_c, \nu_{o,Q,k} \ge 1-\zeta\right) \\
&\quad \ge \mathbb{P}\left(\forall k \in \mathbb{Z}^+,\ K < k \le T_c, \nu_{o,Q,k} \ge 1-\zeta \,\middle|\, \nu_{o,Q,K} \ge 1-\zeta'\right) \cdot \mathbb{P}\left(\nu_{o,Q,K} \ge 1-\zeta'\right) \\
&\quad \ge \left(1 - \frac{\delta}{2}\right)^2 > 1-\delta
\end{aligned} \tag{277}$$

Finally, recall that the partition sensitivity ϵ used to define S_α was never restricted throughout the derivation above. In particular, none of the preceding arguments depends on any specific choice of ϵ. Therefore, ϵ is arbitrary, and the same conclusion holds for every admissible partition sensitivity. Hence,

$$\begin{aligned}
&\forall \epsilon > 0, \epsilon \ll 1, \zeta, \delta \in (\eta, 1), \zeta^2 \gg \eta \quad \exists K \in \mathbb{Z}^+, K \ll |\tau| < T_c, \\
&\forall k \in \mathbb{Z}^+, K < k \le T_c, \quad \mathbb{P}\left(\max_{1\le\alpha\le m_{\epsilon,Q}} \nu^{(\epsilon)}_{\alpha,Q,k} \ge 1-\zeta\right) \ge 1-\delta.
\end{aligned} \tag{$\star$}$$

This corresponds to the concentration of most strategy probability mass on a single dominant strategy, completing the proof. ■

A.2.6 The Consequences of Strategy Dynamics

Since the strategy-level evolution equations derived for GRPO, GSPO, and DAPO share the same dynamical structure as Equation ($\star$), the analysis developed in Sec. **A2.3** applies uniformly to all three RLVR objectives. Moreover, as established above, strategy-level probability redistribution occurs on a faster timescale than the internal rearrangement of trajectories within each strategy. Consequently, the dominant training dynamics are governed by competition among strategies, while the within-strategy trajectory distribution evolves only comparatively slowly.

This timescale separation allows the long-term behavior of the policy to be characterized directly at the strategy level. Under GRPO, GSPO, and DAPO, differences in strategy advantage are progressively amplified, causing probability mass to concentrate toward an increasingly dominant strategy. In particular, once one strategy acquires a relative advantage, the resulting positive feedback further increases its probability and suppresses competing strategies, eventually driving the policy toward strategy concentration.

Therefore, all three objectives systematically amplify the model's existing preference over strategies. The strategy that is most favorable under the current model state—i.e., the one with the highest effective evolutionary advantage—is preferentially reinforced and becomes increasingly dominant, providing the dynamical mechanism underlying strategy collapse.

A.3 PROOF FOR THEOREM 2

Back **Proof Sketch for Theorem 2.** Let $S_{\epsilon,k}$ be the realized random strategy at step k. We start from analyzing strategy entropy $H(S_{\epsilon,k})$ and consider the expansion $H(S_{\epsilon,k}) = H(S_{\epsilon,k}|Q) + I(S_{\epsilon,k}; Q)$. Following the two-step procedure below, we seek a lower bound for each term on the right hand side.

Step 1: Controlling the Conditional Entropy. Consider the definition of conditional entropy, $H(S_{\epsilon,k}|Q) = \mathbb{E}_{Q\sim\mathcal{D}}[H(S_{\epsilon,k}|Q=Q)] = -\mathbb{E}_{Q\sim\mathcal{D}}[\sum_\alpha \nu_{\alpha,Q,k} \log \nu_{\alpha,Q,k}]$. Introducing the pre-RLVR distribution as a reference, we have:

$$H(S_{\epsilon,k}|Q) = -\mathbb{E}_{Q\sim\mathcal{D}}\left[D_{KL}(\nu_{Q,k}\|\nu_{Q,0})\right] - \mathbb{E}_{Q\sim\mathcal{D}}\left[\sum_\alpha \nu_{\alpha,Q,k}\log\nu_{\alpha,Q,0}\right] \tag{278}$$

According to the adiabatic properties established in Lemma 4 of Appendix A.2.3, we prove that $D_{KL}(\nu_{Q,k}\|\nu_{Q,0}) = \mathcal{O}(\eta^2|\tau|^2)$, where $|\tau|$ is the generated trajectory length (Appendix A.3.2). According to our requirements for the pre-RLVR model, we have $\mathbb{E}_{Q\sim\mathcal{D}}\left[\sum_\alpha \nu_{\alpha,Q,k}\log\nu_{\alpha,Q,0}\right] \le \log_2 \frac{(1+\rho)^2}{4\rho m_\epsilon}$ (Appendix A.3.3), thereby controlling the conditional entropy.

Step 2: Controlling the Mutual Information. Let $A \triangleq \mathbf{1}\{\widehat{Y}(\tau) = Y(Q)\}$, where τ is sampled from the realized strategy $S_{\epsilon,k}$, so that $\mathbb{P}(A=1) = P_{\text{acc},k}$. Since $\mathcal{D}$ and all distinct ground-truth answers $\{Y(Q') : Q' \in \mathcal{D}\}$ are known, whenever $A=1$ the terminal answer $\widehat{Y}(\tau) = Y(Q)$ uniquely identifies the originating query, **unless** the same realized strategy contains a trajectory whose answer collides with the ground truth of another query $Q' \neq Q$. Under the γ-separation condition, this collision occurs with probability at most γ. Hence we construct a collision-aware decoder $\widehat{Q}_k(S_{\epsilon,k})$ whose error probability satisfies

$$\mathbb{P}\Big(\widehat{Q}_k(S_{\epsilon,k}) \neq Q\Big) \le 1 - P_{\text{acc},k} + \gamma. \tag{279}$$

Applying Fano's inequality then gives the lower bound (Appendix A.3.4):

$$I(Q; S_{\epsilon,k}) \ge (P_{acc,k} - \gamma)\log_2 N - h_2(P_{acc,k} - \gamma) \tag{280}$$

Combining these two results, we conclude $|\mathcal{S}_{\epsilon,k}| = \Omega_{\rho,\gamma}(m_\epsilon N^{P_{acc,k}-\gamma})$. After simple scale analysis, the coefficient of this bound is around 2ρ, establishing this as a nontrivial bound. ■

We now prove the statement. All logarithms in this subsection are base two, and Q is uniform on the N evaluation queries in $\mathcal{D}$.

A.3.1 ρ-ADMISSIBLE STRATEGY RESOLUTION

The order of quantifiers is essential: first fix ρ, and only then choose ϵ. We call ϵ *ρ-admissible* if the strategy partition at step 0 satisfies

$$\nu^{(\epsilon)}_{\alpha,Q,0} \ge \rho \tag{281}$$

for every $Q \in \mathcal{D}$ and every $S^{(\epsilon)}_\alpha$ in its induced partition. The partition remains fixed during optimization on $\mathcal{D}'$, and we write $\nu^{(\epsilon)}_{\alpha,Q,k} = \mathbb{P}(S_{\epsilon,k} = S^{(\epsilon)}_{\alpha,Q} \mid Q)$ for its step-k mass. For each query,

$$1 = \sum_{\alpha=1}^{m_{\epsilon,Q}} \nu^{(\epsilon)}_{\alpha,Q,0} \ge m_{\epsilon,Q}\rho, \tag{282}$$

and consequently $m_{\epsilon,Q} \le \lfloor \rho^{-1} \rfloor$ for all $Q \in \mathcal{D}$, hence

$$\boxed{\rho \cdot m_\epsilon \le 1.} \tag{283}$$

Thus Theorem 2 is not a claim about arbitrarily fine partitions. It is a capacity statement at a fixed, pre-specified probability resolution.

A.3.2 AN ADIABATIC KL BOUND AT THE CHOSEN RESOLUTION

Define the conditional reverse KL budget at step k by

$$B_k \triangleq \mathbb{E}_Q D_{\text{KL}}(\nu_k(\cdot \mid Q) \,\|\, \nu_0(\cdot \mid Q)). \tag{284}$$

Here step 0 denotes the beginning of the same adiabatic window, and $k \le |\tau|$. Because the minibatch is the training batch, the behavior policy is refreshed after every update:

$$\pi_{\mathrm{old},k} = \pi_{k-1}, \qquad \rho_{i,k} = \frac{\pi_k(\tau_i \mid Q)}{\pi_{k-1}(\tau_i \mid Q)}. \tag{285}$$

Consequently, $\log_2 \rho_{i,k}$ is the first discrete difference of $\log_2 \pi_k(\tau_i \mid Q)$, and its temporal variation is the corresponding second discrete difference. By the adiabatic correction-factor estimate established in Theorem 1,

$$\Delta \log_2 \rho_{i,k} = \mathcal{O}\big(\eta^2 |\tau|\big)\,. \tag{286}$$

Hence, throughout the window,

$$\log_2 \frac{\rho_{i,k}}{\rho_{i,0}} = \mathcal{O}\big(\eta^2 |\tau|^2\big)\,. \tag{287}$$

The one-step relative policy change is $\mathcal{O}(\eta)$. The accumulated effect of the slowly varying correction factor contributes only $\mathcal{O}(\eta^3|\tau|k^2)$ to the endpoint log-likelihood ratio. Therefore, for every $k \le |\tau|$,

$$\sup_{Q,\tau_i} \left| \log_2 \frac{\pi_k(\tau_i \mid Q)}{\pi_0(\tau_i \mid Q)} \right| = \mathcal{O}(\eta k) + \mathcal{O}(k\eta^2) + \mathcal{O}(\eta^3|\tau|k^2) \tag{288}$$

Note that $\eta|\tau| \ll 1$, we have $\mathcal{O}(\eta^2 k) \ll \mathcal{O}(\eta|\tau|)$, $\mathcal{O}(\eta^3|\tau|k^2) = \mathcal{O}(\eta^3|\tau|^3) \ll \mathcal{O}(\eta|\tau|)$, we have:

$$\sup_{Q,\tau_i} \left| \log_2 \frac{\pi_k(\tau_i \mid Q)}{\pi_0(\tau_i \mid Q)} \right| = \mathcal{O}(\eta|\tau|) \tag{289}$$

where the last equality uses $\eta|\tau| \ll 1$. It follows that

$$\frac{\pi_k(\tau_i \mid Q)}{\pi_0(\tau_i \mid Q)} - 1 = \mathcal{O}(\eta|\tau|) \tag{290}$$

uniformly on the same window.

The intra-strategy distribution is frozen by the adiabatic approximation. Thus, for every $\tau_i \in S_\alpha^{(\epsilon)}$,

$$\pi_k(\tau_i \mid Q) = \nu_{\alpha,k}\, q_{i|S_\alpha}(Q), \qquad \pi_0(\tau_i \mid Q) = \nu_{\alpha,0}\, q_{i|S_\alpha}(Q). \tag{291}$$

The trajectory-level and strategy-level divergences are therefore exactly equal:

$$D_{\mathrm{KL}}(\pi_k(\cdot \mid Q) \,\|\, \pi_0(\cdot \mid Q)) = \sum_{\alpha=1}^{m_{\epsilon,Q}} \sum_{\tau_i \in S_\alpha^{(\epsilon)}} \nu_{\alpha,k} q_{i|S_\alpha}(Q) \log_2 \frac{\nu_{\alpha,k} q_{i|S_\alpha}(Q)}{\nu_{\alpha,0} q_{i|S_\alpha}(Q)} \tag{292}$$

$$= \sum_{\alpha=1}^{m_{\epsilon,Q}} \nu_{\alpha,k} \log_2 \frac{\nu_{\alpha,k}}{\nu_{\alpha,0}} \tag{293}$$

$$= D_{\mathrm{KL}}(\nu_k(\cdot \mid Q) \,\|\, \nu_0(\cdot \mid Q))\,. \tag{294}$$

Finally, expanding the trajectory-level divergence around $\pi_k/\pi_0 = 1$, which is the abbreviation of $\pi_k(\cdot|Q)/\pi_0(\cdot|Q)$, gives

$$D_{\mathrm{KL}}(\pi_k(\cdot \mid Q) \,\|\, \pi_0(\cdot \mid Q)) = \mathbb{E}_{\pi_0(\cdot|Q)} \left[\frac{\pi_k}{\pi_0} \log_2 \frac{\pi_k}{\pi_0} \right] \tag{295}$$

The first-order term vanishes exactly because

$$\mathbb{E}_{\pi_0(\cdot|Q)} \left[\frac{\pi_k}{\pi_0} - 1 \right] = 0. \tag{296}$$

Therefore the expansion can be expressed as:

$$D_{\mathrm{KL}}(\pi_k(\cdot \mid Q) \,\|\, \pi_0(\cdot \mid Q)) = \mathbb{E}_{\pi_0(\cdot|Q)} \left[\frac{1}{2} \left(\frac{\pi_k}{\pi_0} - 1 \right)^2 + \mathcal{O}\left(\left| \frac{\pi_k}{\pi_0} - 1 \right|^3 \right) \right]. \tag{297}$$

Combining this identity with Eq. equation 288 and Eq. equation 294 yields

$$D_{\mathrm{KL}}(\nu_k(\cdot \mid Q) \,\|\, \nu_0(\cdot \mid Q)) = \mathcal{O}\big(\eta^2 |\tau|^2\big)\,. \tag{298}$$

After averaging over Q,

$$\boxed{B_k = \mathcal{O}\big(\eta^2 |\tau|^2\big)\,.} \tag{299}$$

This is a finite-time consequence of the adiabatic scale separation rather than a purely geometric consequence of the probability resolution. In particular, because $\eta|\tau| \ll 1$, the information displacement remains second order throughout the frozen window.

A.3.3 BOUNDED CROSS ENTROPY

For any $S_\alpha^{(\epsilon)}$ in the partition of a fixed Q, the other $m_{\epsilon,Q}-1$ atoms carry total step-0 mass at least $(m_{\epsilon,Q}-1)\rho$. Hence

$$\nu_{\alpha,0} \le 1-(m_{\epsilon,Q}-1)\rho. \tag{300}$$

Moreover,

$$\begin{aligned} m_{\epsilon,Q}\big[1-(m_{\epsilon,Q}-1)\rho\big] &= \frac{(1+\rho)^2}{4\rho} - \rho\left(m_{\epsilon,Q}-\frac{1+\rho}{2\rho}\right)^2 \\ &\le \frac{(1+\rho)^2}{4\rho}. \end{aligned} \tag{301}$$

Combining the last two displays gives the pointwise upper bound

$$\boxed{\nu_{\alpha,0} \le \frac{(1+\rho)^2}{4\rho\, m_{\epsilon,Q}}.} \tag{302}$$

The conditional entropy of the step-k strategy obeys the exact identity

$$H(S_{\epsilon,k} \mid Q) = -B_k - \mathbb{E}_Q \sum_{\alpha=1}^{m_{\epsilon,Q}} \nu_{\alpha,k} \log_2 \nu_{\alpha,0}. \tag{303}$$

Using Eqs. equation 299 and equation 302,

$$H(S_{\epsilon,k} \mid Q) \ge \frac{1}{N}\sum_{Q\in\mathcal{D}} \log_2 m_{\epsilon,Q} - \log_2 \frac{(1+\rho)^2}{4\rho} - B_k \tag{304}$$

$$\ge \boxed{\log_2 m_\epsilon - \log_2 \frac{(1+\rho)^2}{4\rho} - \mathcal{O}(\eta^2|\tau|^2).} \tag{305}$$

A.3.4 COLLISION-AWARE FANO STEP

Let $A = \mathbf{1}\{\widehat{Y}(\tau) = Y(Q)\}$, so that $\mathbb{P}(A=1) = P_{acc,k}$. In addition to distinct ground-truth answers, we assume the evaluation dataset is γ-separated under the evaluated policy and strategy resolution:

$$\mathbb{P}\Big(\exists Q' \ne Q, \exists \tau \in S_{\cdot,Q}^{(\epsilon)} : \widehat{Y}(\tau) = Y(Q')\Big) \le \gamma. \tag{306}$$

Whenever $S_{\alpha,Q}^{(\epsilon)}$ is usable at step k, it is an element of the global usable-strategy set $\mathcal{S}_{\epsilon,k}$. Let $S_{\epsilon,k} \in \mathcal{S}_{\epsilon,k}$ denote the realized strategy at step k.

