arXiv:2602.01062v1 [cs.AI] 1 Feb 2026

SetPO：集合级 保持多样性的大语言模型推理策略优化

李晨逸 *1 张远 *♠2 王博 3 马国庆 2 唐伟 2 黄昊阳 2 段楠 2

摘要

可验证奖励强化学习在提升大语言模型（LLM）的推理表现方面展现出显著成效，尤其是在数学任务中。然而，这类改进往往伴随着结果多样性的降低，即模型将概率质量集中在一小部分解法上。受边际收益递减原则的启发，我们提出一种基于核相似度、定义在采样轨迹上的集合级多样性目标函数。我们的方法为每条采样轨迹计算留一边际贡献，并将该目标函数作为即插即用的优势塑形项融入策略优化。我们还在分布扰动框架下研究了单条轨迹对语言模型多样性的贡献。理论分析验证了单调性：越罕见的轨迹，对全局多样性的边际贡献始终越高。在不同模型规模的一系列实验中，我们提出的算法表现出良好效果，在多个基准测试的 Pass@1 和 Pass@K 指标上均持续优于强基线方法。

1. 引言

强化学习（RL）领域近期迅速扩展，在数学推理和代码生成方面尤其取得了显著进展。这一进展主要由可验证奖励强化学习（RLVR）推动，其可扩展性支持大规模的有效训练，从而提升了大语言模型（Ouyang 等人，2022；Guo 等人，2025）的推理能力。在这一范式中，组相对策略优化（GRPO）（Shao 等人，2024）作为近端策略优化（PPO）（Schulman 等人，2017）的有竞争力替代方案应运而生，在无需评论家模型的情况下取得了出色的实证表现。这一设计选择启发了一系列后续研究，它们在 GRPO 的基础上，提升了不同基准测试中的训练稳定性和实证性能（Yu 等人，2025；Zheng 等人，2025；Zhao 等人，2025）。

近期研究主要致力于开发更有效的基于组的策略优化算法，而另一类研究则发现，通过强化学习获得的表面性能提升往往以模型多样性降低为代价。具体而言，近期证据表明，经可验证奖励强化学习微调的模型生成的许多轨迹，也可以通过增加采样预算从基础模型中恢复，这意味着在大规模 rollout 下，基础模型可能构成一个有效的上限（Yue等人，2025）。同样，LLM领域的同期研究也发现，策略性能的提升往往伴随着策略熵下降，这表明性能增强可能源于模型的输出分布集中到更狭窄的解法集合上（Cui等人，2025）。

这些观察结果凸显了有效探索是基于 RLVR 的后训练中缺失的一项关键要素；在提升推理性能的同时维持多样化结果，离不开这一点。然而，与状态-动作空间有限的传统强化学习不同，大语言模型的策略拥有庞大的词汇量，因此在高维空间中运行。由此，搜索空间会随序列长度呈指数级增长，给探索过程带来沉重负担（Gupta 等人，2024）。现有解法通常依赖基于熵的正则化，将其作为强化学习中探索的替代指标（Yao 等人，2025；Cheng 等人，2025）。然而，当熵在词元级定义并沿序列聚合时，可能会造成显著的长度偏差。较长的响应自然会因词元级不确定性增加而累积更高的熵，从而偏向冗长的输出，而不论它们是否真正提供了多样化的解法。此外，词元级熵无法捕捉轨迹级的语义多样性，因为它衡量的是下一个词元预测中的局部不确定性，而非完整输出含义的整体变化。另外，近期一些研究明确以优化 Pass@K 为目标（Chen 等人，2025b；Tang 等人，2025；Walder

*共同贡献，♠项目负责人 [1]北京大学 [2]京东探索研究院，中国 [3]北京理工大学。李晨逸 <lichenyi@stu.pku.edu.cn>，张远< zhangyuan.430@jd.com>，段楠 <duannan@jd.com>。

预印本。*2026*年*2*月*3*日。

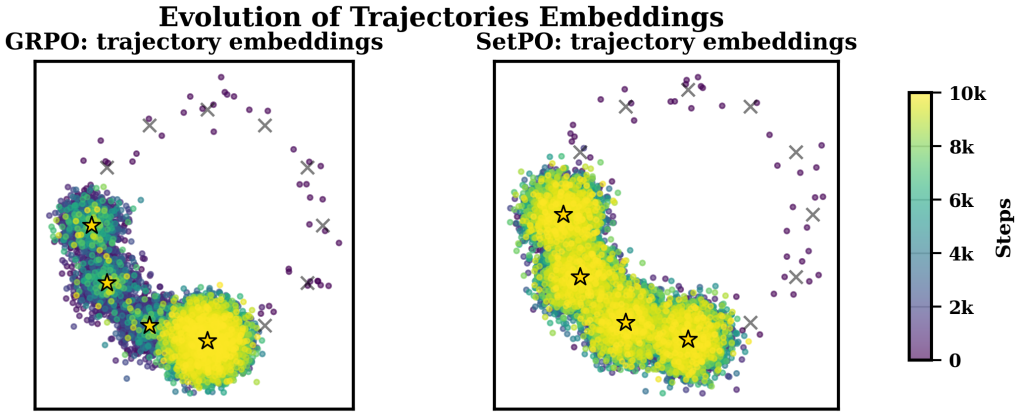

图 *1.* 训练过程中 GRPO 与 SetPO 的轨迹嵌入演化（星号表示正确模式；颜色表示训练步数）。左图（GRPO）：尽管轨迹最初分布于多个模式，但训练逐渐集中到单一主导簇，呈现出明显的模式坍塌。右图（SetPO）：嵌入在整个训练过程中始终分布在多个簇上，并在后期训练步数中继续覆盖正确模式。这表明 SetPO 缓解了模式坍塌，并保持了 rollout 多样性。

& Karkhanis，2025）。然而，由于 Pass@K 主要奖励重复采样下的成功，这类改进未必能转化为单次采样正确率的提升，因而 Pass@1 的收益有限，甚至可能导致性能下降。

在本文中，我们提出 SetPO，一种保持多样性的策略优化算法，促使所学策略表征多模态结果分布。SetPO 为每条轨迹分配集合级边际多样性得分，从而为策略优化提供显式的多样性感知信号。具体而言，对于每个提示，我们采样一组 rollout 样本，并使用预训练嵌入模型计算成对语义相似度。基于这些相似度，我们量化每条轨迹在组内的冗余程度，并通过留一法边际估计器定义其多样性贡献。覆盖不同语义模式的轨迹获得更高得分，而冗余轨迹获得较低得分。随后，我们将这一集合级多样性信号纳入标准奖励最大化，即用加权多样性项增强每条轨迹的训练优势。为验证所提算法的有效性，我们构建了一个包含十二个离散语义模式的玩具多模态老虎机环境，其中仅前四个模式产生正奖励。每个模式由一个 50 维嵌入表示，语义相似度通过余弦相似度衡量。如图 1所示，我们的方法在整个训练过程中比原始 GRPO 保留了更多不同的模式。因此，它缓解了模式坍塌，并产生了更加多样化的结果集合。

我们的主要贡献如下。

- 据我们所知，SetPO 首次引入了明确的集合级多样性目标函数。与传统的成对方法不同，我们的方法采用分布视角：通过汇总语义相似度来评估整个采样集合的整体多样性，而非聚焦于孤立的局部差异。

- 我们将留一边际贡献机制融入策略优化，根据每条轨迹对集合多样性的影响为其分配奖励。我们的理论分析揭示了这一估计量的反冗余特性，表明它会为较稀有样本赋予更高的边际值。

- 我们在涵盖不同模型规模（从 1.5B 到 32B）的综合推理基准测试上评估了所提出的 SetPO 算法。实证结果表明，与强基线方法相比，我们的方法始终表现更优。

2. 相关工作

用于大语言模型推理的强化学习。已有一系列研究（Zhao 等人，2025；Liu 等人，2025；Shrivastava 等人，2025）旨在改进 RLVR 的 GRPO 式优化。DAPO（Yu 等人，2025）引入了若干改动，以提升训练稳定性和样本效率。具体而言，它采用非对称裁剪，使用不同的裁剪比率 ϵ_{low} 和ϵ_{high}，使低概率词元也能获得有效更新。此外，DAPO 会过滤掉所有采样轨迹均正确或均错误的提示，从而将优化重点放在能够提供有意义学习信号的样本上。另外，GSPO（Zheng 等人，2025）采用序列级奖励形式，而非 GRPO 使用的词元级目标函数，从而提升大语言模型的训练稳定性，尤其是在专家混合设置中。其他方法表明，经过适当调优的 REINFORCE 式策略梯度方法也能取得有竞争力的性能，并实现稳定训—

练（Chu 等人，2025；Ahmadian 等人，2024）。例如，REINFORCE++ （Hu，2025）引入全局优势归一化，以提升原始策略梯度方法的训练稳定性。ReMax（Li 等人，2023）通过引入减法基线来改进 REINFORCE 梯度估计器，以降低方差并提升训练稳定性。具体而言，该基线通过贪心采样生成一个响应，并评估其对应的奖励来计算。

强化学习中的多样性。 探索对于提升高维强化学习的性能至关重要（Ladosz 等人，2022）。先前的研究（Yao 等人，2025；Cheng 等人，2025；Wang 等人，2025）主要采用基于熵的正则化来平衡利用与探索。这些方法将熵视为多样性的代理指标，通过最大化大语言模型的平均输出熵，奖励具有更高词元级不确定性的轨迹，以鼓励生成更多样化的结果。（Masood 和 Doshi-Velez，2019）则提出利用最大均值差异距离作为轨迹级正则项，以提升学习到的策略的多样性。近期研究（Lanchantin 等人，2025；Chung 等人，2025；Ismayilzada 等人，2025）则在偏好优化中明确兼顾质量与多样性。

同期研究（Chen 等人，2025a；Li 等人，2025）也探索了用于大语言模型强化学习微调的轨迹级语义多样性度量。尽管这些方法同样依赖轨迹级语义相似度来量化多样性，SetPO 却采用了截然不同的形式化方法。我们提出一种集合级目标函数，将生成的批次作为整体分布进行评估。通过采用留一边际计算，我们明确衡量每条轨迹对整个组信息量的贡献，从而从全局分布视角优化多样性，而非采用先前研究探索的成对分离方法。