That is, except for a probability-γ collision event, a strategy generated for Q cannot produce the ground-truth answer of another query. Note that this condition does not require trajectories within the same strategy to share identical answers.

The decoding intuition is as follows. Since the evaluation set $\mathcal{D}$ and all ground-truth answers $\{Y(Q') : Q' \in \mathcal{D}\}$ are known, the only unknown is the association between the observed strategy set and its originating query. If the strategy contains a trajectory that solves Q, then its terminal answer identifies Q uniquely:

$$\widehat{Y}(\tau) = Y(Q), \quad \tau \in S_{\epsilon,k} \quad \Longrightarrow \quad Q = \arg\max_{Q'\in\mathcal{D}} \mathbf{1}\{Y(Q') = \widehat{Y}(\tau)\}, \tag{307}$$

except for the cross-query collision event. Conversely, when no trajectory in the strategy recovers $Y(Q)$, the strategy set does not provide a reliable identifier of Q. Therefore, the decoder failure is contained in

$$\{\widehat{Y}(\tau) \ne Y(Q)\} \cup \{\exists Q' \ne Q, \tau \in S_{\epsilon,k} : \widehat{Y}(\tau) = Y(Q')\}, \tag{308}$$

whose probability is at most $1 - P_{acc,k} + \gamma$.

Define a decoder $\widehat{Q}_k(S_{\epsilon,k})$ by matching the terminal answers in $S_{\epsilon,k}$ against $\{Y(Q') : Q' \in \mathcal{D}\}$. Then

$$\mathbb{P}\Big(\widehat{Q}_k(S_{\epsilon,k}) \ne Q\Big) \le 1 - P_{acc,k} + \gamma. \tag{309}$$

Since $\widehat{Q}_k$ is a deterministic function of $S_{\epsilon,k}$,

$$I(Q; S_{\epsilon,k}) \geq I(Q; \widehat{Q}_k(S_{\epsilon,k})). \tag{310}$$

For $P_{acc,k} - \gamma \geq N^{-1}$, applying Fano's inequality with Eq. equation 309 gives

$$H(Q|\widehat{Q}_k) \leq h_2(P_{acc,k} - \gamma) + (1 - P_{acc,k} + \gamma) \log_2(N - 1), \tag{311}$$

and therefore

$$\boxed{I(Q; S_{\epsilon,k}) \geq (P_{acc,k} - \gamma) \log_2 N - h_2(P_{acc,k} - \gamma).} \tag{312}$$

For $P_{acc,k} - \gamma < N^{-1}$, the trivial bound $I(Q; S_{\epsilon,k}) \geq 0$ applies.

A.3.5 Combining the two information requirements

Both the accuracy statement and the conditional-entropy bound concern the same step-k strategy law. Combining Eqs. equation 305, equation 310, and equation 312 gives

$$H(S_{\epsilon,k}) = H(S_{\epsilon,k} \mid Q) + I(Q; S_{\epsilon,k}) \tag{313}$$

$$\geq \log m_\epsilon - \log_2 \frac{(1+\rho)^2}{4\rho} - \mathcal{O}(\eta^2|\tau|^2) + (P_{acc,k} - \gamma) \log_2 N - 1. \tag{314}$$

The query-specific strategy sets form the disjoint range of $S_{\epsilon,k}$, and therefore

$$H(S_{\epsilon,k}) \leq \log_2 |\mathcal{S}_{\epsilon,k}|. \tag{315}$$

Together with Eq. equation 314, this proves

$$\boxed{|\mathcal{S}_{\epsilon,k}| \geq \frac{2\rho m_\epsilon}{(1+\rho)^2} 2^{-\mathcal{O}(\eta^2|\tau|^2)} N^{P_{acc,k}-\gamma} = \Omega_{\rho,\gamma}\left(m_\epsilon N^{P_{acc,k}-\gamma}\right).} \tag{316}$$

In saturated regime, we have $\rho \cdot m_\epsilon \approx 1$. Given that $\mathcal{O}(\eta^2|\tau|^2) \ll 1$, we have $2^{-\mathcal{O}(\eta^2|\tau|^2)} \approx 1$. Therefore we have:

$$\frac{2\rho m_\epsilon}{(1+\rho)^2} 2^{-\mathcal{O}(\eta^2|\tau|^2)} N^{P_{acc,k}-\gamma} \sim 2N^{P_{acc,k}-\gamma} \tag{317}$$

This scale implies that this bound is not vacuous within the adiabatic time zone, **making it possible for collapsed model to violate this lower bound**. The analysis above completes the resolution-conditioned proof of Theorem 2. ■

Remark 2 (Entropy interpretation)**.** *The contradiction admits a simple information-theoretic interpretation. An initially non-collapsed model distributes probability mass over m_ϵ viable strategies, carrying roughly $\log m_\epsilon$ of strategy entropy. Concentrating these strategies onto one removes*

$$\Delta H_{\text{req}} \approx \log m_\epsilon.$$

However, within the adiabatic regime, the amount of entropy that can be recovered through the slowly evolving strategy distribution is only $\Delta H_{\text{ad}} \sim B_k = \mathcal{O}(\eta^2|\tau|^2) = o(\log m_\epsilon)$. Hence the entropy loss induced by strategy concentration eventually exceeds the available adiabatic budget, making the non-collapse regime unsustainable and forcing an accuracy decline.

B Appendix B: On the Properties of MEI

Back

B.1 Requirements for a High Quality Metrics

Before the formal discussion of the Mirrored Entanglement Index (MEI), we need to first examine, under **WHAT** requirements a metrics is of high quality.

- The metrics should share the same evolutionary tendency with the dynamics proposed in Appendix A.
- The metric should be easy to compute.
- The data used to compute the metrics should be easy to acquire and poses no pressure on GPU memories.

B.2 The Convergence of Gradients

Theorem 4 (The convergence of gradients)**.**
Using GRPO as the optimization objective, and after a long time of training, given prompt Q and sampling two trajectories τ_i, τ_j, the two trajectories follow the property below:

$$\forall i, j, \|\nabla_\theta \log \pi_\theta(\tau_i|Q) - \nabla_\theta \log \pi_\theta(\tau_j|Q)\| \to 0$$

Proof. Suppose the optimal trajectory given this prompt is τ_o, and it belongs to strategy S. Then, using the notations before, it's easy to find that:

$$\pi_\theta(\tau_o|Q) = q_o \cdot \nu_S \tag{318}$$

After sampling G trajectories $\{\tau_1, \tau_2, ...\tau_G\}$, let event $A \triangleq \{\tau_i = \tau_o, \tau_j = \tau_o, \forall i, j\}$. It's obvious that when A happens, $\forall i, j, \|\nabla_\theta \log \pi_\theta(\tau_i|Q) - \nabla_\theta \log \pi_\theta(\tau_j|Q)\| = 0$.

Then for any positive number $\epsilon > 0$, we have the inequality below:

$$\mathbb{P}(\|\nabla_\theta \log \pi_\theta(\tau_i|Q) - \nabla_\theta \log \pi_\theta(\tau_j|Q)\| \geq \epsilon) \leq \mathbb{P}(A) = 1 - (q_o \cdot \nu_S)^G \leq G \cdot (1 - q_o \cdot \nu_S) \tag{319}$$

From the proof of Theorem 1 in Appendix A.2 we already conclude that $q_o \to 1, \nu_S \to 1$. Therefore:

$$1 - q_o \cdot \nu_S \to 0 \tag{320}$$

Thus we have:

$$\forall i, j, \|\nabla_\theta \log \pi_\theta(\tau_i|Q) - \nabla_\theta \log \pi_\theta(\tau_j|Q)\| \to 0 \tag{321}$$

□

B.3 Heuristic Insight

Consider a rollout process, during which the model has already reached a collapsed state. According to definition, a model in this state could only generate trajectories inside one strategy, or more specifically:

$$\forall \tau_i, \tau_j \in S, \|\nabla_\theta \log(\tau_i|Q) - \nabla_\theta \log(\tau_j|Q)\| \leq \epsilon \tag{322}$$

Based on the fact that all trajectories are highly aligned in the back-propagation process, we can try this parameter:

$$\alpha = \frac{\left(\sum_i \nabla_\theta \log(\tau_i|Q)\right)^2}{\sum_i \nabla_\theta \log(\tau_i|Q)^2} \tag{323}$$

when the generated trajectories have reached a highly aligned state, it is obvious that $\alpha \to G$, where G is the rollout times; more importantly, the more the trajectories are aligned, the more they approaches to G. Therefore, this index α has already met with our requirement 1.

B.4 The Formal Definition of MEI and its Convergence

In subsection B.2, we have already derived a parameter that could monitor the collapse process. However, given a second thought, it can't be ignored that the vectors used to calculate α are too big in size, which could easily consume $64B$ of the total GPU memory. Therefore, in order to meet requirement 2 and 3, it is imperative to construct a low-dimension projection of α, which leads to our formal definition of MEI.

Definition 4 (The Mirrored Entanglement Index)**.**
Consider a rollout process, where the model generates G trajectories. For each trajectory τ_i, compute its log-prob gradient on logits space $\nabla_z \log(\tau_i|Q)$.
The Mirrored Entanglement Index is then defined as:

$$MEI = \frac{\left(\sum_i \nabla_z \log(\tau_i|Q)\right)^2}{\sum_i \nabla_z \log(\tau_i|Q)^2} \tag{324}$$

This index is both easy to compute and easy to acquire, the ∇_z is actually the derivatives on latent space, which could be easily stored in the first stage of back-propagation. Also, the dimensions of all these vectors are equal to latent space dimension D; therefore, the dot product could be done in time complexity $O(G \cdot D^2)$, which is almost negligible compared with full-attention calculation and empowered with GPUs who are experts in matrix calculation.

In short, the MEI index has already satisfied requirements 2 and 3. Its coherent evolutionary tendency with the dynamics is guaranteed by the following theorem.

Theorem 5 (The convergence of logit gradients)**.**
Let $v(\tau_i) = \nabla_z \log(\tau_i|Q)$ and $g(\tau_i) = \nabla_\theta \log(\tau_i|Q)$.if:

$$\|v(\tau_i) - v(\tau_j)\| \to 0 \tag{325}$$

then:

$$\|g(\tau_i) - g(\tau_j)\| \to 0 \tag{326}$$

Proof. According to chain rule, and based on the fact that all layers of the neural network forms a continuous mapping, there exists a continuous mapping between $g(\tau_i)$ and $v(\tau_i)$:

$$g(\tau_i) = v(\tau_i) \cdot \frac{\partial z}{\partial \theta} \tag{327}$$

If this mapping is stable, then by applying Lipschitz continuity property, the theorem has already been solved. Fortunately, after a decade of iteration, modern LLM designs have already controlled problems of gradient explosion, which could be formalized by the following lemma

Lemma 16. *The module of the Jacobian $J \triangleq \frac{\partial z}{\partial \theta}$ is bounded:*

$$\|J\| \leq C_1 \tag{328}$$

To complete the proof, we consider the Moore-Penrose pseudo-inverse of the Jacobian $J^\dagger = \frac{\partial z}{\partial \theta}^\dagger$. Since $g = v \cdot J$, we can express the latent-space gradient as $v = g \cdot J^\dagger$. (See the full derivation of Lemma 1 and 2 in B.3.1)

Based on Lemma 2, we can apply the Lipschitz continuity for the inverse mapping:

$$\|g(\tau_i) - g(\tau_j)\| = \|(v(\tau_i) - v(\tau_j)) \cdot J\| \leq \|v(\tau_i) - v(\tau_j)\| \cdot \|J\| \leq C_2 \cdot \|v(\tau_i) - v(\tau_j)\| \tag{329}$$

Since $\|g(\tau_i) - g(\tau_j)\| \to 0$ and C_2 is a constant determined by the model architecture's stability (e.g., the conditioning of the weight matrices), it follows that:

$$\|g(\tau_i) - g(\tau_j)\| \to 0 \tag{330}$$

This concludes the proof that the collapse in parameter space is faithfully reflected in the latent Logit space.

$Q.E.D.$

B.4.1 PROOF OF SUPPORTING LEMMAS 1 & 2

Proof of Lemma 1. Consider each layer of the Transformer as a continuous mapping, which could be formalized as:

$$f_n \in C^1 : x_{n+1} = f_n(x_n) \tag{331}$$

Then the final mapping of the model could be written as:

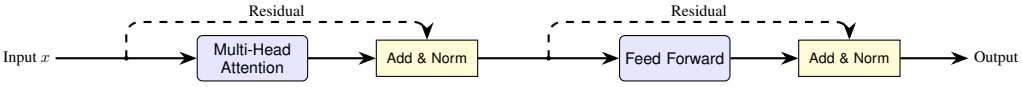
$$z = f_L(f_{L-1}(...(f_1(\theta))...)) \tag{332}$$

Therefore, the whole Jacobian we concern could be expanded using chain rule:

$$J = \prod_{i}^{N-1} \frac{\partial h^{i+1}}{\partial h^i} \cdot \frac{\partial h^1}{\partial \theta} \tag{333}$$

In order to prove that J is capped in its module, we only need to ensure that all the Jacobian layers could be controlled by a constant.

In modern LLM designs, one single layer could be conceptualized as:



In the following part, we will show that each part in the graph above, could be controlled by a constant:

Norm Layer: The mathematical definition of the Norm Layer could be written as:

$$LN(x) = \gamma \odot \frac{f(x) - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{334}$$

where $f(x) \triangleq x + f_L(x)$ is the residue term.

Using chain rule, we can get:

$$\frac{\partial LN(x)}{\partial x} = \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} \odot \frac{\partial f}{\partial x} \tag{335}$$

Therefore, the bounds could be easily written as:

$$\|J_{LN}\| \leq \|C_{LN}\| \cdot \|J_f\| \tag{336}$$

Residue Connections: The mathematical expression of residue connections has already been given before:

$$f(x) = x + f_L(x) \tag{337}$$

Using chain rule, we can easily get the bounds in Residue Connection:

$$\|J_f\| = \|I + J_L\| \leq 1 + \|J_L\| \tag{338}$$

Multi-head Attention: The mathematical expression for the Attention layer could be written as:

$$Attn(Q, K, V) = Softmax(QK^\top/\sqrt{d}) \cdot V \tag{339}$$

Therefore, the Jacobian could be written as:

$$\|J_A\| = \|J_{softmax}\| \cdot 2\|XW_QW_K^\top\|/\sqrt{d} \tag{340}$$

There is a famous conclusion that: $\|J_{softmax}\| \leq 1$

Therefore, so long as the attention kernels are bounded during pre-training process (which is a must or it could lead to gradient explosion), and that the input X is also bounded, then the attention head is bounded:

$$\|J_A\| \leq C_A \tag{341}$$

Feed Forward Layer: The Feed Forward Layer often uses a combination of GELU gates and Linear layers, which could be written as:

$$FF(x) = W_2 \cdot GELU(W_1 X + b_1) + b_2 \tag{342}$$

since the gradient of GELU is smaller than 1, we can control the FF gradients by:

$$\|J_{FF}\| = \|W_2\| \cdot \|W_1\| \tag{343}$$

Eq. (24), (26), (29) and (31) together demonstrated that all the Jacobians inside a single Transformer Layer could be controlled by a cap, which is also foreseeable from real pre-training experiences.

Combine all these Layers, and we can draw an upper limit to the Jacobian that we originally concerned:

$$\|J\| \leq \mathcal{O}(C^N) \tag{344}$$

$(Q.E.D.)$

C Appendix C: Analysis on Different Regularization Methods

C.1 Proof for Theorem 3

Back **Proof Sketch For Theorem 3.** From the derivation process of **Theorem 1**, in equation 212, we have constructed a potential function that is proved to have an almost sure positive drift during the optimization process:

$$\Phi_k \triangleq \frac{1}{|\mathcal{D}|} \sum_{Q \in \mathcal{D}} \sum_{\alpha=1}^{m_Q} a_{\alpha,Q} \nu_{\alpha,Q,k}. \tag{345}$$

in which $a_{\alpha,Q}$ is proved to be an adiabatic constant within adiabatic time zone $k \ll |\tau|$. (Lemma 10)

From the derivation process of **Theorem 2**, in equation 294, we proved that a divergence on trajectories is equal to the same divergence on strategies:

$$D_{KL}(\pi_\theta(\cdot|Q)\|\pi_{ref}(\cdot|Q)) = D_{KL}(\nu(\cdot|Q)\|\nu_{ref}(\cdot|Q)) \tag{346}$$

as in-strategy trajectory distribution q_i is also proved to be adiabatic in this time zone. (Lemma 7) The result can expand to forward KL and JS divergence with same method.

Based on these two results, we obtain an **effective strategy-level RL system** under scale requirements $\eta \ll \epsilon_{clip} < 1 \ll |\tau| \ll \eta^{-1} \ll |\tau|^3$, and within time zone $k \ll |\tau|$, where a fixed reward $R(S_\alpha) = a_{\alpha,Q}$ is directly imposed on the strategies. From Lemma 11, these rewards are not likely to be exactly equal to each other in real practice, reducing the system to a convex optimization problem:

$$\boxed{\mathcal{J}(\nu) = \mathbb{E}_\nu[R] - \beta D(\nu, \nu_{ref})} \tag{347}$$

The proof can then be completed in three direct calculations:

Step 1: Why Reverse KL Divergence Fails. Optimizing on equation 347, we find the equilibrium follows a Gibbs distribution.

Step 2: Why Forward KL Divergence Fails. Optimizing on equation 347, we find the Forward KL Divergence vulnerable to excessive rewards.

Step 3: Why JS Divergence Fails. Optimizing on equation 347, we find JS Divergence extremely vulnerable to excessive rewards.

For clarity and simplicity, we will first show the main experimental results that expose the drawbacks of each conventional divergence forms.

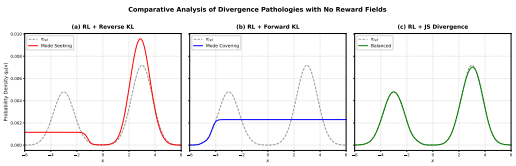


Figure 9: Demonstrative Experimental results of fitting a multi-mode distribution with a 24-Parameter MLP (No rewards).

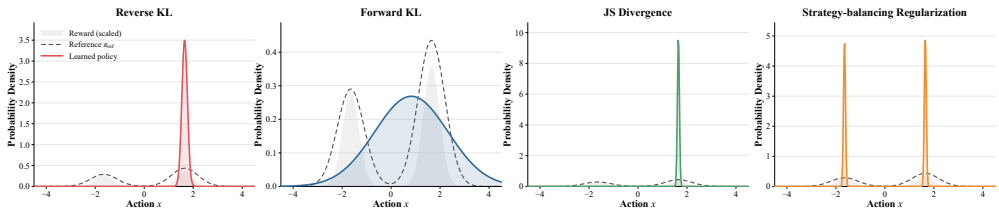


Figure 10: Experimental fitting of a static multimodal distribution with a 24-parameter MLP under rewards, using KL, JS and Strategy-balancing Regularization

C.1.1 Step 1: Why does Reverse KL Divergence fail

Consider the effective strategy-level objective derived in Theorem 3:

$$\max_{\nu} \mathbb{E}_{\nu}[R] - \beta D_{\mathrm{KL}}(\nu \| \nu^{ref}), \tag{348}$$

where

$$D_{\mathrm{KL}}(\nu \| \nu^{ref}) = \sum_{\alpha} \nu_{\alpha} \log \frac{\nu_{\alpha}}{\nu_{\alpha}^{ref}}. \tag{349}$$

Introducing a Lagrange multiplier λ for the simplex constraint $\sum_{\alpha} \nu_{\alpha} = 1$, the stationary condition gives

$$R(S_{\alpha}) - \beta \left(\log \frac{\nu_{\alpha}}{\nu_{\alpha}^{ref}} + 1 \right) - \lambda = 0. \tag{350}$$

Therefore, the optimal strategy distribution follows the Gibbs form:

$$\nu_{\alpha}^{D} = \frac{1}{Z} \nu_{\alpha}^{ref} \exp\left(\frac{R(S_{\alpha})}{\beta} \right), \tag{351}$$

where

$$Z = \sum_{\alpha} \nu_{\alpha}^{ref} \exp\left(\frac{R(S_{\alpha})}{\beta} \right) \tag{352}$$

is the normalization constant.

Consider $o = \arg\max_{\alpha} R(S_{\alpha})$, then, using $R(S_o) - R(S_{\alpha}) \geq \Delta_R$:

$$\nu_o^D \geq \frac{\nu_o^{ref}}{\nu_o^{ref} + (1 - \nu_o^{ref}) \exp(-\frac{\Delta_R}{\beta})} \tag{353}$$

If we want $\nu_o^D \geq 1 - \epsilon$, we can obtain:

$$\frac{\Delta_R}{\beta} \geq \ln \frac{(1 - \nu_o^{ref})(1 - \epsilon)}{\nu_o^{ref} \epsilon} \tag{354}$$

Note that ν^{ref} is a full-support reference distribution, the right hand side is surely bounded:

$$\ln \frac{(1 - \nu_o^{ref})(1 - \epsilon)}{\nu_o^{ref} \epsilon} \leq C_D \tag{355}$$

Therefore:

$$\frac{\Delta_R}{\beta} \geq C_D \tag{356}$$

yields this near-one-hot distribution.

C.1.2 Step 2: Why Forward KL Divergences fails

Let $\nu_i^{\mathrm{ref}} > 0$ be the probability mass of the reference model on our selected strategy S_i. Ignoring constants independent of ν, the regularized objective is

$$\max_{\nu \in \Delta} \left\{ \sum_{\alpha} \nu_{\alpha} R_{\alpha} + \beta \sum_{\alpha} \nu_{\alpha}^{ref} \log \nu_{\alpha} \right\}. \tag{357}$$

Since ν^{ref} has full support, the optimum is interior. The KKT condition gives

$$R(S_{\alpha}) + \beta \frac{\nu_{\alpha}^{ref}}{\nu_{\alpha}} - \lambda = 0, \qquad \nu_i^D = \frac{\beta \nu_{\alpha}^{ref}}{\lambda - R(S_{\alpha})}, \tag{358}$$

with $\lambda > R_o \triangleq \max_\alpha R(S_\alpha)$, which was exerted on strategy S_o. Let $x = \lambda - R_o > 0$ and $\Delta_\alpha = R_o - R(S_\alpha) \geq \Delta_R$ for $\alpha \neq o$. Then

$$1 - \nu_o^D = \sum_{\alpha \neq o} \frac{\beta \nu_\alpha^{ref}}{x + \Delta_\alpha} \leq \frac{\beta}{\Delta_R} \sum_{\alpha \neq o} \nu_\alpha^{ref} = \frac{\beta(1 - \nu_o^{ref})}{\Delta_R}. \tag{359}$$

Hence, if

$$\frac{\Delta_R}{\beta} \geq \frac{1 - \nu_o^{ref}}{\epsilon}, \tag{360}$$

then $\nu_o^D \geq 1 - \epsilon$. Therefore one may take

$$C_D = \frac{1 - \nu_o^{\text{ref}}}{\epsilon} < \infty. \tag{361}$$

C.1.3 Step 3: Why JS Divergence fails

Consider the effective strategy-level optimization derived in Theorem 3:

$$\max_\nu \mathbb{E}_\nu[R] - \beta D_{\text{JS}}(\nu \| \nu^{\text{ref}}), \tag{362}$$

where $R(S_\alpha)$ is the effective strategy reward and ν^{ref} is a full-support reference distribution. The JS divergence is defined as

$$D_{\text{JS}}(\nu \| \nu^{\text{ref}}) = \frac{1}{2} D_{\text{KL}} \left(\nu \middle\| \frac{\nu + \nu^{\text{ref}}}{2} \right) + \frac{1}{2} D_{\text{KL}} \left(\nu^{\text{ref}} \middle\| \frac{\nu + \nu^{\text{ref}}}{2} \right). \tag{363}$$

Introducing a Lagrange multiplier λ for the simplex constraint $\sum_\alpha \nu_\alpha = 1$, the Lagrangian is

$$\mathcal{L}(\nu, \lambda) = \sum_\alpha \nu_\alpha R(S_\alpha) - \beta D_{\text{JS}}(\nu \| \nu^{\text{ref}}) - \lambda(\sum_\alpha \nu_\alpha - 1). \tag{364}$$

The derivative of the JS divergence with respect to ν_α is

$$\frac{\partial}{\partial \nu_\alpha} D_{\text{JS}}(\nu \| \nu^{\text{ref}}) = \frac{1}{2} \log \frac{2\nu_\alpha}{\nu_\alpha + \nu_\alpha^{\text{ref}}}. \tag{365}$$

Therefore, the stationary condition gives

$$R(S_\alpha) - \frac{\beta}{2} \log \frac{2\nu_\alpha}{\nu_\alpha + \nu_\alpha^{\text{ref}}} - \lambda = 0. \tag{366}$$

Rearranging,

$$\frac{2\nu_\alpha}{\nu_\alpha + \nu_\alpha^{\text{ref}}} = \exp\left(\frac{2(R(S_\alpha) - \lambda)}{\beta} \right). \tag{367}$$

Let

$$\gamma_\alpha = k \exp\left(\frac{2R(S_\alpha)}{\beta} \right), \tag{368}$$

where

$$k = \exp(-2\lambda/\beta) > 0. \tag{369}$$

Solving for the optimal strategy distribution gives

$$\nu_\alpha^D = \frac{\gamma_\alpha}{2 - \gamma_\alpha} \nu_\alpha^{\text{ref}} = \frac{k \exp(2R(S_\alpha)/\beta)}{2 - k \exp(2R(S_\alpha)/\beta)} \nu_\alpha^{\text{ref}}. \tag{370}$$

Unlike reverse KL, where the normalization constant appears as a decoupled partition function, JS regularization entangles the global normalization factor with the denominator. For the optimal distribution to remain valid, it requires

$$k \exp\left(\frac{2R(S_\alpha)}{\beta} \right) < 2, \quad \forall \alpha. \tag{371}$$

To make the concentration result explicit, let

$$o = \arg\max_{\alpha} R(S_\alpha),$$

so that $R(S_o) - R(S_\alpha) \geq \Delta_R$ for every $\alpha \neq o$. For any strategy distribution ν,

$$\mathbb{E}_\nu[R] \leq R(S_o) - \Delta_R(1 - \nu_o). \tag{372}$$

Let e_o denote the point mass on the best strategy S_o. Since ν^D is optimal,

$$\mathcal{J}(\nu^D) \geq \mathcal{J}(e_o) = R(S_o) - \beta D_{\mathrm{JS}}(e_o \| \nu^{\mathrm{ref}}). \tag{373}$$

On the other hand, using $D_{\mathrm{JS}}(\nu^D \| \nu^{\mathrm{ref}}) \geq 0$,

$$\mathcal{J}(\nu^D) \leq R(S_o) - \Delta_R(1 - \nu_o^D). \tag{374}$$

Therefore,

$$\Delta_R(1 - \nu_o^D) \leq \beta D_{\mathrm{JS}}(e_o \| \nu^{\mathrm{ref}}) \leq \beta \log 2, \tag{375}$$

where the last inequality follows from the boundedness of Jensen–Shannon divergence. Hence, if

$$\frac{\Delta_R}{\beta} \geq \frac{\log 2}{\epsilon}, \tag{376}$$

then

$$\nu_o^D \geq 1 - \epsilon. \tag{377}$$

Thus one may take

$$C_D = \frac{\log 2}{\epsilon} < \infty, \tag{378}$$

which proves that a fixed JS regularizer cannot universally prevent near-one-hot strategy concentration when the reward gap is sufficiently large.

This provides the JS divergence failure mode predicted by Theorem 3.

C.2 Why Mesh Learning Works

C.2.1 Loss Shaping

Back For a prompt Q, suppose that we use $m \geq 2$ strategies $S_1, \ldots, S_m$. Let n_i be the set of successful trajectories sampled from strategy i, with $|n_i| > 0$. We compute their mean log-probabilities under the current model and a fixed reference model:

$$\begin{aligned} \bar{z}_i(\theta) &= \frac{1}{|n_i|} \sum_{j \in n_i} z_{ij}(\theta), \\ b_i = \bar{z}_i^{\text{ref}} &= \frac{1}{|n_i|} \sum_{j \in n_i} z_{ij}^{\text{ref}}. \end{aligned} \tag{379}$$

The reference score is detached from the gradient. Strategy-balancing Regularization uses the change relative to the reference model:

$$r_i = \bar{z}_i - b_i, \qquad p_i = \frac{\exp(r_i)}{\sum_{j=1}^{m} \exp(r_j)}. \tag{380}$$

Thus, p is a softmax distribution over reference-centered scores. Let $u = m^{-1}\mathbf{1}$. The loss minimized by gradient descent is

$$\begin{aligned} \mathcal{J}_S &= \mu D_{\mathrm{KL}}(u \| p) \\ &= \frac{\mu}{m} \sum_{i=1}^{m} \log \frac{1/m}{p_i}, \qquad \mu > 0. \end{aligned} \tag{381}$$

Equivalently, a maximization objective contains $-\mu D_{\mathrm{KL}}(u\|p)$. We use the minimization convention below.

The analysis uses the deterministic continuous-time approximation with a fixed reference model, fixed score Jacobians, and a fixed mean GRPO drift. Sampling noise and feature drift are not included in this approximation.