3. 多样性感知奖励计算

3.1. 预备知识：基于组的策略优化

基于组的策略优化已成为语言模型对齐中标准近端策略优化的一种稳定且高效的替代方案，其主要优势在于无需单独的评论家网络。该类方法中的代表性方法之一是 GRPO。

形式上，对于从数据集 $\mathcal{D}$中采样的每个提示 x，GRPO从当前策略 $\pi_{\theta_{\text{old}}}$生成一组 G 个输出 $\{o_1, \ldots, o_G\}$。记$\rho_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t}|x,a_{i,<t})}{\pi_{\theta_{\text{old}}}(a_{i,t}|x,a_{i,<t})}$为重要性采样比率，而 $\mathbb{D}_{\text{KL}}$ 是使策略保持接近参考模型π_{ref}的KL散度惩罚。奖励以标量分数 r_i 分配给每个轨迹 o_i 由模型给出。为了在没有学习到的价值函数的情况下降低方差，GRPO通过相对于组统计量对奖励进行归一化，为每个输出计算优势 A_i：$\bar{A}_i = \frac{r_i - \mu}{(\{r_j\}_{j=1}^G)\ \sigma(\{r_j\}_{j=1}^G)+\epsilon}$，其中 μ 和 σ 表示组内奖励的均值和标准差。策略参数 θ 通过最大化以下代理目标进行更新：

$$J(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}(\theta)\,\bar{A}_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\,\bar{A}_i\Big) - \beta\,\text{KL}\big(\pi_\theta(\cdot \mid x, a_{i,<t}) \,\big\|\, \pi_{\text{ref}}(\cdot \mid x, a_{i,<t})\big)\Bigg]. \tag{1}$$

3.2. 语言模型的多样性

我们首先研究轨迹层面上的语言模型多样性度量。给定提示 x，轨迹定义为有限词元序列 $y = (a_1, \ldots\ldots, a_T) \in \Omega$，其中 a_T 是 *EOS* 词元，或者 T 达到最大长度。因此，参数化语言模型会诱导出如下轨迹分布：$P_\theta(y|x) = \prod_{t=1}^{T} \pi_\theta(a_t|x, a_{<t})$。在以下讨论中，在含义明确时，我们省略下标 θ。

不同于关注词元级多样性的方法，我们从语义层面评估完整输出。我们采用有界对称相似性核 $k: \Omega \times \Omega \to [0,1]$，满足 $k(y,y') = k(y',y)$ 且 $k(y,y) = 1$。具体而言，我们令 $k(y,y') = \text{sim}(e(y)\ e(y'))$，并采用预先定义的语义嵌入函数 $e(\cdot): \Omega \to \mathbb{R}^n$，其中 n表示嵌入空间的维度。

对于任意轨迹分布 $P \in \mathcal{P}(\Omega)$，我们将单条轨迹 y 的核化局部质量定义为：

$$m_P(y) := \mathbb{E}_{y' \sim P}\left[k(y,y')\right].$$

从概念上讲，$m_P(y)$估计的是 y周围的语义密度。值较高表示 y 位于拥挤邻域中，意味着存在冗余。相反，值较低则表示 y 位于稀疏区域，意味着相对新颖。随后，我们通过汇总这些局部估计来量化语言模型的整体多样性。设 $g: [0,1] \to \mathbb{R}$ 为连续非递增函数。多样性目标定义如下：

$$\mathcal{F}(P) := \mathbb{E}_{y \sim P}\Big[g\big(m_P(y)\big)\Big]. \tag{2}$$

由于 g 是非递增的，较小的局部质量 $m_P(y)$会产生更大的回报。该机制有效地奖励语义新颖性。在本文中，我们采用如下

塑形函数： $g(x) = -\log(1+x)$. 除数值稳定性之外，该形式体现了边际敏感性递减原则。当 $x \approx 0$ 时，梯度幅值最为陡峭，并随 x 增大而衰减。因此，该函数优先保留高度独特的轨迹，因为对已经拥挤的区域进行调整所获得的奖励会逐渐减小。

3.3. 单条轨迹的留一边际

现在，我们从通过对任意子集进行蒙特卡洛近似来评估语言模型多样性的角度，推导逐样本边际分配。设 $S \subset \Omega$ 是从 $P_\theta(\cdot \mid x)$ 中抽取的任意有限个独立同分布轨迹，其基数为 $|S| \geq 3$。由于 S 是轨迹空间的蒙特卡洛样本，我们可以使用广义组内得分 $D(S)$ 来估计总体多样性 $\mathcal{F}(P)$:

$$D(S) := \frac{1}{|S|} \sum_{y \in S} g\big(\widehat{m}(y; S)\big), \tag{3}$$

其中 $\widehat{m}(y;\ S) := \frac{1}{|S|-1} \sum_{z \in S \setminus \{y\}} k(y, z)$。这里， $\widehat{m}(y;\ S)$ 根据 S 的经验分布估计 y 周围的局部语义密度。由于样本是独立同分布的，对于任意有效子集 S，$D(S)$ 始终是一致估计量。

这一统计性质使我们能够通过比较完整生成批次 $\Omega = \{o_1, \ldots, o_G\}$ 的多样性估计与子集 $\Omega \setminus \{o_i\}$的多样性估计，量化单条轨迹 o_i 的贡献。如果 o_i 具有信息量，那么与不包含它的子集相比，它的存在应能提高近似的分辨率。因此，我们将每条轨迹 o_i 对所考虑的整个集合的多样性贡献 s_i 定义为留一边际差值，具体如下:

$$s_i := D(\Omega) - D\big(\Omega \setminus \{o_i\}\big). \tag{4}$$

在这一表述中， $D(\Omega)$利用所有 G 样本来估计核化密度，而 $D(\Omega \setminus \{o_i\})$则使用其余 $G-1$ 条轨迹重新计算得分。具体而言，缩减集合对应的项计算如下:

$$D(\Omega \setminus \{o_i\}) = \frac{1}{G-1} \sum_{j \neq i} g\big(\widehat{m}(o_j; \Omega \setminus \{o_i\})\big).$$

较大的 s_i 表明，纳入 o_i 会使集合的估计多样性显著提升，这意味着 o_i 捕捉到了其他样本未涵盖的独特语义模式。相反，较小或为负的 s_i 则表明存在冗余。由于强化学习中的组大小 G 通常较小，计算这些递归留一边际量所需的计算开销可以忽略不计。

3.4. 集合级策略优化方法

我们将上述针对特定集合得出的每个样本的边际多样性贡献，作为即插即用塑形项，整合到前文预备知识中介绍的组策略优化框架中。我们的方法仅需标准的一组采样轨迹 $\{o_i\}_{i=1}^G$ 和 $\bar{}$ 基础优势 A_i。

我们根据同一组计算多样性得分 s_i，并构造增广优势:

$$\hat{A}_i := \bar{A}_i + \lambda\, s_i,$$

其中 $\lambda \geq 0$ 用于调节任务表现与多样性之间的权衡。随后，我们将这一增广优势直接代入式（1）中定义的标准 GRPO 目标函数，以替换原有的进制优势。多样性插入项$\bar{}$仅修改优势项 A_i。这种模块化设计确保我们的多样性感知奖励计算适用于广泛的集合级强化学习目标函数，包括 DAPO、GSPO 等，只要它们依赖于基于组的优势估计即可。

4. 理论分析

4.1. 总体视角

为了理解多样性 $\mathcal{F}(P)$ 朝任意单条轨迹的变化，我们从扰动的角度进行考察。首先，我们给出关于核函数和塑形函数的假设如下。

假设 4.1. 核函数 $k(\cdot,\cdot)$ 是可测且对称的，即 $k(x,y) = k(y,x)$。此外，对所有 x, y，均有 $0 \leq k(x,y) \leq 1$ 。塑形函数g：$[0,1] \to \mathbb{R}$ 在 $[0,1]$ 上连续可微，且 g'有界。此外， g 在 $[0,1]$上非递增。

考虑混合更新 $P_\varepsilon = (1-\varepsilon)P + \varepsilon\delta_\tau$，它可以视为单条轨迹 τ 对原始分布的扰动。我们通过盖托导数刻画 τ 的影响。

定理 4.2 (多样性泛函的影响函数)。在假设4.1下， $\mathcal{F}$ 在 P 沿方向 τ 的盖托导数为

$$\mathcal{I}(\tau; P) := \frac{d}{d\varepsilon}\mathcal{F}(P_\varepsilon)\Big|_{\varepsilon=0} = \underbrace{g(m_P(\tau))}_{\textit{intrinsic novelty}} + \underbrace{\mathbb{E}_{y\sim P}[g'(m_P(y))\, k(y,\tau)]}_{\textit{interaction penalty}} - \Psi(P),$$

其中， $\Psi(P)$ 不依赖于 τ。

此处省略具体的技术推导，所有证明均推迟至附录C。影响分为两种

表 1. Pass@1 准确　率，针对 Qwen2.5-Math-7B 的数学基准测试。标有 * 的条目来自原始论文　　er.

方法	GSM8K	MATH500	CollegeMath	AMC23	AIME24	AIME25	Avg
Qwen2.5-Math-7B	53.4	48.1	19.4	41.3	9.9	4.0	29.4
R1-zero-Div*	90.6	76.9	47.5	-	-	-	-
GRPO	89.1	73.5	42.2	53.5	15.4	9.7	47.2
GSPO	89.5	73.0	45.0	57.7	17.4	7.9	48.4
DAPO	92.2	77.6	47.1	59.6	21.7	11.0	51.7
SetPO+GRPO	92.2 (+3.1)	**80.8** (+7.3)	48.3 (+6.1)	60.5 (+7.0)	21.6 (+8.2)	**13.6** (+3.9)	52.8 (+5.6)
SetPO+GSPO	91.6 (+2.1)	75.7 (+2.7)	47.2 (+2.2)	60.9 (+3.2)	21.1 (+3.7)	10.1 (+2.2)	51.1 (+2.7)
SetPO+DAPO	**93.0** (+0.8)	79.2 (+1.6)	**49.5** (+2.2)	**62.3** (+2.7)	**25.3** (+3.6)	13.5 (+2.5)	**53.8** (+2.1)

互补机制：（i）一个内在新颖性项$g(m_P(\tau))$，由于 g 单调不增，因此会严格倾向于较低的局部质量；（ii）一个交互惩罚，用于抑制与现有轨迹高度相似的轨迹，因为 $g'(m_P(y)) < 0$，且核权重 $k(y, \tau)$会将惩罚集中在邻近区域。因此，$\mathcal{F}(P)$遵循边际收益递减原则，并能避免冗余。

4.2. 留一法作为边际贡献

我们的算法在有限候选集上运行，并反复移除对多样性损害最小的元素。需要强调的是，留一法（LOO）得分（4）并非启发式代理，而是移除 o_i 所导致的目标函数边际下降量。以下结果将对此进行精确说明，并表明 s_i 会系统性地惩罚冗余。对于有限集合 Ω，我们将 o_i 的边际贡献刻画如下。