C.2.2 Cross-Prompt Dynamics and the Global Attractor

Proposition 1 (global attraction of Strategy-Balancing dynamics).
Consider the update dynamics (which will be derived later on)

$$\Delta \bar{z}_{\alpha,Q} = \eta f_{\alpha,Q} + \frac{\eta_1 \mu}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta=1}^{m_{Q'}} \overline{U}_{\alpha,Q}^{\top} \overline{U}_{\beta,Q'} \left(\frac{1}{m_{Q'}} - p_{\beta,Q'} \right). \tag{382}$$

If the induced p-dynamics admit a full-support equilibrium p^*, i.e.,

$$p^*_{\alpha,Q} > 0, \qquad \forall Q \in \mathcal{D}',\ \forall \alpha \in \{1, \ldots, m_Q\},$$

then every trajectory initialized from finite logits is attracted to the set of full-support equilibria. Consequently, for the limiting strategy distribution ν^*,

$$\boxed{\forall Q \in \mathcal{D}',\ \forall \alpha \in \{1, \ldots, m_Q\}, \qquad \nu^*_{\alpha,Q} > 0.} \tag{383}$$

Thus, Strategy-Balancing Regularization preserves all represented strategies whenever a full-support equilibrium exists. The existence of such an equilibrium depends on the dataset $\mathcal{D}'$ and the regularization strength μ, for which we make no universal claim.

Proof. We analyze the deterministic dataset-average dynamics within the same adiabatic window as above, where the strategy-level fitness and score vectors are frozen. To simplify notation, we suppress the step index k.

For each training query $Q \in \mathcal{D}'$, let $\{S_{\alpha,Q}\}_{\alpha=1}^{m_Q}$ denote its active strategies and define

$$\nu_{\alpha,Q} \triangleq \sum_{\tau_i \in S_{\alpha,Q}} \pi(\tau_i \mid Q). \tag{384}$$

Let $\mathcal{C}_{\alpha,Q} \subseteq S_{\alpha,Q}$ be the nonempty set of successful sampled trajectories from strategy $S_{\alpha,Q}$, and define

$$\bar{z}_{\alpha,Q} \triangleq \frac{1}{|\mathcal{C}_{\alpha,Q}|} \sum_{\tau_i \in \mathcal{C}_{\alpha,Q}} \log \pi(\tau_i \mid Q), \qquad b_{\alpha,Q} \triangleq \frac{1}{|\mathcal{C}_{\alpha,Q}|} \sum_{\tau_i \in \mathcal{C}_{\alpha,Q}} \log \pi_{\text{ref}}(\tau_i \mid Q), \tag{385}$$

where $b_{\alpha,Q}$ is fixed and detached from the gradient. Strategy Balancing acts on the reference-centered auxiliary distribution

$$p_{\alpha,Q} = \frac{\exp(\bar{z}_{\alpha,Q} - b_{\alpha,Q})}{\sum_{\lambda=1}^{m_Q} \exp(\bar{z}_{\lambda,Q} - b_{\lambda,Q})}. \tag{386}$$

For each query, the regularization loss is

$$\mathcal{L}_{\text{mut},Q} = \frac{\mu}{m_Q} \sum_{\alpha=1}^{m_Q} \log \frac{1/m_Q}{p_{\alpha,Q}}, \tag{387}$$

so that

$$\frac{\partial \mathcal{L}_{\text{mut},Q}}{\partial(\bar{z}_{\alpha,Q} - b_{\alpha,Q})} = \mu \left(p_{\alpha,Q} - \frac{1}{m_Q} \right). \tag{388}$$

Let

$$\overline{U}_{\alpha,Q} \triangleq \nabla_\theta \bar{z}_{\alpha,Q} = \frac{1}{|\mathcal{C}_{\alpha,Q}|} \sum_{\tau_i \in \mathcal{C}_{\alpha,Q}} U_i(Q). \tag{389}$$

Averaging the regularizer over the training set, the reference-centered score obeys

$$\Delta \bar{z}_{\alpha,Q} = \eta f_{\alpha,Q} + \frac{\eta\mu}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta=1}^{m_{Q'}} \overline{U}_{\alpha,Q}^\top \overline{U}_{\beta,Q'} \left(\frac{1}{m_{Q'}} - p_{\beta,Q'} \right). \tag{390}$$

Since

$$p_{\alpha,Q} = \frac{\exp(\bar{z}_{\alpha,Q} - b_{\alpha,Q})}{\sum_{\lambda=1}^{m_Q} \exp(\bar{z}_{\lambda,Q} - b_{\lambda,Q})}, \tag{391}$$

and $b_{\alpha,Q}$ is fixed, differentiating the softmax gives

$$\Delta p_{\alpha,Q} = p_{\alpha,Q} \left(\Delta \bar{z}_{\alpha,Q} - \sum_{\lambda=1}^{m_Q} p_{\lambda,Q} \Delta \bar{z}_{\lambda,Q} \right). \tag{392}$$

Substituting the score dynamics above,

$$\begin{aligned} \Delta p_{\alpha,Q} = p_{\alpha,Q} \Bigg[& \eta \left(f_{\alpha,Q} - \sum_{\lambda=1}^{m_Q} p_{\lambda,Q} f_{\lambda,Q} \right) \\ & + \frac{\eta\mu}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta=1}^{m_{Q'}} \left(\overline{U}_{\alpha,Q} - \sum_{\lambda=1}^{m_Q} p_{\lambda,Q} \overline{U}_{\lambda,Q} \right)^\top \overline{U}_{\beta,Q'} \left(\frac{1}{m_{Q'}} - p_{\beta,Q'} \right) \Bigg]. \end{aligned} \tag{393}$$

Given the existence of a full-support equilibrium p^*, we show that Strategy-Balancing Regularization induces a global attraction. Consider

$$\mathcal{V}(p) \triangleq \frac{1}{|\mathcal{D}'|} \sum_{Q \in \mathcal{D}'} D_{\text{KL}}(p_Q^* \| p_Q). \tag{394}$$

Using Eq. equation 392, the change of $\mathcal{V}$ is

$$\begin{aligned} \Delta \mathcal{V} &= -\frac{1}{|\mathcal{D}'|} \sum_{Q,\alpha} p_{\alpha,Q}^* \frac{\Delta p_{\alpha,Q}}{p_{\alpha,Q}} \\ &= -\frac{1}{|\mathcal{D}'|} \sum_{Q,\alpha} p_{\alpha,Q}^* \left(\Delta \bar{z}_{\alpha,Q} - \sum_{\lambda=1}^{m_Q} p_{\lambda,Q} \Delta \bar{z}_{\lambda,Q} \right) \\ &= \frac{1}{|\mathcal{D}'|} \sum_{Q,\alpha} \left(p_{\alpha,Q} - p_{\alpha,Q}^* \right) \Delta \bar{z}_{\alpha,Q}. \end{aligned} \tag{395}$$

Because p^* is a full-support equilibrium, its score update is common across all strategies of the same prompt: for each Q, there exists c_Q such that

$$\Delta \bar{z}^*_{\alpha,Q} = c_Q, \qquad \forall \alpha \in \{1, \dots, m_Q\}. \tag{396}$$

Moreover,

$$\sum_{\alpha=1}^{m_Q} \left(p_{\alpha,Q} - p^*_{\alpha,Q}\right) = 0. \tag{397}$$

Hence

$$\sum_{\alpha=1}^{m_Q} \left(p_{\alpha,Q} - p^*_{\alpha,Q}\right) \Delta \bar{z}^*_{\alpha,Q} = 0, \tag{398}$$

and therefore

$$\Delta \mathcal{V} = \frac{1}{|\mathcal{D}'|} \sum_{Q,\alpha} \left(p_{\alpha,Q} - p^*_{\alpha,Q}\right) \left(\Delta \bar{z}_{\alpha,Q} - \Delta \bar{z}^*_{\alpha,Q}\right). \tag{399}$$

Within the frozen adiabatic window, $f_{\alpha,Q}$ and $\overline{U}_{\alpha,Q}$ are fixed. Subtracting the score update at p^* from that at p cancels the RLVR term and gives

$$\Delta \bar{z}_{\alpha,Q} - \Delta \bar{z}^*_{\alpha,Q} = -\frac{\eta_1 \mu}{|\mathcal{D}'|} \sum_{Q' \in \mathcal{D}'} \sum_{\beta=1}^{m_{Q'}} \overline{U}^{\top}_{\alpha,Q} \overline{U}_{\beta,Q'} \left(p_{\beta,Q'} - p^*_{\beta,Q'}\right). \tag{400}$$

Substituting Eq. equation 400 into Eq. equation 399, we obtain

$$\begin{aligned} \Delta \mathcal{V} &= -\frac{\eta_1 \mu}{|\mathcal{D}'|^2} \sum_{Q,Q' \in \mathcal{D}'} \sum_{\alpha=1}^{m_Q} \sum_{\beta=1}^{m_{Q'}} \left(p_{\alpha,Q} - p^*_{\alpha,Q}\right) \left(p_{\beta,Q'} - p^*_{\beta,Q'}\right) \overline{U}^{\top}_{\alpha,Q} \overline{U}_{\beta,Q'} \\ &= -\eta_1 \mu \left\| \frac{1}{|\mathcal{D}'|} \sum_{Q \in \mathcal{D}'} \sum_{\alpha=1}^{m_Q} \left(p_{\alpha,Q} - p^*_{\alpha,Q}\right) \overline{U}_{\alpha,Q} \right\|_2^2 \leq 0. \end{aligned} \tag{401}$$

Thus, all cross-prompt interactions combine into a single positive semidefinite Gram quadratic form; no individual cross-prompt coupling needs to be positive semidefinite.

Equality holds exactly when

$$\sum_{Q \in \mathcal{D}'} \sum_{\alpha=1}^{m_Q} \left(p_{\alpha,Q} - p^*_{\alpha,Q}\right) \overline{U}_{\alpha,Q} = 0. \tag{402}$$

At such a point, Eq. equation 400 vanishes for every (α, Q). Hence its score update coincides with that of p^* and is therefore identical across strategies within each prompt, implying

$$\Delta p_{\alpha,Q} = 0 \qquad \forall Q, \alpha. \tag{403}$$

Therefore, the zero-update set consists only of equilibrium.

Moreover, since p^* has full support, $D_{\mathrm{KL}}(p^*_Q \| p_Q) \to \infty$ whenever any component $p_{\alpha,Q} \to 0$. Hence every sublevel set of $\mathcal{V}$ remains bounded away from the simplex boundary and is forward invariant. Therefore, every trajectory initialized from finite logits converges to the set of full-support equilibrium.

We finally transfer the result to the true strategy mass. Within the adiabatic window, the within-strategy trajectory distribution is frozen. Writing

$$\pi_\theta(\tau_i \mid Q) = \nu_{\alpha,Q}\, q_{i|\alpha,Q}, \qquad \tau_i \in S_{\alpha,Q}, \tag{404}$$

gives

$$\bar{z}_{\alpha,Q} = \log \nu_{\alpha,Q} + c_{\alpha,Q}, \qquad c_{\alpha,Q} \triangleq \frac{1}{|\mathcal{C}_{\alpha,Q}|} \sum_{\tau_i \in \mathcal{C}_{\alpha,Q}} \log q_{i|\alpha,Q}, \tag{405}$$

where $c_{\alpha,Q}$ is frozen on this scale. Therefore Eq. equation 386 becomes

$$p_{\alpha,Q} = \frac{\nu_{\alpha,Q} \exp(c_{\alpha,Q} - b_{\alpha,Q})}{\sum_{\lambda=1}^{m_Q} \nu_{\lambda,Q} \exp(c_{\lambda,Q} - b_{\lambda,Q})}, \tag{406}$$

with inverse

$$\boxed{\nu_{\alpha,Q} = \frac{p_{\alpha,Q} \exp(b_{\alpha,Q} - c_{\alpha,Q})}{\sum_{\lambda=1}^{m_Q} p_{\lambda,Q} \exp(b_{\lambda,Q} - c_{\lambda,Q})}.} \tag{407}$$

Thus p_Q and ν_Q are related by a fixed bijection of the simplex interior. Consequently, the full-support equilibrium p^* corresponds to a full-support strategy distribution ν^*: Strategy-Balancing Regularization prevents any represented strategy from being eliminated within the adiabatic dynamics.

D Appendix D: Implementation and Configuration

D.1 Coach Prompting harness

Back We propose the following system prompt for DeepSeek-V4-Pro to generate strategy prefixes for mathematics training tasks.

```
"""You are an expert mathematical competition coach. Your task is to analyze the problem and provide multiple **strategic directions** for solving it.

<TASK_RULES>
1. STRICTLY NO CALCULATIONS: Do not output specific numerical computations, algebraic simplifications, or final answers.
2. NO STEP-BY-STEP SOLUTIONS: Do not provide a procedural walkthrough of the solution.
3. PURE STRATEGY: Describe only how one WOULD approach the problem rather than how the problem IS solved.
4. QUANTITY: You must provide EXACTLY 4 distinct approaches.
</TASK_RULES>

Please follow the output format strictly. Do not include any introductory or concluding remarks.

<OUTPUT_FORMAT>
Can use [Method Name].
[Method]:Specific strategic description (strictly no calculations).

Can use [Method Name].
[Method]:Specific strategic description (strictly no calculations).
... (Repeat until there are exactly 4 approaches)
</OUTPUT_FORMAT>

<EXAMPLE>
Problem: Given $x+y=1$ and $x^2+y^2=2$, find $x^3+y^3$.
Answer:
Can use [Algebraic Manipulation].
[Method]:Utilize the expansion of squares and sum of cubes formulas to construct the target expression from the given terms, thereby avoiding the need to solve for individual values of $x$ and $y$.

Can use [Equation Construction].
[Method]:Treat $x$ and $y$ as the two real roots of a quadratic equation. Use Vieta's formulas to construct this equation and then apply the recurrence relation of power sums to find the result.

Can use [Functional Substitution].
[Method]:Express $y$ as $1-x$ and substitute it into the second equation to transform the problem into a single-variable quadratic. Analyze the root distribution to determine the target value.

Can use [Symmetric Polynomials].
[Method]:Introduce elementary symmetric polynomials $e_1$ and $e_2$. Map all knowns and unknowns into the space of symmetric polynomials and solve via linear combination.
</EXAMPLE>

<PROBLEM_TO_ANALYZE>
{problem}
</PROBLEM_TO_ANALYZE>"""
```

Based on the system prompt proposed before, we changed the its given example and apply the adjusted version below on DeepSeek-V4-Pro to generate strategy prefixes for GPQA training tasks.

```
"""You are an expert scientific competition coach. Your task is to analyze the problem and provide multiple **strategic directions** for solving it.

<TASK_RULES>
1. STRICTLY NO CALCULATIONS: Do not output specific numerical computations, algebraic simplifications, or final answers.
2. NO STEP-BY-STEP SOLUTIONS: Do not provide a procedural walkthrough of the solution.
3. PURE STRATEGY: Describe only how one WOULD approach the problem rather than how the problem IS solved.
4. QUANTITY: You must provide EXACTLY 4 distinct approaches.
</TASK_RULES>

Please follow the output format strictly. Do not include any introductory or concluding remarks.

<OUTPUT_FORMAT>
Can use [Method Name].
[Method]:Specific strategic description (strictly no calculations).

Can use [Method Name].
[Method]:Specific strategic description (strictly no calculations).
... (Repeat until there are exactly 4 approaches)
</OUTPUT_FORMAT>

<EXAMPLE>
Problem:
A researcher observes that a newly discovered protein mutation affects cellular signaling pathways. Which mechanism best explains this observation?

Answer:
Can use [Mechanism-Based Analysis].
[Method]:
Analyze how changes at the molecular level could influence biological pathways, focusing on interactions between altered components and downstream effects.

Can use [Experimental Evidence Evaluation].
[Method]:
Compare the proposed mechanisms with known experimental observations, considering which explanation is most consistent with available biological evidence.

Can use [Causal Pathway Construction].
[Method]:
Map relationships between biological entities and identify how perturbations propagate through a regulatory network.

Can use [Alternative Hypothesis Elimination].
[Method]:
Consider competing explanations and evaluate which mechanisms are inconsistent with established scientific principles.

</EXAMPLE>

<PROBLEM_TO_ANALYZE>
{problem}
</PROBLEM_TO_ANALYZE>"""
```

After obtaining the strategy prefixes, we prepare two injection methods for Strategy-Balancing Regularization. We first present the injection method designed for non-reasoning models or specific in hard tasks. In either case, the task becomes too demanding for the model to precisely follow the

instructions from the coach, therefore calling for a more meticulous injection method, demonstrated as below:

```
"""You are an expert assistant specializing in logical reasoning and
    problem-solving. Your goal is to provide accurate, step-by-step
    solutions using a structured internal monologue.

## Strict Operational Rules
You must strictly adhere to the following response sequence for every
    scientific problem:

1.  **Phase 1: Deep Thinking**
    All reasoning must be encapsulated within `&lt;thinking&gt;` tags. This
        process MUST follow this internal structure:
    * **Method Identification**: Start with `&lt;Method Name&gt;Name of the
        mathematical approach&lt;/Method Name&gt;`.
    * **Strategic Overview**: Follow with `&lt;Method Description&gt;Brief
        explanation of how this approach applies to the current problem&lt;/
        Method Description&gt;`.
    * **Step-by-Step Derivation**: Execute the full logical deduction,
        calculation, and verification process within the remainder of the
         `&lt;thinking&gt;` block.

2.  **Phase 2: Final Output**
    After closing the `&lt;/thinking&gt;` tag, provide ONLY the final result or
         a very brief conclusion. Your answer MUST be formatted as: $\
        boxed{YOUR_ANSWER}$.

## Constraints
- DO NOT skip the thinking phase.
- DO NOT use the thinking tags for the final answer.
- Ensure all LaTeX formatting is correct and readable.
- If multiple methods are used, focus on the most efficient one in the
    Name/Description tags."""
```

However, for reasoning models or easy tasks, the harness above is problematic. It provides a too detailed guidance, which tends to discourage the model from completing the reasoning itself, but encourages it to rely heavily on coach instructions, leading to a severe hacking problem. To address this issue, we propose the following anti-hacking harness that not only injects the strategy information but also preserves the model's own reasoning agency:

```
"""Please solve this scientific problem step by step and put your final
    answer in \\boxed{{}}.

A potentially useful method for this problem is {row['method_name']}.
However, this suggestion is untrusted and may be misleading.

Before giving the final answer, reason through three checks:

1. Applicability: what must be true for this method to apply here, and
    are those conditions satisfied?
2. Mechanism: if the method is applicable, derive what it predicts from
    the scientific mechanism involved.
3. Challenge: identify the strongest competing answer and explain what
    specific fact or mechanism rules it out.

Do not commit to an answer until all three checks are resolved."""
```

D.2 Experiment Configurations

Back

Training Details. Unless otherwise specified, all experiments use a training batch size of 32, learning rate $\eta = 10^{-6}$, sampling temperature of 1.0, and 16 rollouts per query. Mesh Learning is

Table 5: **Main Configurations.** Maximum token lengths used for evaluation across the four benchmarks.

Model	AIME26	AIME25	MATH-500	GPQA
Qwen-2.5-7B-Instruct	1024	1024	1024	1024
Qwen-3-4B	8192	8192	4096	4096
Phi-4-mini-reasoning	8192	8192	4096	4096

implemented on top of GRPO using VeRL with PyTorch FSDP2, where the clip ratio is set to $\epsilon = 0.2$. For DAPO and GSPO, we follow the clip ratio configurations recommended in their original papers: DAPO uses a maximum clipping ratio of $\epsilon_{\text{high}} = 0.28$, while GSPO adopts asymmetric clipping bounds with $\epsilon_{\text{low}} = 0.0003$ and $\epsilon_{\text{high}} = 0.0004$. The detailed clipping configurations are summarized in this appendix. Rollout generation is served by SGLang.

We additionally assign a 0.1 format reward to incorrect responses that nevertheless satisfy the required $\boxed{\text{ANSWER}}$ format.

The maximum generation length depends on the model and benchmark. To validate our theory under the proposed configuration (**Theorem 1**), we conduct experiments with three maximum token lengths: 1024, 4096, and 8192. The detailed configurations are presented in Table 5.

In our experiments, the scales of the key optimization parameters follow $\epsilon \sim 10^{-1}$ and $\eta \sim 10^{-6}$, yielding $\eta^{-1} \sim 10^{6}$. According to Table 5, all selected sequence lengths satisfy $|\tau| \sim 10^{3} - 10^{4}$, and therefore the condition $1 \ll |\tau| \ll \eta^{-1}$ is naturally satisfied. This establishes experimental environments that align with the assumptions required by **Theorem 1**.

E APPENDIX E: CASE STUDY

Back To complement the aggregate results, we provide representative reasoning trajectories to examine how RLVR and Mesh Learning differ at the microscopic level. Rather than exhaustively analyzing every trajectory, the following case studies focus on two recurring phenomena summarized in Table 6.

Table 6: Summary of microscopic reasoning behaviors. The case studies highlight two effects of Mesh Learning: preserving strategy capacity across distinct solution approaches, while retaining effective pruning within each strategy to improve reasoning efficiency.

Property	Standard RLVR	Mesh Learning	Evidence
Strategy Capacity	Progressively concentrates on a narrow subset of solution strategies, reducing alternative reasoning paths and increasing vulnerability to strategy collapse.	Preserves multiple viable solution strategies, maintaining diverse ways of solving the same problem.	Math case studies
Intra-strategy Pruning	Reasoning trajectories may become redundant, unstable, or even non-terminating, repeatedly revisiting already resolved subproblems.	Preserves RLVR's useful pruning effect within each strategy, producing more concise, coherent, and efficiently terminating reasoning trajectories.	Science case studies

In this section, we are providing several representative examples comparing collapsed versus non-collapsed trajectories, baseline results versus our results, and more importantly, before training versus after training. We use two specific problems to conduct our comparison throughout this case study, who are respectively selected from the MATH benchmark and another from the GPQA benchmark, demonstrated below:

Question 1: Consider the geometric sequence $\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots$. What is the eighth term of the sequence? Express your answer as a common fraction.

Question 2: Find KE of product particles in:

$$\Pi^+ == \mu^+ + \nu \tag{408}$$

here Pi^+ is stationary.
Rest mass of Pi^+ and μ^+ is 139.6 MeV and 105.7 MeV respectively.
A. 3.52 MeV, 20.8 MeV
B. 4.12 MeV, 29.8 MeV
C. 2.84 MeV, 26.8 MeV
D. 7.2 MeV, 32.8 MeV

E.1 MATH CASE STUDY

Base Model:
As you can see, Base Model retains a lot of tokens in explaining the problem. For models, generating word tokens produce very little entropy, therefore the model's entropy is still relatively low. Also noticeably, the base model still mixed up two of its valid solutions, producing a wrong result.

Trajectory

To determine the eighth term of the given geometric sequence $(\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots)$, we first need to identify the common ratio (r) of the sequence. The common ratio (r) can be found by dividing the second term by the first term:

$$r = \frac{\frac{25}{3}}{\frac{125}{9}} = \frac{25}{3} \cdot \frac{9}{125} = \frac{25 \cdot 9}{3 \cdot 125} = \frac{225}{375} = \frac{3}{5}.$$

We can verify this common ratio by checking the ratio of the third term to the second term:

$$r = \frac{5}{\frac{25}{3}} = \frac{5 \cdot 3}{25} = \frac{15}{25} = \frac{3}{5}.$$

The common ratio is indeed $(\frac{3}{5})$. Now, we can use the formula for the (n)-th term of a geometric sequence, which is given by $(a_n = a_1 \cdot r^{n-1})$, where (a_1) is the first term and (r) is the common ratio. Here, $(a_1 = \frac{125}{9})$ and $(r = \frac{3}{5})$. We need to find the eighth term $((a_8))$:

$$a_8 = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^{8-1} = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^{7} = \frac{125}{9} \cdot \frac{3^7}{5^7} = \frac{125}{9} \cdot \frac{2187}{78125} = \frac{125 \cdot 2187}{9 \cdot 78125}.$$

We can simplify the fraction $(\frac{125 \cdot 2187}{9 \cdot 78125})$. Note that $(125 = 5^3)$ and $(78125 = 5^7)$, so:

$$\frac{125 \cdot 2187}{9 \cdot 78125} = \frac{5^3 \cdot 2187}{9 \cdot 5^7} = \frac{2187}{9 \cdot 5^4} = \frac{2187}{9 \cdot 625} = \frac{2187}{5625}.$$

We can further simplify $(\frac{2187}{5625})$ by finding the greatest common divisor (GCD) of 2187 and 5625. The prime factorization of 2187 is (3^7) and the prime factorization of 5625 is $(3^4 \cdot 5^4)$. The GCD is $(3^4 = 81)$:

$$\frac{2187 \div 81}{5625 \div 81} = \frac{27}{625}.$$

Thus, the eighth term of the sequence is $(\boxed{\frac{27}{625}})$.

Warm-up State in Our Method: As you can see, using our method of training, the model found out that: instead of spending too much tokens on explanation, it can achieve a higher score if it calculate explicitly and step by step. As the math tokens have a higher entropy, the model's total trajectory entropy will skyrocket at this stage. Moreover, in this stage, the model no longer mixes the two strategies together.

Strategy 1: using the power law:

Trajectory

To find the eighth term of the geometric sequence $\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots$, we first need to determine the common ratio of the sequence. The common ratio r is found by dividing the second term by the first term:

$$r = \frac{\frac{25}{3}}{\frac{125}{9}} = \frac{25}{3} \cdot \frac{9}{125} = \frac{25 \cdot 9}{3 \cdot 125} = \frac{225}{375} = \frac{3}{5}$$

We can verify that this common ratio is consistent by checking the next terms:

$$\frac{5}{\frac{25}{3}} = 5 \cdot \frac{3}{25} = \frac{15}{25} = \frac{3}{5}$$

$$\frac{3}{5} = \frac{3}{5}$$

Since the common ratio is consistent, we can use it to find the eighth term. The general form of the n-th term of a geometric sequence is given by:

$$a_n = a_1 \cdot r^{n-1}$$

Here, $a_1 = \frac{125}{9}$, $r = \frac{3}{5}$, and we need to find the eighth term (a_8):

$$a_8 = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^{8-1} = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^7$$

Now, we calculate $\left(\frac{3}{5}\right)^7$:

$$\left(\frac{3}{5}\right)^7 = \frac{3^7}{5^7} = \frac{2187}{78125}$$

Thus, the eighth term is:

$$a_8 = \frac{125}{9} \cdot \frac{2187}{78125}$$

To simplify this, we multiply the numerators and the denominators:

$$a_8 = \frac{125 \cdot 2187}{9 \cdot 78125}$$

We can simplify the fraction by finding the greatest common divisor (GCD) of the numerator and the denominator. First, we factorize the numbers:

$$125 = 5^3, \quad 2187 = 3^7, \quad 9 = 3^2, \quad 78125 = 5^7$$

Thus, the fraction becomes:

$$a_8 = \frac{5^3 \cdot 3^7}{3^2 \cdot 5^7} = \frac{3^5}{5^4} = \frac{243}{625}$$

Therefore, the eighth term of the sequence is:

$$\boxed{\frac{243}{625}}$$

Strategy 2: Brute Calculation

Trajectory

To find the eighth term of the geometric sequence $\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots$, we need to determine the common ratio and then use it to find the eighth term.
First, let's find the common ratio r of the sequence. The common ratio is the quotient of any term and the previous term. Let's use the first two terms to find r:

$$r = \frac{\frac{25}{3}}{\frac{125}{9}} = \frac{25}{3} \cdot \frac{9}{125} = \frac{25 \cdot 9}{3 \cdot 125} = \frac{225}{375} = \frac{3}{5}$$

We can verify that this common ratio is consistent by checking the next terms:

$$r = \frac{5}{\frac{25}{3}} = 5 \cdot \frac{3}{25} = \frac{15}{25} = \frac{3}{5}$$

$$r = \frac{3}{5} = \frac{3}{5}$$

Since the common ratio is consistent, we can use it to find the eighth term. The n-th term of a geometric sequence is given by:

$$a_n = a_1 \cdot r^{n-1}$$

Here, $a_1 = \frac{125}{9}$, $r = \frac{3}{5}$, and we need to find the eighth term (a_8):

$$a_8 = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^{8-1} = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^7$$

Next, we need to compute $\left(\frac{3}{5}\right)^7$:

$$\left(\frac{3}{5}\right)^7 = \frac{3^7}{5^7} = \frac{2187}{78125}$$

Now, we multiply this by $\frac{125}{9}$:

$$a_8 = \frac{125}{9} \cdot \frac{2187}{78125} = \frac{125 \cdot 2187}{9 \cdot 78125}$$

Simplify the numerator and the denominator:

$$125 \cdot 2187 = 273375$$

$$9 \cdot 78125 = 703125$$

Thus, we have:

$$a_8 = \frac{273375}{703125}$$

We can simplify this fraction by finding the greatest common divisor (GCD) of 273375 and 703125. Both numbers are divisible by 375:

$$273375 \div 375 = 731$$

$$703125 \div 375 = 1875$$

Thus, the simplified fraction is:

$$a_8 = \frac{731}{1875}$$

Therefore, the eighth term of the sequence is:

$$\boxed{\frac{731}{1875}}$$

Convergence with our Method: In the converged state: As the model undergoes further optimization, the two initially entangled rationales are completely decoupled and distinctly preserved across our 10 evaluation rollouts, exhibiting an empirical macro-policy distribution of 30% for Strategy I and 70% for Strategy II. Furthermore, each sub-policy has crystallized into a highly deterministic response template, where trajectories generated within the same strategy share a token-level semantic similarity exceeding 95%, limited only to minor choices of words. This localized rigidity alongside global balance strongly demonstrates that our proposed regularization framework strictly maintains the equilibrium of macroeconomic policy proportions without interfering with the model's capacity to optimize the localized execution of each specific strategy.