定理 4.3 (LOO 贡献的精确分解)。 对于 $|\Omega| \geq 3$， s_i 可精确分解为三个部分：

$$s_i = \underbrace{\frac{1}{|\Omega|} g(m_i)}_{\mathcal{T}_{self}} + \underbrace{\left(\frac{1}{|\Omega|} - \frac{1}{|\Omega|-1}\right) \sum_{j \neq i} g(m_j)}_{\mathcal{T}_{norm}} + \underbrace{\frac{1}{|\Omega|-1} \sum_{j \neq i} \left[g(m_j) - g(m_j^{(-i)}) \right]}_{\mathcal{T}_{interaction}},$$

其中 $m_j^{(-i)} = \widehat{m}(o_j; \Omega \setminus \{o_i\})$。

具体的技术推导在此省略，所有证明和完整分析请参见附录 D。此外，每个交互求和项仅通过成对相似度 o_i 依赖于 $k(o_j, o_i)$。

这一分解说明了为何 s_i 是一个稳健的信号。第一项 $\mathcal{T}_{\text{self}}$ 会奖励那些在多样性塑造函数 g 下各自具有新颖性的点。与此同时，

交互项 $\mathcal{T}_{\text{interaction}}$ 衡量了o_i 对邻域结构造成的拥挤程度。移除冗余点通常会降低其邻居的局部质量，从而提高它们的 $g(\cdot)$ 值。这会使 $D(\Omega \setminus \{o_i\})$ 增大，进而得到一个更小的 s_i。这一机制恰好捕捉了留一法构造所期望的收益递减行为。

此外，我们考察当前样本与其他样本之间每个相似度对应的边际贡献。

定理 4.4(s_i的严格反冗余性)。 对于任意 $t \neq i$，若假设 4.1 成立，则留一法贡献随成对相似度严格递减 $k(o_i, o_t)$：

$$\frac{\partial s_i}{\partial k(o_i, o_t)} = \frac{1}{|\Omega|(|\Omega|-1)} \Big(g'(m_i) + g'(m_t) \Big) < 0.$$

在我们的目标函数下，与任何其他元素的相似度增加，都必须降低 o_i 的边际贡献。因此，迭代移除最小-s_i 元素可证明是一种冗余移除过程。最后，我们确立一个全局排序保证。如果某个点与集合其余部分的相似度处处更低，那么留一法必然会将其排在更高位置。

定理 4.5 (支配排序)。 设 $o_r, o_c \in \Omega$。若假设 4.1 成立，且对所有 $k(o_r, o_j) \leq k(o_c, o_j)$ 均有 $j \notin \{r, c\}$，且至少一个此类 j满足严格不等式，则有 $s_r > s_c$。

总之，定理 4.2 量化了注入单条轨迹如何通过其影响扰动总体多样性泛函。当轨迹位于低质量区域时，它会受到青睐；而根据它与现有分布之间相似性诱导的重叠，它会受到惩罚。在有限候选集上，定理 4.3–4.5 进一步确立，留一边际 s_i（定义为移除 o_i 时同一目标函数的精确边际下降量）会随每个成对相似度 $k(o_i, o_t)$严格递减。因此， s_i 提供了一种有原则的去冗余排序信号，其优化与提升

表 2。 Pass@1 准确率在 Qwen2.5-Math-1.5B 数学基准测试中的表现。标有 * 的条目来自原始论文。

方法	GSM8K	MATH500	CollegeMath	AMC23	AIME24	AIME25	Avg
Qwen2.5-Math-1.5B	39.4	36.4	6.6	29.5	4.6	1.8	19.7
R1-zero-Div*	83.2	70.4	43.9	-	-	-	-
GRPO	83.0	67.2	43.1	47.5	11.5	8.2	43.4
GSPO	83.5	67.2	43.0	48.8	11.1	5.6	43.2
DAPO	85.9	69.5	44.3	50.1	12.0	6.6	44.7
SetPO+GRPO	**86.9** (+3.9)	**72.9** (+5.7)	45.7 (+2.6)	51.3 (+3.8)	**13.4** (+1.9)	**9.7** (+1.5)	**46.7** (+3.3)
SetPO+GSPO	85.0 (+1.5)	69.4 (+2.2)	44.7 (+1.7)	50.6 (+1.8)	13.2 (+2.1)	7.4 (+1.8)	45.2 (+2.0)
SetPO+DAPO	86.6 (+0.7)	71.7 (+2.2)	**46.5** (+2.2)	**53.2** (+3.1)	**13.4** (+1.4)	8.9 (+2.3)	**46.7** (+2.0)

诱导经验分布下的多样性目标函数，而非充当一种启发式方法。

5. 数值实验

5.1. 一般数学推理

骨干模型。 我们采用两个不同规模的指令微调骨干模型，以验证我们方法的泛化能力：Qwen2.5-Math-7B、Qwen2.5-Math-1.5B（Yang 等人，2024）。我们的实现基于开源的 veRL 框架（Sheng 等人，2024），以确保训练高效且 rollout 管理灵活。

基线方法。我们将提出的 SetPO 方法+GRPO、SetPO+GSPO 和 SetPO+DAPO 与四种具有代表性的强化学习基线方法进行比较：GRPO（Shao 等人，2024）、GSPO（Zheng 等人，2025）、DAPO（Yu 等人，2025）和 R1-Zero-Div（Yao 等人，2025）。为确保比较严格且公平，我们对所有方法采用完全一致的实验流程；具体而言，我们使用相同的超参数，包括学习率、批量大小和总训练步数。

训练配置。所有模型均使用标准的 GSM8K 训练集划分进行训练（Cobbe 等人，2021）。我们使用 AdamW 优化器，以恒定学习率 3×10^{-6} 训练模型 2 个 epoch，全局批次大小为 64。在在线数据生成阶段，我们将每个提示的 rollout 次数设为 6，并采用 0.9 的采样温度，以鼓励在解空间中进行多样化探索。更多具体的实验设置见附录E。

评估配置。 我们在一系列全面的数学推理基准测试上评估性能，包括GSM8K测试、MATH500（Hendrycks 等人，2021）、奥林匹克竞赛（He 等人，2024）、大学数学基准（Tang 等人，2024）、AMC 2023（美国数学协会，2023），以及 AIME 24/25（美国数学协会，2025）。为确保统计稳健性，所有报告结果均取自五个随机种子的平均值。除非另有说明，我们报告基于 Avg@64 得出的 Pass@1 准确率（即对 64 个采样生成结果取平均），推理时温度设为 0.5。对于大学数学基准，由于其评估集规模显著更大，为提高效率，我们采用 Avg@16。

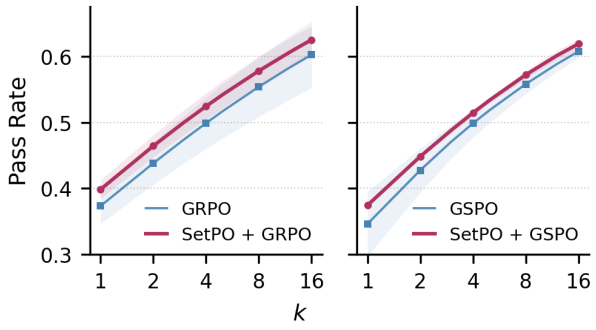

图 2. SetPO+GRPO 和 SetPO+GSPO 在奥林匹克竞赛基准上的表现。阴影区域表示多次运行之间的方差。两种方法都提升了性能并降低了变异性，表明 SetPO 能稳定训练动态。

数值结果。表1 和2分别总结了 7B 和 1.5B模型的 Pass@1 结果。总体而言，在不同模型规模和基准测试难度下，SetPO 相较于基线方法取得了显著提升。具体来说，SetPO 在不同基准测试和模型规模下，始终优于 GRPO、GSPO 和 DAPO。如表1所示，SetPO+DAPO 的整体表现最佳，平均 Pass@1 准确率为 53.8%。值得注意的是，将 SetPO 应用于 GRPO 时提升最为显著，平均准确率从 47.2% 提高到 52.8%（+5.6）。在具有挑战性的基准测试中，SetPO 带来了显著的相对提升；例如，在 AIME24 上，准确率从 15.7% 提高到 21.6%，相当于

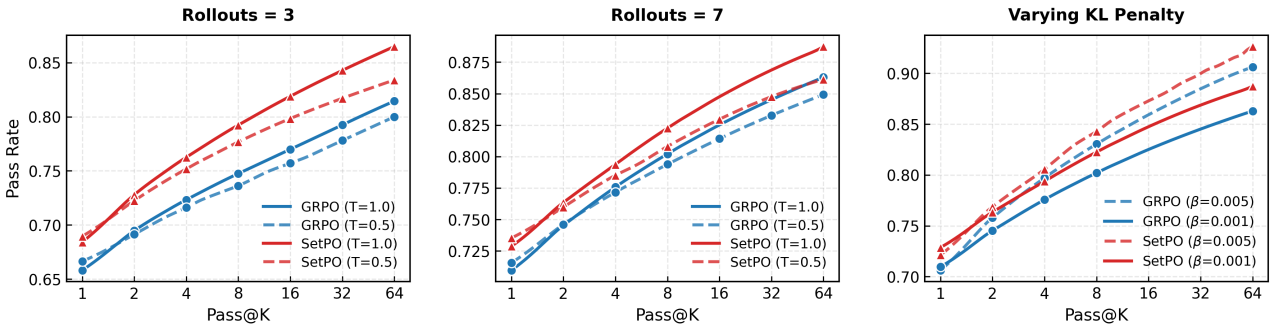

图*3*。 在不同解码温度、训练展开次数和KL惩罚设置下，SetPO（SetPO+GRPO）与GRPO基线方法在Countdown任务上的Pass@K表现。SetPO在这些设置下始终取得更高的Pass@K。

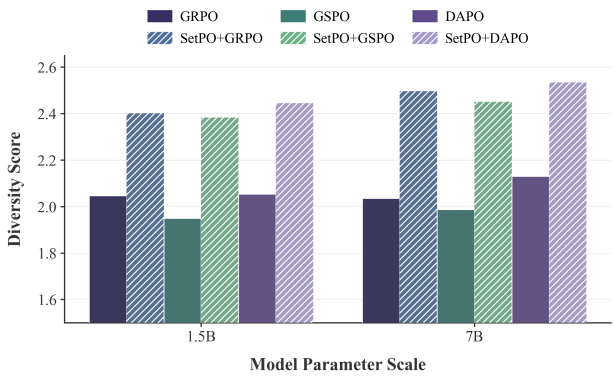

图*4*。 AIME24基准上的多样性得分，根据Gemini-2.5-Flash生成的16个解法逐题计算。得分范围为1（多样性最低）至5（多样性最高）。

相对提升超过 40%。表2进一步表明，1.5B模型也呈现出类似趋势，其中 SetPO+DAPO 的整体平均准确率最高，达到 46.7%。此外，如图2所示，我们的方法显著降低了基线方法的方差，并在不同随机种子下持续带来性能提升。