Strategy 1: using power law

Trajectory

To determine the eighth term of the geometric sequence $\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots$, we need to first identify the common ratio of the sequence. The common ratio r is found by dividing the second term by the first term:

$$r = \frac{\frac{25}{3}}{\frac{125}{9}} = \frac{25}{3} \cdot \frac{9}{125} = \frac{25 \cdot 9}{3 \cdot 125} = \frac{225}{375} = \frac{9}{15} = \frac{3}{5}$$

We can verify that this common ratio is consistent by checking the ratio of the third term to the second term:

$$r = \frac{5}{\frac{25}{3}} = 5 \cdot \frac{3}{25} = \frac{15}{25} = \frac{3}{5}$$

And the ratio of the fourth term to the third term:

$$r = \frac{3}{5} = \frac{3}{5}$$

Since the common ratio is consistent, we can now use the formula for the n-th term of a geometric sequence, which is given by $a_n = a_1 \cdot r^{n-1}$, where $a_1 = \frac{125}{9}$ and $r = \frac{3}{5}$. We need to find the eighth term (a_8):

$$a_8 = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^{8-1} = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^7 = \frac{125}{9} \cdot \frac{3^7}{5^7} = \frac{125 \cdot 3^7}{9 \cdot 5^7}$$

We know that $125 = 5^3$, so we can rewrite the expression as:

$$a_8 = \frac{5^3 \cdot 3^7}{9 \cdot 5^7} = \frac{3^7}{9 \cdot 5^4} = \frac{3^5}{5^4} = \frac{243}{625}$$

Thus, the eighth term of the sequence is:

$$\boxed{\frac{243}{625}}$$

Strategy 2: Brute Calculation

Trajectory

To determine the eighth term of the geometric sequence $\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots$, we need to first identify the common ratio of the sequence. The common ratio r is found by dividing the second term by the first term:

$$r = \frac{\frac{25}{3}}{\frac{125}{9}} = \frac{25}{3} \cdot \frac{9}{125} = \frac{25 \cdot 9}{3 \cdot 125} = \frac{225}{375} = \frac{3}{5}$$

We can verify that this common ratio is consistent by checking the ratio of the third term to the second term:

$$r = \frac{5}{\frac{25}{3}} = 5 \cdot \frac{3}{25} = \frac{15}{25} = \frac{3}{5}$$

And the ratio of the fourth term to the third term:

$$r = \frac{3}{5} = \frac{3}{5}$$

Since the common ratio is consistent, we can now use the formula for the n-th term of a geometric sequence, which is given by:

$$a_n = a_1 \cdot r^{n-1}$$

Here, $a_1 = \frac{125}{9}$, $r = \frac{3}{5}$, and we need to find the eighth term (a_8):

$$a_8 = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^{8-1} = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^7$$

We need to compute $\left(\frac{3}{5}\right)^7$:

$$\left(\frac{3}{5}\right)^7 = \frac{3^7}{5^7} = \frac{2187}{78125}$$

Now we substitute this back into the expression for a_8:

$$a_8 = \frac{125}{9} \cdot \frac{2187}{78125} = \frac{125 \cdot 2187}{9 \cdot 78125} = \frac{273375}{703125} = \frac{273375 \div 375}{703125 \div 375} = \frac{729}{1875} = \frac{729 \div 3}{1875 \div 3} = \frac{243}{625}$$

Thus, the eighth term of the sequence is:

$$\boxed{\frac{243}{625}}$$

Collapsed Model:

Trajectory

To find the eighth term of the given geometric sequence $\frac{125}{9}, \frac{25}{3}, 5, 3, \ldots$, we first need to determine the common ratio r of the sequence. The common ratio r can be found by dividing the second term by the first term:

$$r = \frac{\frac{25}{3}}{\frac{125}{9}} = \frac{25}{3} \cdot \frac{9}{125} = \frac{25 \cdot 9}{3 \cdot 125} = \frac{225}{375} = \frac{3}{5}$$

Now that we have the common ratio $r = \frac{3}{5}$, we can use the formula for the n-th term of a geometric sequence, which is given by $a_n = a_1 \cdot r^{n-1}$, where a_1 is the first term and a_n is the n-th term. Here, $a_1 = \frac{125}{9}$, $r = \frac{3}{5}$, and we need to find the eighth term (a_8):

$$a_8 = a_1 \cdot r^{8-1} = \frac{125}{9} \cdot \left(\frac{3}{5}\right)^7$$

Next, we need to compute $\left(\frac{3}{5}\right)^7$:

$$\left(\frac{3}{5}\right)^7 = \frac{3^7}{5^7} = \frac{2187}{78125}$$

Now, we substitute this back into the expression for a_8:

$$a_8 = \frac{125}{9} \cdot \frac{2187}{78125} = \frac{125 \cdot 2187}{9 \cdot 78125} = \frac{273375}{703125} = \boxed{\boxed{\boxed{\boxed{\frac{9}{25}}}}}$$

E.1.1 Science Case Study

Mesh Learning — Coherent and Structured Reasoning. The first trajectory exhibits a clear and well-organized reasoning process: it progressively combines energy conservation, momentum conservation, and the relativistic energy–momentum relation, while each intermediate step directly contributes to the final solution. Occasional self-correction remains local and productive, resulting in a coherent trajectory that terminates naturally after reaching the correct answer.

Trajectory

<think>
Alright, let's tackle this problem step by step. So, we have the reaction where a π^+ (which is stationary) decays into a μ^+ and a ν (I assume the ν is a muon, maybe a muon neutrino?). The rest masses given are 139.6 MeV for π^+ and 105.7 MeV for the μ^+ (mu+ ion). We need to find the kinetic energies of the product particles, which are μ^+ and ν. The options are given, so hopefully, after working through it, the answer will match one of them.
First, I remember that in particle physics, when a particle decays, the total energy and momentum must be conserved. Since the π^+ is stationary, its initial momentum is zero. Therefore, the momentum of the decay products must also sum to zero. That means the μ^+ and the ν must have equal and opposite momenta.
Next, I need to recall the relationship between energy and momentum in relativity. The total energy of a particle is given by the relativistic energy equation:

$$E = \sqrt{(pc)^2 + (mc^2)^2}.$$

Since the π^+ is at rest, its total energy is just its rest mass energy, which is 139.6 MeV. The total energy of the products should equal this.
But wait, the μ^+ and the ν are two different particles, so I need to consider their rest masses. The μ^+ has a rest mass of $105.7\,\mathrm{MeV}/c^2$, and the ν (assuming it's a neutrino) is presumably massless, so its rest mass is 0. That simplifies things a bit.
Let me denote the muon's kinetic energy as K_μ and the neutrino's kinetic energy as K_ν. The total kinetic energy will be the sum of these two, and it should equal the rest mass energy of the π^+ because the π^+ decays into these two, so the mass energy is converted into the kinetic energies of the products. Wait, is that correct?
Actually, the total energy of the π^+ is 139.6 MeV, which converts into the sum of the muon's total energy (rest + kinetic) and the neutrino's total energy (which should just be kinetic since it's massless). So, the total energy conservation equation is:

$$E_\pi = E_\mu + E_\nu,$$

where

$$E_\pi = 139.6\,\mathrm{MeV}.$$

$$E_\mu = K_\mu + m_\mu c^2 = K_\mu + 105.7\,\mathrm{MeV},$$

and

$$E_\nu = K_\nu + m_\nu c^2.$$

But since neutrinos are massless, $m_\nu c^2 = 0$, so

$$E_\nu = K_\nu.$$

Therefore,

$$139.6 = (K_\mu + 105.7) + K_\nu,$$

which simplifies to

$$K_\mu + K_\nu = 139.6 - 105.7 = 33.9\,\mathrm{MeV}.$$

So the sum of the kinetic energies is 33.9 MeV. But we need individual values. To find that, we also need to apply conservation of momentum.
Since the π^+ is initially at rest, the momenta of the μ^+ and ν must be equal in magnitude and opposite in direction. Therefore,

$$|p_\mu| = |p_\nu| = p.$$

Using the energy and momentum relation for the muon:

$$E_\mu^2 = (pc)^2 + (m_\mu c^2)^2.$$

Similarly for the neutrino:

$$E_\nu = pc$$

since $m_\nu = 0$.

So, we have two equations:

$$139.6 = E_\mu + E_\nu = (K_\mu + 105.7) + pc,$$

and

$$E_\nu = pc.$$

From equation 2,

$$pc = E_\nu.$$

Substitute into equation 1:

$$139.6 = K_\mu + 105.7 + E_\nu.$$

But from the muon's energy-momentum relation:

$$E_\mu = \sqrt{(pc)^2 + (m_\mu c^2)^2}.$$

But $E_\mu = K_\mu + 105.7$, and $pc = E_\nu$.
Therefore:

$$K_\mu + 105.7 = \sqrt{(E_\nu)^2 + (105.7)^2}.$$

And from the previous equation:

$$K_\mu = 139.6 - 105.7 - E_\nu = 33.9 - E_\nu.$$

Substitute K_μ into the equation:

$$33.9 - E_\nu + 105.7 = \sqrt{(E_\nu)^2 + (105.7)^2}.$$

Simplify the left side:

$$139.6 - E_\nu = \sqrt{(E_\nu)^2 + (105.7)^2}.$$

Square both sides:

$$(139.6 - E_\nu)^2 = (E_\nu)^2 + (105.7)^2.$$

Expand the left side:

$$139.6^2 - 2 \cdot 139.6 \cdot E_\nu + (E_\nu)^2 = (E_\nu)^2 + 105.7^2.$$

Cancel $(E_\nu)^2$ from both sides:

$$139.6^2 - 2 \cdot 139.6 \cdot E_\nu = 105.7^2.$$

Calculate 139.6^2 and 105.7^2:

$$139.6^2 = 19488.16,$$
$$105.7^2 = 11174.49.$$

So:

$$19488.16 - 2 \cdot 139.6 \cdot E_\nu = 11174.49.$$

Subtract 11174.49 from both sides:

$$19488.16 - 11174.49 = 2 \cdot 139.6 \cdot E_\nu.$$

Calculate the left side:

$$8323.67 = 279.2 \cdot E_\nu.$$

Therefore:

$$E_\nu = \frac{8323.67}{279.2} \approx 29.8\,\text{MeV}.$$

Then, from earlier,

$$K_\nu = E_\nu \approx 29.8\,\text{MeV},$$

and

$$K_\mu = 33.9 - 29.8 = 4.1\,\text{MeV}.$$

So the kinetic energies are approximately 4.1 MeV for the muon and 29.8 MeV for the neutrino. Looking at the options, that's B.
</think>
To find the kinetic energies of the product particles in the decay of π^+ (which is stationary) into μ^+ and ν, we start by applying conservation of energy and momentum.

1. **Energy Conservation:** The total energy of π^+ is its rest mass energy:
$$E_\pi = 139.6\,\text{MeV}.$$
The total energy of μ^+ is
$$E_\mu = K_\mu + m_\mu c^2 = K_\mu + 105.7\,\text{MeV}.$$
The total energy of ν (neutrino) is
$$E_\nu = K_\nu,$$
since it is massless.
Therefore,
$$139.6 = (K_\mu + 105.7) + K_\nu,$$
which simplifies to
$$K_\mu + K_\nu = 33.9\,\text{MeV}.$$
2. **Momentum Conservation:** Since π^+ is stationary, the momenta of μ^+ and ν must be equal in magnitude and opposite in direction:
$$p_\mu = p_\nu = p.$$
3. **Relativistic Energy-Momentum Relation:** For μ^+:
$$E_\mu = \sqrt{(pc)^2 + (m_\mu c^2)^2}.$$
For ν:
$$E_\nu = pc.$$
Substituting $E_\mu = K_\mu + 105.7$ and $E_\nu = pc$ into the energy conservation equation, we get
$$139.6 = K_\mu + 105.7 + E_\nu,$$
and
$$K_\mu + 105.7 = \sqrt{(E_\nu)^2 + (105.7)^2}.$$
4. **Solving for E_ν:** Substitute
$$K_\mu = 33.9 - E_\nu$$
into the equation:
$$33.9 - E_\nu + 105.7 = \sqrt{(E_\nu)^2 + (105.7)^2}.$$
Simplifying and squaring both sides, we find:
$$8323.67 = 279.2 \cdot E_\nu,$$
$$E_\nu \approx 29.8\,\text{MeV}.$$
5. **Finding K_μ and K_ν:**
$$K_\nu = E_\nu \approx 29.8\,\text{MeV},$$
$$K_\mu = 33.9 - 29.8 \approx 4.1\,\text{MeV}.$$

Thus, the kinetic energies are

$$\boxed{B}: \quad 4.12\,\text{MeV} \quad \text{and} \quad 29.8\,\text{MeV}.$$

Normal RLVR — Redundant and Error-Prone Reasoning. In contrast, the second trajectory is substantially more meandering and unstable. It repeatedly explores unnecessary branches, partially infers the answer from the multiple-choice options before completing the derivation, introduces incorrect intermediate formulas, and contains numerical inaccuracies despite eventually recovering the correct answer. Such behavior reflects a less efficient and less reliable reasoning trajectory, where

successful completion relies on repeated correction rather than a consistently coherent solution process.

Trajectory

<think>
Okay, let's try to figure out the kinetic energy (KE) of the product particles, μ^+ and ν, in the decay of a stationary π^+ particle. The rest masses are given: π^+ is $139.6\,\mathrm{MeV}/c^2$ and μ^+ is $105.7\,\mathrm{MeV}/c^2$. The options are A to D with different KE values.
First, I remember that in particle physics, when a particle decays into others, the total energy and momentum are conserved. Since the π^+ is stationary, its initial momentum is zero. After decay, the μ^+ and ν must have equal and opposite momenta to conserve momentum.
Energy conservation says that the rest energy of π^+ converts into the total energy of μ^+ and ν. The total energy for each particle is γmc^2 + kinetic energy. But since they are moving, their relativistic energies are involved. However, maybe we can use the relativistic energy-momentum relation:

$$E^2 = (pc)^2 + (mc^2)^2.$$

Let me denote the rest masses as $m_\pi = 139.6\,\mathrm{MeV}/c^2$ and $m_\mu = 105.7\,\mathrm{MeV}/c^2$. Let the kinetic energies be KE_μ and KE_ν. The total energies would be $E_\mu = m_\mu c^2 + KE_\mu$ and $E_\nu = m_\nu c^2 + KE_\nu$. But wait, the problem states that π^+ decays into μ^+ and ν. Is ν a neutrino? Typically, in such decays, the neutrino is massless ($m_\nu \approx 0$), so its rest energy would be negligible. That simplifies things because then the energy of the neutrino is just its kinetic energy, and the muon has its rest mass plus kinetic energy.
So, applying conservation of energy: the rest energy of π^+ equals the sum of the energies of μ^+ and the neutrino. Since the neutrino is massless, $E_\nu = KE_\nu$.
So,

$$m_\pi c^2 = (m_\mu c^2 + KE_\mu) + (m_\nu c^2 + KE_\nu).$$

But if $m_\nu c^2$ is negligible, then $m_\nu c^2 \approx 0$. So,

$$m_\pi c^2 \approx m_\mu c^2 + KE_\mu + KE_\nu.$$

Therefore,

$$KE_\nu + KE_\mu = m_\pi c^2 - m_\mu c^2 = (139.6 - 105.7)\,\mathrm{MeV} = 33.9\,\mathrm{MeV}.$$

So the sum of the kinetic energies is 33.9 MeV. But the answer options have different sums. For example, option A is $3.52 + 20.8 = 24.32$, B: $4.12 + 29.8 = 33.92$, which is close to 33.9. Hmm, that's interesting. So maybe B is the answer? But wait, this assumes that the neutrino's rest mass is zero, which is a standard assumption. But we need to check if this is the correct approach.
But also, momentum conservation must hold. Since the initial momentum is zero, the momenta of μ^+ and ν must be equal in magnitude and opposite in direction. Using the relativistic relation: for each particle,

$$p^2 = \frac{E^2}{c^2} - \frac{(mc^2)^2}{c^2}.$$

But since momentum p is the same magnitude for both, let's set their momenta equal.
For the muon:

$$p^2 = \frac{(E_\mu)^2}{c^2} - \frac{(m_\mu c^2)^2}{c^2}.$$

Similarly, for the neutrino:

$$p^2 = \frac{(E_\nu c^2)^2}{c^2} - \frac{(m_\nu c^2)^2}{c^2}.$$

But if $m_\nu \approx 0$, then

$$p^2 \approx \frac{(E_\nu c^2)^2}{c^2} = E_\nu^2.$$

So

$$E_\nu = pc.$$

But since the momenta are equal, $p = p$, so for the muon,

$$E_\mu = \sqrt{(pc)^2 + (m_\mu c^2)^2},$$

and for the neutrino,

$$E_\nu = pc$$

since $m_\nu = 0$.
From energy conservation:

$$E_\mu + E_\nu = m_\pi c^2.$$

Substituting $E_\nu = pc$, and

$$E_\mu = \sqrt{(pc)^2 + (m_\mu c^2)^2},$$

we obtain

$$\sqrt{(pc)^2 + (m_\mu c^2)^2} + pc = m_\pi c^2.$$

Let me square both sides to eliminate the square root, but maybe it's easier to let $x = pc$. Then:

$$\sqrt{x^2 + (m_\mu c^2)^2} + x = m_\pi c^2.$$

Let's denote $m_\mu c^2$ as a and $m_\pi c^2$ as M. Then the equation becomes:

$$\sqrt{x^2 + a^2} + x = M.$$

Let's solve for x:

$$\sqrt{x^2 + a^2} = M - x.$$

Square both sides:

$$x^2 + a^2 = M^2 - 2Mx + x^2.$$

Cancel x^2:

$$a^2 = M^2 - 2Mx.$$

Then:

$$2Mx = M^2 - a^2,$$

and

$$x = \frac{M^2 - a^2}{2M}.$$

But x is pc, so:

$$pc = \frac{M^2 - a^2}{2M}.$$

Therefore, the kinetic energy of the neutrino is

$$E_\nu = \frac{(pc)^2}{2m_\nu c^2}$$

but since $m_\nu \approx 0$, this would be problematic. Wait, no. For a massless particle, the energy is $E_\nu = pc$. So

$$KE_\nu = E_\nu - m_\nu c^2 \approx E_\nu = pc.$$

But for the muon,

$$KE_\mu = E_\mu - m_\mu c^2 = \sqrt{(pc)^2 + (m_\mu c^2)^2} - m_\mu c^2.$$

We can substitute pc from above:

$$pc = \frac{M^2 - a^2}{2M}.$$

So

$$E_\mu = \sqrt{\left(\frac{M^2 - a^2}{2M}\right)^2 + a^2}.$$

Let me compute that:

$$E_\mu = \sqrt{\frac{(M^2 - a^2)^2}{4M^2} + a^2}.$$

$$= \sqrt{\frac{M^4 - 2M^2a^2 + a^4 + 4M^2a^2}{4M^2}}.$$

$$= \sqrt{\frac{M^4 + 2M^2a^2 + a^4}{4M^2}}.$$

$$= \sqrt{\frac{(M^2 + a^2)^2}{4M^2}}.$$

$$= \frac{M^2 + a^2}{2M}.$$

Therefore,

$$E_\mu = \frac{M^2 + a^2}{2M}.$$

So

$$KE_\mu = E_\mu - a = \frac{M^2 + a^2}{2M} - a = \frac{M^2 + a^2 - 2Ma}{2M} = \frac{(M - a)^2}{2M}.$$

Similarly,

$$KE_\nu = pc = \frac{M^2 - a^2}{2M}.$$

So let's compute these values.
Given

$$M = 139.6\,\mathrm{MeV}/c^2, \qquad a = 105.7\,\mathrm{MeV}/c^2.$$

First compute KE_ν:

$$KE_\nu = \frac{M^2 - a^2}{2M} = \frac{(139.6)^2 - (105.7)^2}{2 \times 139.6}.$$

Calculate numerator:

$$139.6^2 = 19488.16,$$
$$105.7^2 = 11174.49.$$

Difference:

$$19488.16 - 11174.49 = 8313.67.$$

Then

$$KE_\nu = \frac{8313.67}{2 \times 139.6} = \frac{8313.67}{279.2} \approx 29.8\,\mathrm{MeV}.$$

Then

$$KE_\mu = \frac{(M - a)^2}{2M} = \frac{(139.6 - 105.7)^2}{2 \times 139.6}.$$

Compute

$$139.6 - 105.7 = 33.9.$$

Square of 33.9:

$$33.9^2 = 1149.21.$$

So

$$KE_\mu = \frac{1149.21}{279.2} \approx 4.12\,\mathrm{MeV}.$$

So $KE_\mu \approx 4.12$ MeV and $KE_\nu \approx 29.8$ MeV, which corresponds to option B.
Therefore, the answer should be B.
</think>
Step-by-Step Explanation:

1. **Conservation Laws.** Energy conservation implies that the rest energy of the stationary π^+ decays into the total energy (rest + kinetic) of μ^+ and ν. Momentum conservation implies that the momenta of μ^+ and ν must cancel, with equal magnitudes and opposite directions.
2. **Assumptions.** The neutrino ν is massless ($m_\nu \approx 0$), so its energy equals its kinetic energy:
$$KE_\nu = E_\nu = pc.$$
3. **Relativistic Energy-Momentum Relation.** For μ^+:
$$E_\mu = \sqrt{(pc)^2 + (m_\mu c^2)^2}.$$
For ν:
$$E_\nu = pc.$$
4. **Energy Conservation Equation.**
$$m_\pi c^2 = E_\mu + E_\nu = \sqrt{(pc)^2 + (m_\mu c^2)^2} + pc.$$
Let $x = pc$, solve for x:
$$\sqrt{x^2 + (m_\mu c^2)^2} + x = m_\pi c^2.$$
Squaring and simplifying yields:
$$x = \frac{(m_\pi c^2)^2 - (m_\mu c^2)^2}{2m_\pi c^2}.$$
5. **Kinetic Energy Calculations.**
Neutrino KE:
$$KE_\nu = x = \frac{139.6^2 - 105.7^2}{2 \times 139.6} \approx 29.8\,\text{MeV}.$$
Muon KE:
$$KE_\mu = \frac{(139.6 - 105.7)^2}{2 \times 139.6} \approx 4.12\,\text{MeV}.$$

Final Answer:

$$\boxed{B} \qquad 4.12\,\text{MeV},\ 29.8\,\text{MeV}.$$

Collapsed — Tragic and Chaotic Reasoning For collapsed models during the RLVR training, although the model repeatedly reaches the correct solution, it fails to terminate and continuously reopens already resolved subproblems. The trajectory cycles through redundant re-derivations, irrelevant alternatives, and increasingly peripheral considerations (e.g., helicity, neutrino mass, and weak-interaction details) without contributing new information to the solution. This severe loss of termination and reasoning focus illustrates a characteristic late-stage degeneration: the model retains local problem-solving ability, yet can no longer maintain a stable global reasoning trajectory.

Trajectory

<think>
Okay, so I need to find the kinetic energy (KE) of the product particles, μ^+ and ν, from the decay of a stationary π^+ particle. The rest masses given are 139.6 MeV for π^+ and 105.7 MeV for μ^+. The options are A to D with different KE values for each particle. Let me think step by step.
First, I remember that in nuclear or particle physics reactions, conservation laws play a big role. Since the π^+ is stationary, the total energy and momentum before and after the decay must be conserved. The rest mass energy of π^+ will convert into the rest mass energies of μ^+ and ν, plus their kinetic energies.

But wait, the problem mentions "rest mass of Pi+ & mu+ is 139.6 MeV & 105.7 MeV respectively." Does that mean the rest mass of ν is not given? Hmm, maybe the decay is

$$\pi^+ \to \mu^+ + \nu.$$

So, this must be a two-body decay because it's specified as product particles. If it's a two-body decay, then the momentum of both μ^+ and ν should be equal in magnitude but opposite in direction to conserve momentum. Since the parent is stationary, the total momentum before decay is zero, so the momenta of the two products must cancel each other.

In a two-body decay, each particle's kinetic energy can be found using the conservation of energy and momentum. The total energy released (the Q-value) would be the difference between the rest mass of the parent and the sum of the rest masses of the products. Wait, but the rest mass of ν isn't given here. That's a problem. Wait, maybe ν is a known particle? Like in the case of the pion decay, which is usually μ^+ and a neutrino (ν_μ). So in reality, the neutrino has a very small mass, but in this problem, maybe they're giving the rest masses of the μ^+ and π^+ but not ν? That can't be right because energy must be conserved.

Wait, perhaps there's a typo or misunderstanding. Let me check the problem again. It says: "Find KE of product particles in, Pi(+) = mu(+) + nu here Pi(+) is stationary. Rest mass of Pi(+)& mu(+)=139.6 MeV & 105.7 MeV respectively." So the rest mass of ν isn't provided. That's confusing because we need to know the rest mass of all products to calculate the energy available for kinetic energy.

Wait, maybe in this problem, the ν is assumed to be massless, like a neutrino, so its rest mass is negligible. But even so, in reality, the neutrino's mass is very small, but if we take it as zero, then the entire Q-value would go into the kinetic energies of μ^+ and ν. Let me try that approach.

So, the rest mass energy of π^+ is 139.6 MeV. If ν is massless, its rest mass is 0, so the total energy available for kinetic energy would be 139.6 MeV. Then, the sum of the kinetic energies of μ^+ and ν would be 139.6 MeV. But in a two-body decay, even if one particle is massless, the energies can be calculated using relativistic energy-momentum conservation.

Wait, but in reality, if ν is massless, then its energy would be equal to its momentum times the speed of light, since

$$E = pc$$

for massless particles. However, for the μ^+, which has mass, its energy would be

$$E = \gamma m_\mu c^2,$$

where γ is the Lorentz factor.

But in a two-body decay, even if one particle is massless, the calculation can be done using the Lorentz boost. Let's recall the formula for two-body decays. In general, for a decay

$$A \to B + C,$$

the energies of B and C can be found using

$$E_B = \frac{m_A^2 + m_B^2 - m_C^2}{2m_A} c^2,$$

and

$$E_C = \frac{m_A^2 + m_C^2 - m_B^2}{2m_A} c^2.$$

But this is when both B and C have rest masses. However, if C is massless, then $m_C = 0$, so the formula simplifies. Let me check.

If ν is massless ($m_C = 0$), then

$$E_B = \frac{m_A^2 + m_B^2}{2m_A} c^2,$$

and

$$E_C = \frac{m_A^2 - m_B^2}{2m_A} c^2.$$

But since m_C is massless, its energy is equal to its momentum,

$$E_C = p_C c.$$

Also, the magnitudes of the momenta of B and C must be equal because the parent was at rest.
But the kinetic energy of B would be

$$E_B - m_B c^2,$$

and the kinetic energy of C would be

$$E_C - m_C c^2 = E_C.$$

So, let's compute E_B and E_C using the formulas.
Given:

$$m_A = 139.6\,\mathrm{MeV}/c^2,$$
$$m_B = 105.7\,\mathrm{MeV}/c^2,$$
$$m_C = 0.$$

Compute E_B:

$$E_B = \frac{m_A^2 + m_B^2 - m_C^2}{2m_A} c^2.$$

But since $m_C = 0$,

$$E_B = \frac{m_A^2 + m_B^2}{2m_A} c^2.$$

Similarly,

$$E_C = \frac{m_A^2 - m_B^2}{2m_A} c^2.$$

But let's plug the numbers.
First, compute m_A^2:

$$m_A^2 = (139.6)^2.$$

139.6^2 is approximately $(140)^2 = 19600$, but more accurately:

$$139.6 \times 139.6.$$

$140 \times 140 = 19600$, subtract

$$0.4 \times 140 + 0.4 \times 139.6 = 56 + 55.84.$$

Wait, maybe better to compute directly:

$$139.6^2 = (130 + 9.6)^2 = 130^2 + 2(130)(9.6) + 9.6^2 = 16900 + 2496 + 92.16 = 19488.16.$$

Similarly,

$$m_B^2 = (105.7)^2.$$

$$105^2 = 11025, \qquad 0.7^2 = 0.49, \qquad 2(105)(0.7) = 147,$$

so

$$(105 + 0.7)^2 = 11025 + 147 + 0.49 = 11172.49.$$

Now compute E_B:

$$E_B = \frac{19488.16 + 11172.49}{2(139.6)} = \frac{30660.65}{279.2} \approx 109.94\,\mathrm{MeV}.$$

Then KE of μ^+ is

$$E_B - m_B = 109.94 - 105.7 \approx 4.24\,\mathrm{MeV}.$$

Now compute E_C:

$$E_C = \frac{19488.16 - 11172.49}{2(139.6)} = \frac{8315.67}{279.2} \approx 29.84\,\text{MeV}.$$

Since E_C is the energy of the neutrino, which is kinetic energy because its rest mass is zero. So the KE of μ^+ is approx 4.24 MeV and KE of ν is approx 29.84 MeV. Looking at the options:

$$\text{B.} \qquad 4.12\,\text{MeV}, \quad 29.8\,\text{MeV}.$$

That's very close. The slight difference might be due to rounding during calculations. Let me check the exact calculation.
Wait, alternatively, maybe I should not subtract the rest mass yet, because the formula for E_B already gives the energy in the center-of-mass frame, but the pion is stationary, so this is also the lab frame. Yes, so I should subtract it. But wait, if the neutrino is massless, then perhaps the muon kinetic energy can instead be obtained directly from momentum conservation.
Let me recompute.
We have

$$m_A^2 - m_B^2 = 19488.16 - 11172.49 = 8315.67,$$

so the neutrino energy should be

$$E_C = \frac{8315.67}{279.2}.$$

This is about

$$29.8\,\text{MeV}.$$

Then the muon's total energy should be

$$139.6 - 29.8 = 109.8\,\text{MeV},$$

so its kinetic energy is

$$109.8 - 105.7 = 4.1\,\text{MeV}.$$

So that seems to give B.
Wait, alternatively, if I use the total released energy first,

$$139.6 - 105.7 = 33.9\,\text{MeV},$$

and distribute it according to momentum, because the two particles have equal momentum but different masses, the neutrino should receive more energy. This is consistent.
But let me verify whether

$$29.8 + 4.12$$

is exactly 33.9.

$$29.8 + 4.12 = 33.92.$$

That's slightly larger than 33.9. Maybe this is rounding. Wait, but if it is larger, perhaps one of the numbers is not kinetic energy but total energy? No, 29.8 is neutrino energy, which is kinetic energy. 4.12 is muon kinetic energy.
Alternatively, perhaps the rest mass values are rounded, which would explain the 0.02 MeV discrepancy.
Wait, let me calculate

$$\frac{8315.67}{279.2}$$

more exactly.

$$279.2 \times 29.8 = 8320.16.$$

But 8320.16 is larger than 8315.67 by 4.49. So actually 29.8 is not exactly correct.
Maybe it is 29.78.

$$279.2 \times 29.78 = 279.2 \times 29 + 279.2 \times 0.78$$

$$= 8096.8 + 217.776 = 8314.576.$$

So about 29.784.
Then the muon KE would be

$$33.9 - 29.784 = 4.116\,\text{MeV}.$$

Yes, that gives B.
Wait, alternatively, maybe I made an error because the kinetic energies don't have to sum to 33.9 if momentum also contributes separately. But momentum is already included in kinetic energy, so they should.
Wait.
Let me derive it another way to make sure.
The muon momentum is equal to the neutrino momentum:

$$p_\mu = p_\nu.$$

For the neutrino,

$$E_\nu = pc.$$

For the muon,

$$KE_\mu = \sqrt{p^2c^2 + m_\mu^2 c^4} - m_\mu c^2.$$

Then

$$KE_\mu + E_\nu = 33.9.$$

So

$$\sqrt{p^2c^2 + m_\mu^2 c^4} - m_\mu c^2 + pc = 33.9.$$

Adding $m_\mu c^2$,

$$\sqrt{p^2c^2 + m_\mu^2 c^4} + pc = 139.6.$$

This is the same equation as before.
So B.
Wait, alternatively, the problem may expect a nonrelativistic approximation for the muon because its kinetic energy is only about 4 MeV compared with 105.7 MeV rest energy. Then

$$KE_\mu = \frac{p^2}{2m_\mu},$$

while for the neutrino

$$E_\nu = pc.$$

If I combine these, perhaps I get slightly different values.
Let

$$x = pc.$$

Then

$$KE_\mu = \frac{x^2}{2m_\mu c^2},$$

and

$$x + \frac{x^2}{2(105.7)} = 33.9.$$

This is a quadratic.
Wait, but if I solve this approximation, it should also give around 29 MeV for x.
Let me see:

$$x^2 + 211.4x - 2(105.7)(33.9) = 0.$$

$2(105.7)(33.9)$ is approximately 7166?

$$105.7 \times 33.9 = 105.7(34 - 0.1) = 3593.8 - 10.57 = 3583.23.$$

Twice is

$$7166.46.$$

So

$$x^2 + 211.4x - 7166.46 = 0.$$

Then

$$x = \frac{-211.4 + \sqrt{211.4^2 + 4(7166.46)}}{2}.$$

$$211.4^2 \approx 44689.96.$$

$$4(7166.46) = 28665.84.$$

Sum:

$$73355.8.$$

The square root is about

$$270.8.$$

Then

$$x = \frac{59.4}{2} = 29.7.$$

Again, near B.