多样性评估。 为了评估生成答案的多样性，我们使用Gemini-2.5-Flash作为大语言模型评估器。评分采用五分制，从多样性最低的1分到最高的5分。完整的评估提示见附录H。图 4 显示，在1.5B和7B模型规模下，SetPO增强方法的多样性得分始终高于基线方法，表明我们的算法有效提升了最终输出的多样性。

5.2. Countdown评估

任务描述。 Countdown任务是对算术推理和约束满足能力的严格基准测试。在此任务中，模型会获得一个目标数字和一组输入数字。目标是生成一个由输入数字和基本算术运算构成、计算结果等于目标数字的有效数学表达式。此任务考验模型在离散动作空间中进行搜索并验证自身推理步骤的能力。

模型与训练。 我们使用 Qwen2.5-3B 作为骨干模型，在倒计时数据集（Pang等人，2023）上开展实验。在训练过程中，我们将所有模型的全局批次大小设置为 128，最大生成长度设置为 2048。为严格遵循所需的答案格式，我们的基于规则的奖励函数加入了权重为 0.1 的辅助格式奖励项。我们探索了不同的超参数设置，具体考察了 $\{3, 7\}$ 的展开次数以及 $\{0.001, 0.005\}$ 的 KL散度系数。详细配置见附录 E。

评估协议。 在推理过程中，我们报告 $k \in \{1, \ldots, 64\}$ 的无偏 Pass@k 指标。为全面分析模型的行为，我们通过在 $T = 0.5$ 和 $T = 1.0$ 下评估性能，研究采样温度的影响，并讨论样本量变化如何影响实证结论。

性能优势。 如图3所示，在 Countdown任务上，SetPO 的表现始终优于 GRPO 基线，展现出更高的样本效率和解答质量。在 3 轮和 7 轮 rollout 设置下，这一性能优势都保持稳定。值得注意的是，随着 k 增大，性能差距也随之扩大，这表明与基线相比，SetPO 尤其擅长覆盖更多奖励模式。此外，我们的方法在多种采样配置下都保持着明显优势。较高的温度通常会促进更广泛的探索，从而提升两种方法的通过率，但在温度为0。5 和 1。0时，SetPO 的表现始终优于 GRPO。最后，图3右侧面板所示的不同 KL惩罚分析进一步验证了 SetPO 的稳定性，结果表明它

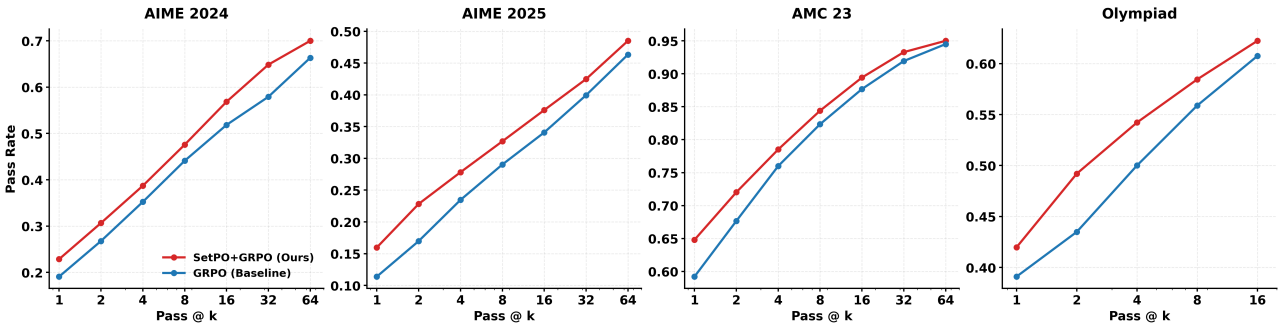

图*5*。 Qwen2.5-32B 上 SetPO+GRPO 与 GRPO 基线的 Pass@k 性能。对于所有 k，SetPO+GRPO 均持续优于 GRPO，表明其在不同采样预算下都能稳定提升性能。

表*3*。 Countdown任务的多样性统计。多样性宽度表示至少能产生两种不同输出的问题数量。平均模态数表示每个问题的不同答案数量的均值。采样温度设为 1.0。

设置	多样性宽度	平均值 Mode
rollout 7 SetPO+GRPO	**415**	**3.5**
rollout 7 GRPO	351	2.3
rollout 3 SetPO+GRPO	**403**	**3.2**
rollout 3 GRPO	321	2.1

无论正则化强度如何，都能稳健地提升性能，优于基线。

多样性分析。 为系统地评估输出的多样性，我们分析表3中呈现的统计数据。我们引入两个主要指标：（1）多样性宽度，定义为模型生成至少两个不同正确解的样本数量；（2）平均模态数，计算在被识别为具有多样性的实例中，每个实例的唯一有效答案数量的平均值。

实证结果表明，与 GRPO 基线相比，整合 SetPO 可显著扩展解空间。在 7-rollout 设置下，SetPO+GRPO 将多样性宽度从 351 提升至 415，表明其具备发现替代推理路径的强大能力。此外，平均模态数从 2.3 上升至 3.5。这一量化改进证实，SetPO+GRPO 能够主动促进对更广泛有效解法的覆盖。

5.3. 扩展至 32B模型

实验设置。 我们将实验扩展至 32B 参数规模，使用 Qwen2.5-32B（Team等人，2024）作为骨干模型，进一步验证了我们方法的可扩展性。模型在 Simple-RL Zoo（Zeng等人，2025）的高难度子集上进行微调。为确保优化稳定且在这一规模下有效探索，我们将全局批次大小设为 1024，PPO小批次大小设为 256。在推理 rollout 阶段，我们以 0.1 的采样温度为每个提示生成 6 条轨迹，以兼顾多样性与准确性。

数值性能。 图 5验证了 SetPO 在 32B 参数模型上的可扩展性：在全部四个数学基准测试中，SetPO 均持续优于 GRPO 基线。具体而言，SetPO 在单样本和多样本设置下都展现出稳定优势。在 Pass@1 设置中，我们的方法相比基线取得了明显的准确率提升，尤其是在更具挑战性的 AIME 2025 和奥林匹克竞赛任务上，这表明其推理质量更高。在 AIME24 和 AIME25 等高难度基准测试中，与原始 GRPO 相比，SetPO 在 Pass@1 上几乎提升了 5 个百分点。此外，随着采样预算 k 增加，SetPO 仍持续保持这一性能差距。这些结果表明，SetPO 不仅提升了单个解法的正确性，还增强了复杂推理任务中生成分布的多样性与稳健性。

6. 结论

我们提出了 SetPO，这是一种集合级策略优化方法，通过依据轨迹在滚动生成集合中的留一边际贡献给予奖励，来促进输出的多样性。SetPO 估计由语义轨迹相似度诱导的局部核密度，并将每条轨迹受密度影响的边际效应转化为可解释的留一法多样性奖励，可将其纳入标准的基于组的策略优化中。我们分析了多样性泛函在分布扰动下的性质，给出了相应界限，并确立了一种单调机制，表明较稀有模态能够带来更大的多样性增益。在数学推理和组合任务上针对多种模型规模开展的大量实验表明，SetPO 始终优于强基线方法，在保持解多样性的同时取得了更高的准确率。

影响声明

本文旨在推动强化学习领域的发展。我们的工作可能会带来许多潜在的社会影响，但我们认为没有哪一项需要在此特别强调。

参考文献

Ahmadian, A., Cremer, C., Gall´e, M., Fadaee, M., Kreutzer, J., Pietquin, O., Ust¨un, A., and Hooker, S. 回归基础：重新审视用于从人类反馈中学习的大语言模型强化学习风格优化。*arXiv* 预印本 *arXiv:2402.14740*，2024。

Chen, M. 对基于代码训练的大语言模型进行评估。*arXiv* 预印本 *arXiv:2107.03374*，2021。

Chen, M., Chen, G., Wang, W., and Yang, Y. Seed-GRPO：通过语义熵增强的 GRPO 实现不确定性感知策略优化。*arXiv* 预印本 *arXiv:2505.12346*，2025a。

Chen, Z., Qin, X., Wu, Y., Ling, Y., Ye, Q., Zhao, W. X., 和 Shi, G. 面向大型推理模型、用于自适应平衡探索与利用的 Pass@ k 训练。*arXiv* 预印本 *arXiv:2508.10751*，2025b。

Cheng, D., Huang, S., Zhu, X., Dai, B., Zhao, W. X., Zhang, Z., 和 Wei, F. 通过探索进行推理：强化学习应用于大语言模型的熵视角，2025。*arXiv* 预印本 *arXiv:2506.14758*，2025。

Chu, X., Huang, H., Zhang, X., Wei, F., 和 Wang, Y. GPG：一种用于模型推理的简单而强大的强化学习基线。*arXiv* 预印本 *arXiv:2504.02546*，2025。

Chung, J. J. Y., Padmakumar, V., Roemmele, M., Sun, Y., 和 Kreminski, M. 修改大型语言模型的后训练，以实现多样化的创意写作。*arXiv* 预印本 *arXiv:2503.17126*，2025。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. 和 Schulman, J. 训练验证器来解答数学文字题，2021 年。网址https://arxiv.org/abs/2110.14168。

Cui, G., Zhang, Y., Chen, J., Yuan, L., Wang, Z., Zuo, Y., Li, H., Fan, Y., Chen, H., Chen, W. 等。推理语言模型强化学习的熵机制。*arXiv* 预印本 *arXiv:2505.22617*，2025 年。

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X. 等。DeepSeek-R1 通过强化学习激励大语言模型进行推理。《自然》，645(8081):633–638，2025。

Gupta, N., Narasimhan, H., Jitkrittum, W., Rawat, A. S., Menon, A. K. 和 Kumar, S. 语言模型级联：词元级不确定性及其他。*arXiv* 预印本 *arXiv:2404.10136*，2024。He, C., Luo, R., Bai, Y., Hu, S., Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y. 等。奥林匹克竞赛基准：一项具有挑战性的基准，旨在通过奥林匹克竞赛级别的双语多模态科学问题推动通用人工智能发展。载于计算语言学协会第*62*届年会论文集（第*1*卷：长篇论文），第3828–3850页，2024。

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D. 和 Steinhardt, J. 使用数学数据集衡量数学问题求解能力。*arXiv* 预印本 *arXiv:2103.03874*，2021 年。

Hu, J. Reinforce++：一种用于对齐大语言模型的简单高效方法。*arXiv* 预印本 *arXiv:2501.03262*，2025。

Ismayilzada, M., Laverghetta, A., Luchini, S. A., Patel, R., Bosselut, A., Van Der Plas, L. 和 Beaty, R. E. 创意偏好优化。计算语言学协会：*2025*年自然语言处理实证方法会议论文集，第 9580–9609 页，2025。

Ladosz, P., Weng, L., Kim, M. 和 Oh, H. 深度强化学习中的探索：综述。《信息融合》，85:1–22，2022。

Lanchantin, J., Chen, A., Dhuliawala, S., Yu, P., Weston, J., Sukhbaatar, S. 和 Kulikov, I. 多样化偏好优化。*arXiv* 预印本 *arXiv:2501.18101*，2025。

Li, T., Zhang, Y., Yu, P., Saha, S., Khashabi, D., Weston, J., Lanchantin, J. 和 Wang, T. 协同提升语言模型生成的多样性与质量。*arXiv* 预印本 *arXiv:2509.02534*，2025。

Li, Z., Xu, T., Zhang, Y., Lin, Z., Yu, Y., Sun, R. 和 Luo, Z.-Q. ReMax：一种简单、有效且高效的用于对齐大型语言模型的强化学习方法。*arXiv* 预印本 *arXiv:2310.10505*，2023。

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S. 和 Lin, M. 理解类 r1-zero 训练：一种批判性视角。*arXiv* 预印本 *arXiv:2503.20783*，2025。

Masood, M. A. 和 Doshi-Velez, F. 多样性诱导策略梯度：使用最大均值差异寻找一组多样化策略。*arXiv* 预印本 *arXiv:1906.00088*，2019。

美国数学协会。美国数学竞赛（amc），2023年。网址 https://maa.org/student-programs/amc/。

美国数学协会。美国数学邀请赛（aime），2025年。网址 https://maa.org/maa-invitational-competitions/。

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal,S., Slama, K.,Ray, A., 等。通过人类反馈训练语言模型遵循指令。神经信息处理系统进展，35:27730–27744，2022。

Pang, J.-C., Wang, P., Li, K., Chen, X.-H., Xu, J., Zhang, Z., and Yu, Y. 通过强化学习反思实现语言模型自我改进，2023。网址https://arxiv.org/abs/2305.14483。

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. 近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347*，2017。

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., 和 Guo, D. Deepseekmath：推动开放语言模型数学推理的极限，2024。网址https://arxiv.org/abs/2402.03300。

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., 和 Wu, C. Hybridflow：一种灵活高效的基于人类反馈的强化学习框架。*arXiv* 预印本 *arXiv: 2409.19256*，2024。

Shrivastava, V., Awadallah, A., Balachandran, V., Garg, S., Behl, H., 和 Papailiopoulos, D. 多采样，少思考：面向简洁推理的组过滤策略优化。*arXiv* 预印本 *arXiv:2508.09726*，2025。

Tang, Y., Zheng, K., Synnaeve, G., 和 Munos, R. 使用强化学习优化语言模型以适应推理时目标。*arXiv* 预印本 *arXiv:2503.19595*，2025。

Tang, Z., Zhang, X., Wang, B., 和 Wei, F. MathScale：扩展用于数学推理的指令微调。*arXiv*预印本 *arXiv:2403.02884*，2024。

Team, Q. 等. Qwen2 技术报告。*arXiv* 预印本 *arXiv:2407.10671*，2(3)，2024。

Walder, C. 和 Karkhanis, D. Pass@ k 策略优化：解决更困难的强化学习问题。*arXiv* 预印本 *arXiv:2505.15201*，2025。

Wang, S.、Yu, L.、Gao, C.、Zheng, C.、Liu, S.、Lu, R.、Dang, K.、Chen, X.、Yang, J.、Zhang, Z.等。超越二八定律：高熵少数词元驱动面向大语言模型推理的有效强化学习。*arXiv*预印本 *arXiv:2506.01939*，2025。

Yang, A.、Zhang, B.、Hui, B.、Gao, B.、Yu, B.、Li, C.、Liu, D.、Tu, J.、Zhou, J.、Lin, J.、Lu, K.、Xue, M.、Lin, R.、Liu, T.、Ren, X.与Zhang, Z. Qwen2.5-Math技术报告：通过自我改进迈向数学专家模型，2024。网址 https://arxiv.org/abs/2409.12122。

Yao, J., Cheng, R., Wu, X., Wu, J., 和 Tan, K.C. 面向大语言模型推理的多样性感知策略优化。*arXiv* 预印本 *arXiv:2505.23433*，2025。

Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., 等。DAPO：面向大规模应用的开源大语言模型强化学习系统。*arXiv* 预印本 *arXiv:2503.14476*，2025。

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., 和 Huang, G。强化学习真的能激励大语言模型超越基础模型的推理能力吗？*arXiv* 预印本 *arXiv:2504.13837*，2025。

Zeng, W., Huang, Y., Liu, Q., Liu, W., He, K., Ma, Z., 和 He, J。SimpleRL-Zoo：在真实环境中研究并驾驭开放基础模型的零强化学习。*arXiv* 预印本 *arXiv:2503.18892*，2025。

Zhao, Y., Liu, Y., Liu, J., Chen, J., Wu, X., Hao, Y., Lv, T., Huang, S., Cui, L., Ye, Q., 等. 几何平均策略优化。*arXiv* 预印本 *arXiv:2507.20673*，2025。

Zheng, C., Liu, S., Li, M., Chen, X.-H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., 等。群组序列策略优化。*arXiv* 预印本 *arXiv:2507.18071*，2025。

A. DAPO 和 GSPO 简介

DAPO。解耦裁剪与动态采样策略优化（DAPO）是一种 GRPO 风格的强化学习算法，专为可验证推理任务设计。在这类任务中，对于每个提示 q，从行为策略 $\pi_{\theta_{\text{old}}}$中采样一组响应 $\{o_i\}_{i=1}^{G}$，并通过采用非对称裁剪范围的代理目标进行优化，以缓解熵坍塌（“Clip-Higher”）。具体而言，DAPO 最大化

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\,\bar{A}_{i,t},\ \text{clip}(r_{i,t}(\theta),\,1-\epsilon_{\text{low}},\,1+\epsilon_{\text{high}})\,\bar{A}_{i,t}\Big)\right],$$

受动态采样约束

$$0 < \Big|\{\, o_i : \text{is\_equivalent}(a, o_i)\,\}\Big| < G.$$

此处，$r_{i,\,t}(\theta) = \frac{\pi_\theta(o_{i,\,t}|q,\,o_{i,\,<t})}{\pi_{\theta_{\text{old}}}(o_{i,\,t}|q,\,o_{i,\,<t})}$ 是词元级重要性比率，而 $\hat{A}_{i,\,t}$ 是组归一化优势（同一响应中的各词元共享），由结果奖励 $\{R_i\}_{i=1}^{G}$ 通过 $\hat{A}_{i,\,t} = \frac{R_i - \text{均值}(\{R_i\})}{\text{标准差}(\{R_i\})}$ 计算得出。

GSPO。组序列策略优化（GSPO）通过在序列级别定义重要性比率并应用序列级裁剪，解决了 GRPO 中词元级重要性比率与序列级奖励不匹配的问题。GSPO 最大化

$$J_{\text{GSPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\min\Big(s_i(\theta)\,\bar{A}_i,\ \text{clip}(s_i(\theta),\,1-\epsilon,\,1+\epsilon)\,\bar{A}_i\Big)\right],$$

其中，基于组的优势为 $\bar{A}_i = \frac{r_{(x,\,y_i)} - \text{均值}(\{r(x,\,y_i)\})}{\text{标准差}(\{r(x,\,y_i)\})}$，而（长度归一化的）序列级重要性比率为

$$s_i(\theta) = \left(\frac{\pi_\theta(y_i \mid x)}{\pi_{\theta_{\text{old}}}(y_i \mid x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\log\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}\right).$$

这种设计使优化单位（序列）与奖励粒度相一致，从而降低方差并提升稳定性，尤其适用于较长的响应和大规模训练。

B. 我们的算法

SetPO+GRPO 使用的目标函数如下：

$$\begin{aligned} J(\theta) =& \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\rho_{i,t}(\theta)\,\hat{A}_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\,\hat{A}_i\Big) - \\ & \beta\,\text{KL}\big(\pi_\theta(\cdot \mid x, a_{i,<t}) \,\big\|\, \pi_{\text{ref}}(\cdot \mid x, a_{i,<t})\big)\Bigg], \end{aligned} \tag{5}$$

其中，$\hat{A}_i$ 将原始优势 A_i 与我们的留一边际量相结合，该边际量由式（4）计算得出。SetPO 的完整算法+见算法1。

C. $\mathcal{F}(P)$ 的理论框架：轨迹多样性

本节将确立多样性度量的数学基础。我们首先定义大语言模型的多样性度量，推导其对扰动的敏感性，为有限步更新建立严格的下界，最后讨论其与解码温度相关的热力学性质。

算法1 集合级策略优化

要求： 提示数据集 $\mathcal{D}$、策略 π_θ、参考模型 π_{ref}、组大小 G、多样性系数 λ、KL 系数 β、学习率 η。

当未达到最大迭代次数**时**
  从 $\mathcal{D}$中采样一个提示批次 $\Omega = \{x_1, \ldots, x_M\}$。
  对于每个提示 $x \in \Omega$ **执行**
    采样：
    生成 G 输出 $A = \{o_1, \ldots, o_G\}$ 其中 $o_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$。
    奖励计算：
    计算每个输出的任务奖励 $\{r_1, \ldots, r_G\}$。
    计算基础优势 A_i（例如通过组归一化：$A_i = \frac{r_i - \text{均值}(r)}{\text{标准差}(r) + \epsilon}$）。
    多样性估计（留一法）：
    计算成对相似度矩阵 $K \in \mathbb{R}^{G \times G}$，其中 $K_{ij} = k(o_i, o_j)$。
    计算组得分 $D(A) = \frac{1}{G}\sum_{i=1}^{G} g(\widehat{m}_i)$。
    for $i = 1$到 G**执行**
      计算约简集得分 $D(\Omega \setminus \{o_i\})$ 使用公式 3 （重新计算局部质量）。
      计算多样性贡献 $s_i = D(\Omega) - D(\Omega \setminus \{o_i\})$。
    结束 for
    计算增广优势：$\hat{A}_i \leftarrow \bar{A}_i + \lambda \cdot s_i$。
  结束循环
  策略优化：
  最大化代理目标 $J(\theta)$ 在收集的批次上依据 (5) 进行优化：$\theta \leftarrow \theta + \eta \nabla_\theta J(\theta)$。
结束循环