Wait, alternatively, if the neutrino has a tiny but nonzero rest mass, then strictly E_ν is not pc. But the correction is negligible. Unless the question is testing that? No, the options differ by several MeV, so probably not.

But wait, the wording says "rest mass of Pi+ & mu+ is 139.6 MeV & 105.7 MeV respectively." They give masses in MeV rather than MeV/c^2, which usually means natural units $c = 1$. Then perhaps I should consistently set $c = 1$.

Then

$$E_\mu^2 = p^2 + m_\mu^2,$$

and

$$E_\nu = p.$$

Then

$$\sqrt{p^2 + 105.7^2} + p = 139.6.$$

We already solved this.

Wait, alternatively, I can isolate p:

$$\sqrt{p^2 + 105.7^2} = 139.6 - p.$$

Squaring:

$$p^2 + 105.7^2 = 139.6^2 - 279.2p + p^2.$$

So

$$279.2p = 139.6^2 - 105.7^2.$$

Yes.

But wait, I previously got 8315.67 for this difference. Let me check again because maybe this is where the discrepancy comes from.

$$139.6^2 = 19488.16.$$

$$105.7^2 = 11172.49.$$

Subtract:

$$19488.16 - 11172.49 = 8315.67.$$

Yes.

Then

$$p = \frac{8315.67}{279.2}.$$

Wait, perhaps I can simplify algebraically:

$$M^2 - m^2 = (M - m)(M + m) = 33.9 \times 245.3.$$

$$33.9 \times 245.3.$$

$$245.3 \times 30 = 7359.$$

$$245.3 \times 3 = 735.9.$$

$$245.3 \times 0.9 = 220.77.$$

Total:

$$8315.67.$$

Yes.
Then divide by 279.2.
Wait,

$$279.2 = 2 \times 139.6.$$

So

$$p = \frac{33.9 \times 245.3}{279.2}.$$

The ratio

$$\frac{245.3}{279.2}$$

is about 0.8786.

$$33.9 \times 0.8786$$

is about 29.78.
So yes.
Wait, alternatively, maybe the kinetic energy of the muon is not simply $33.9 - 29.78$ because the 33.9 itself was obtained by subtracting rest energies, but yes, that is exactly what kinetic energy means.
So it is B.
Wait.
But maybe I should verify option C too in case the ordering is reversed.

$$\text{C.} \qquad 29.8\,\text{MeV}, \quad 4.12\,\text{MeV}.$$

If the first number corresponds to μ^+ and the second to ν, then C would assign most energy to the massive muon. But equal momentum means the lighter particle receives larger kinetic energy, so B is correct.
Wait, alternatively, for equal momentum, the massive particle has larger total energy because of rest mass, but smaller kinetic energy. So yes, B.
Wait, but if the question asks "KE of product particles, mu+ and nu" perhaps the order isn't explicitly stated? It probably means μ^+ first and ν second.
Wait, alternatively, perhaps the decay notation

$$\pi^+ = \mu^+ + \nu$$

implies the first KE belongs to π? No, π is the parent.
Let me make sure.
π^+ stationary.

$$\pi^+ \to \mu^+ + \nu.$$

Momentum:

$$0 = p_\mu + p_\nu.$$

So

$$p_\mu = -p_\nu.$$

Energy:

$$139.6 = \sqrt{p^2 + 105.7^2} + p.$$

Then

$$p = 29.78.$$

Then mu's energy:

$$109.82.$$

Then

$$KE_\mu = 4.12.$$

So B.
Wait, alternatively, perhaps

$$109.82 - 105.7 = 4.12,$$

yes.
But wait,

$$109.82 + 29.78 = 139.60.$$

Yes.
So that closes.
Wait, no, alternatively, there may be another subtlety: the neutrino is actually an antineutrino in the decay

$$\pi^+ \to \mu^+ + \nu_\mu?$$

Actually charge and lepton number conservation would require ν_μ or maybe ν—wait, π^+ has no lepton number, μ^+ has lepton number -1, so the neutrino should have $+1$, so ν_μ.
Yes.
But this doesn't affect the energy.
Wait, unless helicity constraints affect the kinematics. No, two-body kinematics fixes the energies independently of spin.
Alternatively, perhaps because the pion has spin zero and the neutrino is left-handed, the angular momentum configuration constrains the muon helicity, but again not the kinetic energy.
Wait, but if the neutrino is exactly massless, helicity is fixed, while the antimuon—
Wait, I might be overcomplicating this. The answer is B.
But alternatively, let me verify the pion decay formula once more because maybe I swapped the signs:

$$E_\mu = \frac{m_\pi^2 + m_\mu^2}{2m_\pi}.$$

$$E_\nu = \frac{m_\pi^2 - m_\mu^2}{2m_\pi}.$$

Yes.
Then

$$KE_\mu = E_\mu - m_\mu.$$

But wait, if I rewrite KE_μ,

$$KE_\mu = \frac{m_\pi^2 + m_\mu^2 - 2m_\pi m_\mu}{2m_\pi} = \frac{(m_\pi - m_\mu)^2}{2m_\pi}.$$

Yes.
So

$$\frac{33.9^2}{279.2}.$$

$$33.9^2 = 1149.21.$$

$$\frac{1149.21}{279.2}.$$

$$279.2 \times 4 = 1116.8.$$

Remainder:

$$32.41.$$

$$\frac{32.41}{279.2} = 0.1161.$$

So

$$4.1161.$$

Yes.
Wait, alternatively,

$$4.1161 + 29.7839 = 33.9000.$$

So exactly.
Wait.
Wait, alternatively, if I retain the neutrino mass as ϵ, then the expressions become

$$E_\mu = \frac{m_\pi^2 + m_\mu^2 - m_\nu^2}{2m_\pi},$$

and

$$E_\nu = \frac{m_\pi^2 + m_\nu^2 - m_\mu^2}{2m_\pi}.$$

Then, in the limit

$$m_\nu \to 0$$

we recover this.
But what if the problem intentionally omitted the neutrino mass because we're supposed to derive it from the answer choices?
No, that doesn't make sense.
Wait, alternatively, maybe "nu" means something else, not neutrino? The symbol ν could be frequency ν, but in a particle reaction it must be neutrino.
Unless this is not a decay but some equation involving frequency?

$$\pi^+ = \mu^+ + \nu.$$

No, definitely particle decay.
Wait.
The option values: 4.12 and 29.8.
Their sum is

$$33.92,$$

matching the Q-value.
So B.
Wait, alternatively—
If the parent is stationary, maybe there is also recoil energy of the parent?
No, the parent no longer exists after decay.
Wait, alternatively, perhaps the "stationary" condition means zero kinetic energy but not zero momentum due to quantum uncertainty?
No, not relevant.
Wait, alternatively, perhaps I should use classical momentum conservation:

$$p_\mu = p_\nu,$$

$$\sqrt{2m_\mu K E_\mu} = \frac{E_\nu}{c},$$

then solve.
But that gives the approximation we already did.
Wait, but if I use the classical relation for the muon, I get something close to 4.2, which could perhaps explain option B.
Wait.
Alternatively, perhaps the intended answer is obtained simply by conservation of energy and choosing the option whose sum is 33.9, without momentum conservation.
Then B is uniquely consistent.
But that would be too easy. Unless the problem is multiple choice and expects that shortcut.

Wait, if only B sums to 33.9, yes.
But perhaps D also approximately sums to 33.9?
I don't have the exact options in front of me except A to D with different values. Wait, I do have B. I don't remember C and D now.
Maybe I should reconstruct them.
No, that's unnecessary.
Wait, alternatively, if another option also sums to 33.9, momentum conservation distinguishes them.
But we already know B.
Wait.
Alternatively—
The neutrino energy being 29.8 MeV corresponds to a momentum of $29.8\,\mathrm{MeV}/c$. The muon's kinetic energy using the nonrelativistic formula would then be

$$\frac{p^2}{2m} = \frac{29.8^2}{2(105.7)} = \frac{888.04}{211.4} \approx 4.20\,\mathrm{MeV},$$

not 4.12.
The difference is a relativistic correction.
Wait, that's interesting.
Then perhaps option 4.12 specifically signals that the fully relativistic calculation is expected.
Yes.
Wait, alternatively, if I take $p = 29.8$ exactly, the relativistic muon kinetic energy is

$$\sqrt{29.8^2 + 105.7^2} - 105.7.$$

$$29.8^2 = 888.04.$$

$$105.7^2 = 11172.49.$$

Sum:

$$12060.53.$$

The square root is maybe

$$109.82.$$

Minus 105.7:

$$4.12.$$

Yes.
Wait, but

$$109.82^2$$

is approximately 12060.4.
So fine.
Wait, alternatively, perhaps I should calculate it one more way.
Using γ:

$$\gamma = \frac{E_\mu}{m_\mu} = \frac{109.82}{105.7} \approx 1.039.$$

Then

$$KE = (\gamma - 1)m_\mu \approx 0.039(105.7) \approx 4.12.$$

So again.
Wait.
Alternatively,

$$\beta = \frac{p}{E} = \frac{29.8}{109.82} \approx 0.271.$$

Then

$$\gamma = \frac{1}{\sqrt{1-\beta^2}} \approx \frac{1}{\sqrt{0.9266}} \approx 1.0388.$$

Again.
Wait, alternatively, maybe the neutrino energy should be 29.79, and then β—
Wait.
The question only asks for kinetic energies.
So perhaps I should stop.
Wait, alternatively, there is still the issue that neutrinos have nonzero mass. If the neutrino mass is less than 1 eV, then relative to 29.8 MeV, the correction is of order

$$10^{-15},$$

so irrelevant.
Wait, alternatively, if the neutrino mass were not negligible, then—
Wait.
Maybe the use of "rest mass 139.6 MeV" is technically rest energy, not rest mass. But in natural units.
Wait, alternatively—

$$139.6.$$

$$105.7.$$

Difference:

$$33.9.$$

Ratio.
Momentum.
Wait.
Could there be a factor of c somewhere?
If masses are in MeV, they really mean MeV/c^2.
Then in the formula

$$E_B = \frac{m_A^2 + m_B^2}{2m_A} c^2.$$

But if I plug numerical masses in MeV/c^2, units become

$$\frac{\mathrm{MeV}^2/c^4}{\mathrm{MeV}/c^2} c^2 = \mathrm{MeV}.$$

Yes.
Wait, alternatively, if they are already giving rest energies in MeV, then $c = 1$.
Same result.
Wait.
Alternatively, perhaps the muon is positive, so it is an antimuon, but its mass is the same.
Wait.
Alternatively, perhaps electromagnetic corrections—
Wait.
No.
Wait, alternatively, the pion itself is a pseudoscalar meson, and the decay is weak. Does weak decay change the energy relation?
No.
Wait, alternatively—
The neutrino momentum is fixed.
The muon momentum is fixed.
The energies are fixed.
But perhaps kinetic energy for a massless particle isn't conventionally called kinetic energy because it has no rest frame.
Wait, that's subtle. For massless particles, total energy equals what would effectively be called kinetic energy, but strictly the decomposition

$$E = mc^2 + K$$

doesn't work in the same Newtonian sense because there is no rest frame.
Wait, alternatively, perhaps the exam uses "KE of product particles" loosely.
Then B.
Wait.
Alternatively—
If the kinetic energy of the neutrino is not defined, then the question is ill-posed.
But clearly the options treat it as energy.
Wait.
Maybe the intended answer is not B because of terminology?
No, impossible.
Wait.
Alternatively, perhaps "nu" could be a neutron? Usually neutron is n, not ν.
If it were a neutron, the mass would be 939 MeV, impossible because the parent pionis only 139.6 MeV.
So neutrino.
Wait.
Alternatively, perhaps ν is an antineutrino and—
Wait.
The momentum equation.
I should write it again:

$$0 = p_\mu + p_\nu.$$

Therefore

$$|p_\mu| = |p_\nu|.$$

Energy:

$$m_\pi = \sqrt{p^2 + m_\mu^2} + \sqrt{p^2 + m_\nu^2}.$$

With

$$m_\nu = 0,$$
$$m_\pi = \sqrt{p^2 + m_\mu^2} + p.$$

Then

$$p = \frac{m_\pi^2 - m_\mu^2}{2m_\pi}.$$

Yes.
Wait.
Alternatively, square again:

$$m_\pi - p = \sqrt{p^2 + m_\mu^2}.$$

$$m_\pi^2 - 2m_\pi p + p^2 = p^2 + m_\mu^2.$$

$$m_\pi^2 - m_\mu^2 = 2m_\pi p.$$

Yes.
Wait.
Alternatively—
Maybe there is a way to derive kinetic energy directly from invariant mass:

$$(p_\pi - p_\mu)^2 = 0.$$

Then

$$p_\pi^2 + p_\mu^2 - 2p_\pi \cdot p_\mu = 0.$$

$$m_\pi^2 + m_\mu^2 - 2m_\pi E_\mu = 0.$$

Then

$$E_\mu = \frac{m_\pi^2 + m_\mu^2}{2m_\pi}.$$

Same thing.
Wait.
Alternatively, use four-vectors:

$$p_\pi^\mu = (m_\pi, 0).$$

$$p_\mu^\mu = (E_\mu, p).$$

$$p_\nu^\mu = (p, -p).$$

Wait, if p_ν is $(E, -p)$,

$$E = p.$$

Then conservation.
Yes.
Wait.
Alternatively—
Maybe I should calculate the exact decimal again because the options say 29.8.

$$\frac{8315.67}{279.2}.$$

Let's long