C.1. 预备知识与定义

设 Ω 为轨迹空间。我们采用一个有界、对称的相似性核 $k : \Omega \times \Omega \to [0, 1]$，满足 $k(y, y) = 1$。对于任意分布 $P \in \mathcal{P}(\Omega)$，$y$ 处的核化局部质量 定义为：

$$m_P(y) := \mathbb{E}_{y' \sim P}[k(y, y')] \in [0, 1]. \tag{6}$$

该量估计了分布 P 在轨迹 y附近的语义密度。全局多样性泛函定义如下：

$$\mathcal{F}(P) := \mathbb{E}_{y \sim P}\Big[g\big(m_P(y)\big)\Big], \tag{7}$$

其中，形状函数 $g(x) : [0, 1] \to \mathbb{R}$。本文采用的一种经典选择是 $g(x) = -\log(1 + x)$，它光滑且严格递减。

假设C.1. 核函数 $k(\cdot, \cdot)$ 是可测且对称的，即 $k(x, y) = k(y, x)$。此外，对所有 x, y，均有 $0 \le k(x, y) \le 1$。形状函数 $g(x) : [0, 1] \to \mathbb{R}$ 是连续、严格递减的塑形函数，并且存在 $L < \infty$，使得对所有 $a, b \in [0, 1]$，$\|g'(a) - g'(b)\| \le L\|a - b\|$。

对于某条轨迹 $\tau \in \Omega$，考虑混合扰动

$$P_\varepsilon := (1 - \varepsilon)P + \varepsilon\,\delta_\tau, \qquad \varepsilon \in [0, 1]. \tag{8}$$

我们研究有限-ε 变化 $\mathcal{F}(P_\varepsilon) - \mathcal{F}(P)$ 及其与原始稀有度 $m_P(\tau)$ 的关系。对于任意 $y \in \Omega$，均有

$$\begin{aligned} m_{P_\varepsilon}(y) = \mathbb{E}_{y' \sim P_\varepsilon} k(y, y') &= (1 - \varepsilon)\mathbb{E}_{y' \sim P} k(y, y') + \varepsilon \mathbb{E}_{y' \sim \delta_\tau} k(y, y') \\ &= (1 - \varepsilon) m_P(y) + \varepsilon k(y, \tau) = m_P(y) + \varepsilon\Big(k(y, \tau) - m_P(y)\Big). \end{aligned} \tag{9}$$

定义相似度偏移量

$$d_\tau(y) := k(y, \tau) - m_P(y).$$

然后（9）变为

$$m_{P_\varepsilon}(y) = m_P(y) + \varepsilon\, d_\tau(y). \tag{10}$$

特别是，使用 $k(\tau,\ \tau) = 1$，可得

$$m_{P_\varepsilon}(\tau) = (1-\varepsilon)m_P(\tau) + \varepsilon = m_P(\tau) + \varepsilon\big(1 - m_P(\tau)\big).$$

从该表达式可以直接看出，较稀有样本（较小的 $m_P(\tau)$）会获得更大的注入增益，因为增量$m_{P_\varepsilon}(\tau) - m_P(\tau) = \varepsilon\big(1 - m_P(\tau)\big)$ 会随着 $m_P(\tau)$ 的减小而增大。

接下来，我们对 $\mathcal{F}(P)$ 的有限-ε 多样性变化进行分解。根据定义，

$$\mathcal{F}(P_\varepsilon) = \mathbb{E}_{y\sim P_\varepsilon}\Big[g\big(m_{P_\varepsilon}(y)\big)\Big] = (1-\varepsilon)\mathbb{E}_{y\sim P}\Big[g\big(m_{P_\varepsilon}(y)\big)\Big] + \varepsilon\, g\big(m_{P_\varepsilon}(\tau)\big).$$

Dec将 $\mathcal{F}(P)$ 分解为一个 ϵ 部分和另一个 $(1-\epsilon)$ 部分，并重新排列，可得精确恒等式 y

$$\mathcal{F}(P_\varepsilon) - \mathcal{F}(P) = \underbrace{\varepsilon\Big(g\big(m_{P_\varepsilon}(\tau)\big) - \mathcal{F}(P)\Big)}_{\text{main term: contribution of injected } \tau} + \underbrace{(1-\varepsilon)\,\mathbb{E}_{y\sim P}\big[g\big(m_{P_\varepsilon}(y)\big) - g\big(m_P(y)\big)\big]}_{\text{perturbation on existing samples}}. \tag{11}$$

利用 $m_{P_\varepsilon}(y) = m_P(y) + \varepsilon d_\tau(y)$ 由式（10）可得，

$$\mathcal{F}(P_\varepsilon) - \mathcal{F}(P) = \varepsilon\Big(g\big(m_{P_\varepsilon}(\tau)\big) - \mathcal{F}(P)\Big) + (1-\varepsilon)\,\mathbb{E}_{y\sim P}\big[g\big(m_P(y) + \varepsilon d_\tau(y)\big) - g\big(m_P(y)\big)\big]. \tag{12}$$

C.2. 无穷小分析：影响函数

现在我们分析 $\mathcal{F}$ 对无穷小变化的响应。这刻画了分布空间中多样性的“梯度”，指出应当提高哪些轨迹的权重，才能最大限度地增加多样性。

考虑混合扰动 $P_\varepsilon := (1-\varepsilon)P + \varepsilon\,\delta_\tau$，并取 $\varepsilon \to 0^+$ 时的极限。

定理 C.2 （多样性影响函数）。 $\mathcal{F}$ 在 P 处沿轨迹 τ 方向的 *Gâteaux* 导数为：

$$\mathcal{I}(\tau; P) := \frac{d}{d\varepsilon}\mathcal{F}(P_\varepsilon)\Big|_{\varepsilon=0} = \underbrace{g(m_P(\tau))}_{\textit{Intrinsic Novelty}} + \underbrace{\mathbb{E}_{y\sim P}\left[g'(m_P(y))k(y,\tau)\right]}_{\textit{Interaction Cost}} - \Psi(P), \tag{13}$$

其中 $\Psi(P)$ 是一个独立于 τ 的标量基线。

证明。 回顾在（12）中推导出的有限差分的精确分解：

$$\mathcal{F}(P_\varepsilon) - \mathcal{F}(P) = \varepsilon\Big(g\big(m_{P_\varepsilon}(\tau)\big) - \mathcal{F}(P)\Big) + (1-\varepsilon)\,\mathbb{E}_{y\sim P}\big[g\big(m_P(y) + \varepsilon d_\tau(y)\big) - g\big(m_P(y)\big)\big].$$

将等式两边除以 ε 并取 $\varepsilon \to 0^+$ 的极限，我们分别分析右侧的两项。

注入项。 根据核的连续性，$\lim_{\varepsilon\to 0} m_{P_\varepsilon}(\tau) = m_P(\tau)$。因此，第一项变为：

$$\lim_{\varepsilon\to 0^+} \frac{\varepsilon\Big(g\big(m_{P_\varepsilon}(\tau)\big) - \mathcal{F}(P)\Big)}{\varepsilon} = g\big(m_P(\tau)\big) - \mathcal{F}(P). \tag{14}$$

扰动项。 对于第二项，我们将极限作用于期望。由于 g 连续可微且定义域有界，控制收敛定理允许我们交换极限与积分：

$$\begin{aligned}\lim_{\varepsilon\to 0^+}(1-\varepsilon)\mathbb{E}_{y\sim P}\left[\frac{g\big(m_P(y) + \varepsilon d_\tau(y)\big) - g\big(m_P(y)\big)}{\varepsilon}\right] &= 1\cdot\mathbb{E}_{y\sim P}\left[\frac{d}{d\varepsilon}\Big[g\big(m_P(y) + \varepsilon d_\tau(y)\big)\Big]_{\varepsilon=0}\right] \\ &= \mathbb{E}_{y\sim P}\Big[g'\big(m_P(y)\big)\cdot d_\tau(y)\Big]. \end{aligned}\tag{15}$$

代入相似度偏移量的定义 $d_\tau(y) = k(y,\tau) - m_P(y)$ 到（15）中，我们展开交互作用：

$$\mathbb{E}_{y\sim P}\Big[g'\big(m_P(y)\big)\big(k(y,\tau) - m_P(y)\big)\Big] = \mathbb{E}_{y\sim P}\Big[g'\big(m_P(y)\big)k(y,\tau)\Big] - \mathbb{E}_{y\sim P}\Big[g'\big(m_P(y)\big)m_P(y)\Big].$$

将注入项和扰动项的结果相加：

$$\begin{aligned}\mathcal{I}(\tau;P) &= \Big(g(m_P(\tau)) - \mathcal{F}(P)\Big) + \Big(\mathbb{E}_{y\sim P}[g'(m_P(y))k(y,\tau)] - \mathbb{E}_{y\sim P}[g'(m_P(y))m_P(y)]\Big) \\ &= g(m_P(\tau)) + \mathbb{E}_{y\sim P}[g'(m_P(y))k(y,\tau)] - \underbrace{\Big(\mathcal{F}(P) + \mathbb{E}_{y\sim P}[g'(m_P(y))m_P(y)]\Big)}_{\Psi(P)}.\end{aligned}$$

项 $\Psi(P)$ 仅依赖于参考分布 P，且相对于 τ 为常数。 □

为确定 $\mathcal{I}(\tau; P)$ 与 $m_P(\tau)$ 之间的相关性，我们证明：只要 g严格递减，交互机制便会内在地与内在新颖性协同，从而惩罚冗余。

由于 $\Psi(P)$ 独立 于 τ，对于排序或基于梯度的重新加权，通常使用未中心化得分即可 e

$$\tilde{\mathcal{I}}(\tau;P) := g(m_P(\tau)) + \mathbb{E}_{y\sim P}[g'(m_P(y))\,k(y,\tau)]\,.$$

交互项取决于 τ 邻域内由 $k(y,\tau)$ 加权的 $m_P(y)$。为在不依赖脆弱的相对同质性假设的情况下得到局部单调性陈述，我们在有效邻域上采用绝对光滑性控制，并显式考虑核尾部。

假设 C.3（核阈值 η 处的绝对局部光滑性）。固定阈值 $\eta \in (0,1)$。对于 $\tau \in \Omega$，定义有效邻域

$$B_{\tau,\eta} := \{y \in \Omega :\ k(y,\tau) \geq \eta\},$$

并假设存在 $\delta = \delta(\tau,\eta) \geq 0$，使得

$$\sup_{y\in B_{\tau,\eta}} |m_P(y) - m_P(\tau)| \leq \delta(\tau,\eta).$$

引理 C.4（带显式余项界的代理约化）。在假设 *C.1* 和 *C.3*下，对任意 $\eta \in (0,1)$以及任意 $\tau \in \Omega$，

$$\tilde{\mathcal{I}}(\tau;P) = \phi(m_\tau) + r_\tau, \qquad \phi(m) := g(m) + mg'(m), \quad \textit{where} \quad m_\tau := m_P(\tau),$$

且余项满足

$$|r_\tau| \leq L\,(\delta(\tau,\eta)\,m_\tau \ +\ \eta). \tag{16}$$

证明。记作

$$\begin{aligned}\tilde{\mathcal{I}}(\tau;P) &= g(m_\tau) + \mathbb{E}[g'(m_P(y))k(y,\tau)] \\ &= g(m_\tau) + g'(m_\tau)\mathbb{E}[k(y,\tau)] + \mathbb{E}[(g'(m_P(y)) - g'(m_\tau))k(y,\tau)]\,.\end{aligned}$$

由于 $\mathbb{E}[k(y,\tau)] = m_\tau$，前两项等于 $\phi(m_\tau)$。因此余项为

$$r_\tau = \mathbb{E}[(g'(m_P(y)) - g'(m_\tau))k(y,\tau)]\,.$$

将期望拆分为 $B_{\tau,\ \eta}$ 及其补集。在 $B_{\tau,\ \eta}$，假设 C.3 以及 g' 的利普希茨连续性给出 $|g'(m_P(y))\ -g'(m_\tau)| \leq L|m_P(y)\ - m_\tau| \leq L\delta$，因此

$$\Big|\mathbb{E}\big[(g'(m_P(y)) - g'(m_\tau))k(y,\tau)\mathbf{1}_{B_{\tau,\eta}}\big]\Big| \leq L\delta\,\mathbb{E}[k(y,\tau)] = L\delta\,m_\tau.$$

在 $B^c_{\tau,\ \eta}$，我们有 $k(y,\ \tau) < \eta$ 以及 $|g'(m_P(y)) - g'(m_\tau)| \leq L$，因此

$$\Big|\mathbb{E}\Big[(g'(m_P(y)) - g'(m_\tau))k(y,\tau)\mathbf{1}_{B^c_{\tau,\eta}}\Big]\Big| \leq L\,\eta.$$

结合这两个界即可证明（16）。 □

定理 C.5 （曲率条件下的局部单调性）。 此外，假设 g 在 $(0,1]$ 上二次连续可微，且满足

$$mg''(m) < -2g'(m), \qquad \forall m \in (0,1]. \tag{17}$$

此外，固定任意 $a \in (0,1]$ 并定义

$$c_a := \min_{m\in[a,1]} \big(-\phi'(m)\big) > 0.$$

对于任意满足 $m_P(\tau_1)$ $m_P(\tau_2)$ $\in [a,1]$的 τ_1、 τ_2，我们有

$$\tilde{\mathcal{I}}(\tau_1;P) - \tilde{\mathcal{I}}(\tau_2;P) \le -c_a\big(m_P(\tau_1) - m_P(\tau_2)\big) + \big(|r_{\tau_1}| + |r_{\tau_2}|\big), \tag{18}$$

其中 r_τ 是引理C.4中的余项。特别地，当

$$m_P(\tau_1) - m_P(\tau_2) > \frac{|r_{\tau_1}| + |r_{\tau_2}|}{c_a},$$

该排序与稀有性一致：

$$m_P(\tau_1) > m_P(\tau_2) \quad \Longrightarrow \quad \tilde{\mathcal{I}}(\tau_1;P) < \tilde{\mathcal{I}}(\tau_2;P).$$

因此，在局部平滑误差 $\delta(\tau,\eta)$ 与尾部阈值 η 较小的情形下，交互项与内在新颖性项协同作用，以惩罚冗余。

证明。 由 ϕ、 $\phi'(m) = 2g'(m) + mg''(m)$ 以及（17）可知 ϕ 严格递减。在紧区间 $[a,1]$ 上， ϕ' 连续且为负，因此 $c_a = \min^{m\in[a,1]}(-\phi'(m)) > 0$。根据中值定理，对于 m_1， $m_2 \in [a,1]$，

$$\phi(m_1) - \phi(m_2) \le -c_a(m_1 - m_2).$$

将此应用于 $m_i = m_P(\tau_i)$，并使用分解式（C.4）：

$$\tilde{\mathcal{I}}(\tau_1;P) - \tilde{\mathcal{I}}(\tau_2;P) = \phi(m_1) - \phi(m_2) + (r_{\tau_1} - r_{\tau_2}) \le -c_a(m_1 - m_2) + |r_{\tau_1}| + |r_{\tau_2}|,$$

即为（18）。 □

推论 C.6 (软对数稳定性与有界梯度)。 对于 $g(m) = -\log(1+m)$，曲率条件 (17) 对所有 $m \ge 0$均成立。此外，$|g'(m)| \le 1$ 在 $[0,1]$, 上，因此影响分数在低质量区域中不会发生梯度爆炸。

证明。 对于 $g(m) = -\log(1+m)$，有 $g'(m) = -(1+m)^{-1}$ 且 $g''(m) = (1+m)^{-2}$，因此

$$\phi'(m) = 2g'(m) + mg''(m) = \frac{-2(1+m)+m}{(1+m)^2} = \frac{-(m+2)}{(1+m)^2} < 0,$$

这意味着（17）。此外， $|g'(m)| = (1+m)^{-1} \le 1$ 在 $[0,1]$上。 □

C.3. 有限扰动分析与误差界

虽然影响函数刻画了无穷小更新（$\varepsilon \to 0^+$）的最优方向，但实际的生成算法采用离散扰动（$\varepsilon > 0$）。在本小节中，我们为有限扰动下的多样性增益建立严格的界。我们证明，轨迹冗余 $m_P(\tau)$ 决定了预期增益的大小以及近似误差的紧致程度。

定理 C.7 （有限扰动界）。记 $M := \max_{x\in[0,1]} |g'(x)|$。对于任意轨迹 $\tau \in \Omega$ 以及扰动尺度 $\varepsilon \in [0,1]$，多样性的有限变化 $\Delta\mathcal{F} = \mathcal{F}(P_\varepsilon) - \mathcal{F}(P)$ 满足以下界：

$$\mathcal{M}(\tau,\varepsilon) - \mathcal{E}(\tau,\varepsilon) \;\le\; \Delta\mathcal{F} \;\le\; \mathcal{M}(\tau,\varepsilon) + \mathcal{E}(\tau,\varepsilon),$$

其中，主信号 $\mathcal{M}$ 和误差半径 $\mathcal{E}$ 定义如下：

$$\mathcal{M}(\tau,\varepsilon) := \varepsilon\left(g\big((1-\varepsilon)m_P(\tau) + \varepsilon\big) - \mathcal{F}(P)\right),$$
$$\mathcal{E}(\tau,\varepsilon) := M\cdot\varepsilon\cdot\big(m_P(\tau) + \kappa(P)\big).$$

此处， $\kappa(P) := \mathbb{E}_{y,\ y'\sim P}[k(y,y')]$ 表示该分布的全局平均相似度。

证明。 回顾式（11）中推导出的有限差分的精确分解：

$$\Delta\mathcal{F} = \mathcal{M}(\tau,\varepsilon) + B(\varepsilon), \tag{19}$$

其中，残差项为 $B(\varepsilon) = (1-\varepsilon)\mathbb{E}_{y\sim P}[g(m_{P_\varepsilon}(y)) - g(m_P(y))]$。要证明该定理，只需界定其绝对值 $|B(\varepsilon)|$。

根据假设C.1，代入 线性更新规则 $m_{P_\varepsilon}(y) = m_P(y) + \varepsilon d_\tau(y)$，其中 $d_\tau(y) = k(y, \tau) - m_P(y)$，我们得到 t:

$$|g(m_{P_\varepsilon}(y)) - g(m_P(y))| \le M \cdot \varepsilon \cdot |d_\tau(y)|.$$

由于核非负（$k \ge 0$），直接应用三角不等式可得：

$$|d_\tau(y)| = |k(y,\tau) - m_P(y)| \le k(y,\tau) + m_P(y).$$

对 $y \sim P$ 取期望：

$$\mathbb{E}_{y\sim P}|d_\tau(y)| \le \mathbb{E}_{y\sim P}[k(y,\tau)] + \mathbb{E}_{y\sim P}[m_P(y)] = m_P(\tau) + \kappa(P). \tag{20}$$

因此，有

$$|B(\varepsilon)| \le (1-\varepsilon)\cdot \mathbb{E}_{y\sim P}\left[M \cdot \varepsilon \cdot |d_\tau(y)|\right] \le M \cdot \varepsilon \cdot (m_P(\tau) + \kappa(P)) = \mathcal{E}(\tau,\varepsilon).$$

双侧界可由 $-|B(\varepsilon)| \le B(\varepsilon) \le |B(\varepsilon)|$立即得到。 □

定理C.7 揭示了一种结构对偶性，其中冗余 $m_P(\tau)$ 同时决定了一次更新的潜在收益与理论风险。具体而言，最小化 $m_P(\tau)$ 会同时增强新颖性信号 $\mathcal{M}$ 并严格压缩误差半径 $\mathcal{E}$。这意味着，较为罕见的轨迹能够提供更优的权衡。它们以更高的确定性带来更高的预期增益。为严格地将这一观察付诸实践，我们关注认证下界，其定义为$J(\tau,\varepsilon) := \mathcal{M}(\tau,\varepsilon) - \mathcal{E}(\tau,\varepsilon)$，其作为改进的最坏情况保证。

定理C.8 (认证增益的单调性)。 若 $\epsilon > 0$，则认证下界 $J(\tau,\varepsilon)$ 是局部质量 $m_P(\tau)$的严格递减函数。

证明。令 $m := m_P(\tau) \in [0,1]$。我们分析下界函数 J关于 m的敏感性 ：

$$J(m) = \varepsilon\Big[g\big((1-\varepsilon)m + \varepsilon\big) - \mathcal{F}(P) - Mm - M\kappa(P)\Big].$$

对 $J(m)$ 求导，得到：

$$\frac{\partial J}{\partial m} = \varepsilon\left[(1-\varepsilon)g'\big((1-\varepsilon)m+\varepsilon\big) - M\right].$$

该导数严格为负，这一点可由系统的性质直接得出。 □

这证实了冗余会招致“双重惩罚”：既削弱多样性信号，又放大扰动噪声。

D. 留一法贡献的离散分析 s_i

定义。 令 $\Omega = \{o_1,\dots,o_G\}$，且 $G \ge 3$。假设相似性核是对称的： $k(o_p, o_q) = k(o_q, o_p)$，并且在其定义域上有界。对于任意有限集合 S，且 $|S| = n \ge 2$，定义留一局部质量

$$m_u(S) = \frac{1}{|S|-1}\sum_{v\in S,\, v\ne u} k(o_u, o_v),$$

以及多样性得分

$$D(S) = \frac{1}{|S|}\sum_{u\in S} g\big(m_u(S)\big),$$

其中 g 在相关范围内连续可微且严格递减： $g'(x) < 0$.

对于完整集合 Ω，我们记作 $m_i := m_i(\Omega)$。对于留一集合 $A_{-i} := \Omega \setminus \{o_i\}$，我们记作

$$m_j^{(-i)} := m_j(A_{-i}), \qquad D_{-i} := D(A_{-i}).$$

你的算法使用留一法贡献

$$s_i := D(\Omega) - D_{-i}.$$

引理 D.1 （移除下的精确密度更新律）。 对于任意 $j \neq i$,

$$m_j^{(-i)} = \frac{1}{G-2} \sum_{p \in \Omega,\, p \neq j,i} k(o_j, o_p) = \frac{(G-1)m_j - k(o_j, o_i)}{G-2} = m_j + \frac{m_j - k(o_j, o_i)}{G-2}.$$

等价地,

$$m_j^{(-i)} - m_j = \frac{m_j - k(o_j, o_i)}{G-2}.$$

证明。 设 $R_{ji} := \sum_{p \in \Omega\ p \neq j,\ i} k(o_j,\ o_p)$。则 $m_j = \frac{1}{G-1}\big(k(o_j,\ o_i) + R_{ji}\big)$ 且 $m_j^{(-i)} = \frac{1}{G-2} R_{ji}$。消去 $R_{ji} = (G-1)m_j - k(o_j,\ o_i)$ 并代入。 □

定理 D.2 （ s_i的精确分解）。 令 $G = |\Omega| \geq 3$。则

$$s_i = \underbrace{\frac{1}{G} g(m_i)}_{\mathcal{T}_{self}} + \underbrace{\left(\frac{1}{G} - \frac{1}{G-1}\right) \sum_{j \neq i} g(m_j)}_{\mathcal{T}_{norm}} + \underbrace{\frac{1}{G-1} \sum_{j \neq i} \left[g(m_j) - g(m_j^{(-i)})\right]}_{\mathcal{T}_{interaction}}.$$

此外，根据引理 *D.1*，每个交互求和项仅通过 $k(o_j,\ o_i)$ 依赖于 $m_j^{(-i)} = m_j + \frac{m_j - k(o_j, o_i)}{G-2}$.

证明。 展开 $D(\Omega) = \frac{1}{G}\big(g(m_i) + \sum_{j \neq i} g(m_j)\big)$ 和 $D_{-i} = \frac{1}{G-1} \sum_{j \neq i} g(m_j^{(-i)})$，并对各项重新分组。 □

定理 D.3 （关于成对相似度的严格单调性）。 固定 $i \neq t$ 并记 $k_{it} := k(o_i,\ o_t)$。则

$$\frac{\partial s_i}{\partial k_{it}} = \frac{1}{G(G-1)} \Big(g'(m_i) + g'(m_t)\Big) < 0.$$

证明。 D_{-i} 不涉及 o_i，因此独立于 k_{it}。因此 $\frac{\partial s_i}{\partial k_{it}} = \frac{\partial D(\Omega)}{\partial k_{it}}$。在 $D(\Omega)$中，只有 m_i 和 m_t 依赖于 k_{it}和 $\frac{\partial m_i}{\partial k_{it}} = \frac{1}{G-1}$、$\frac{\partial m_t}{\partial k_{it}} = \frac{1}{G-1}$。因此

$$\frac{\partial D(\Omega)}{\partial k_{it}} = \frac{1}{G} \left(g'(m_i) \frac{1}{G-1} + g'(m_t) \frac{1}{G-1}\right) = \frac{1}{G(G-1)} (g'(m_i) + g'(m_t)) < 0,$$

因为 $g' < 0$。 □

定理D.4 （逐点支配下的稀有—冗余排序）。 设 $r \neq c$ 为Ω中的两个下标。假设 o_r 对所有其他点逐点不比 o_c 更相似：

$$k(o_r, o_j) \leq k(o_c, o_j) \quad \forall j \in \Omega \setminus \{r, c\},$$

且对于至少一个这样的 j，严格不等式成立。接下来

$$s_r > s_c.$$

证明。 由于 $s_i = D(\Omega) - D_{-i}$ 和 $D(\Omega)$ 是共有的，

$$s_r - s_c = D_{-c} - D_{-r}.$$

展开

$$D_{-c} = \frac{1}{G-1}\left(g(m_r^{(-c)}) + \sum_{j\neq r,c} g(m_j^{(-c)})\right), \quad D_{-r} = \frac{1}{G-1}\left(g(m_c^{(-r)}) + \sum_{j\neq r,c} g(m_j^{(-r)})\right),$$

因此

$$s_r - s_c = \frac{1}{G-1}\left(g(m_r^{(-c)}) - g(m_c^{(-r)}) + \sum_{j\neq r,c}\left[g(m_j^{(-c)}) - g(m_j^{(-r)})\right]\right).$$

现在注意：

$$m_r^{(-c)} = \frac{1}{G-2}\sum_{j\neq r,c} k(o_r, o_j), \quad m_c^{(-r)} = \frac{1}{G-2}\sum_{j\neq r,c} k(o_c, o_j),$$

因此 $m_r^{(-c)} \leq m_c^{(-r)}$根据假设成立；若有任何一项为严格不等式，则整体也为严格不等式。对于 $j \neq r,\ c$，对两种移除情况应用引理 D.1 并作差：

$$m_j^{(-c)} - m_j^{(-r)} = \frac{(G-1)m_j - k(o_j, o_c)}{G-2} - \frac{(G-1)m_j - k(o_j, o_r)}{G-2} = \frac{k(o_j, o_r) - k(o_j, o_c)}{G-2} \leq 0,$$

若某处的支配关系是严格的，则至少有一个 j 也再次严格成立（利用 k的对称性）。由于 g 严格递减， $s_r - s_c$ 展开式中的每个方括号内的差值均非负，且至少有一个严格为正，因此 $s_r - s_c > 0$。 □

推论 D.5 (差距的加权刻画)。 存在 ξ_0，位于 $m_r^{(-c)}$ 和 $m_c^{(-r)}$ 之间；且存在 $\{\xi_j\}_{j\neq r,\ c}$ 位于$m_j^{(-c)}$ 和 $m_j^{(-r)}$ 之间，使得

$$s_r - s_c = \frac{1}{(G-1)(G-2)}\sum_{j\neq r,c}(w_0 + w_j)\Big(k(o_c, o_j) - k(o_r, o_j)\Big), \quad w_0 := -g'(\xi_0) > 0,\ \ w_j := -g'(\xi_j) > 0.$$

因此，只要 o_c 相对于 o_r 的加权相似性优势为正，即使逐点支配不成立， $s_r > s_c$ 也成立。

证明。 对 $g(m_r^{(-c)}) - g(m_c^{(-r)})$ 以及定理 $g(m_j^{(-c)}) - g(m_j^{(-r)})$ 展开式中的每个 D.4应用中值定理，然后利用

$$m_r^{(-c)} - m_c^{(-r)} = \frac{1}{G-2}\sum_{j\neq r,c}(k(o_r, o_j) - k(o_c, o_j)), \quad m_j^{(-c)} - m_j^{(-r)} = \frac{k(o_j, o_r) - k(o_j, o_c)}{G-2},$$

并化简。 □

推论 D.6 (移除下的稀释与拥塞)。 对于任意 $j \neq i$，移除 o_i 会使 o_j 变得更密集 （质量增加），当且仅当 o_i 对 o_j 起稀释作用：

$$m_j^{(-i)} > m_j \iff k(o_j, o_i) < m_j.$$

因此，当且仅当移除 $k(o_j, o_i) < m_j$ 时，邻居得分会降低：

$$g(m_j^{(-i)}) < g(m_j) \iff k(o_j, o_i) < m_j.$$

证明。 由引理 D.1以及 g的单调性可立即得出。 □

E. 实验设置

E.1. 训练设置

我们按照（Yao 等人，2025）中的实验设置，在 GSM8K 上训练 Qwen-1.5B 和 Qwen-7B 模型，以确保公平比较。SetPO+GRPO、SetPO+GSPO 和 SetPO+DAPO 的超参数分别汇总于表 4、表 5 和表 6。针对 Qwen-32B 的扩展实验，其详细设置见表 7。Countdown 实验的相应超参数列于表 8。在所有设置中，核函数定义为 $k(o_i, o_j) = \text{clamp}(\text{sim}(e_i, e_j))$，其中 o_i 和 o_j 表示输出答案，e_i、e_j 是它们对应的语义嵌入。这里，sim 表示余弦相似度，并应用截断操作以确保核值保持在范围 $[0, 1]$内。

表 *4.* SetPO+GRPO 和 GRPO 的超参数与实验设置

参数	值	参数	值
基础模型	Qwen2.5-Math-7B/1.5B	算法	GRPO
数据集	GSM8K	嵌入模型	Qwen3-Embedding-0.6B
	训练配置		
全局批次大小	64	学习率	3×10^{-6}
KL 系数	1×10^{-4}	总训练轮数	2
生成样本数 (N)	6	训练生成温度	0.9
裁剪比率	0.2	最大响应长度	756
边际权重	0.05	形状函数	$-\log(1+x)$
	推理与评估配置		
推理温度	0.5	Top-p	0.9

表*5.* SetPO+GSPO 和 GSPO 的超参数与实验设置

参数	值	参数	值
基础模型	Qwen2.5-Math-7B/1.5B	算法	GSPO
数据集	GSM8K	嵌入模型	Qwen3-Embedding-0.6B
	训练配置		
全局批次大小	64	学习率	1×10^{-6}
KL 系数	1×10^{-4}	总训练轮数	2
下限裁剪比率 (ϵ_{low})	0.0003	上限裁剪比率 (ϵ_{high})	0.0004
温度	0.9	最大响应长度	756
rollout 次数	6	小批量大小	64
边际权重	0.1	形状函数	$-\log(1+x)$
	推理与评估配置		
推理温度	0.5	Top-p	0.9

E.2. 评估设置

遵循评估协议（Chen，2021），我们采用 Pass@k 指标的无偏估计量。直接仅生成 k 个样本来衡量准确率，通常会导致较高的方差。因此，我们生成更多 n 个样本，并通过解析计算在大小为 k 的随机子集中找到至少一个正确解的概率。设 n表示样本总数，c表示正确解的数量。无偏估计量定义如下：

$$\text{Pass@}k := 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}} \tag{21}$$

表 6. SetPO+DAPO 和 DAPO 的超参数与实验设置

参数	值	参数	值
基础模型	Qwen2.5-Math-7B/1.5B	算法	DAPO
数据集	GSM8K	嵌入模型	Qwen3-Embedding-0.6B
训练配置			
全局批次大小	64	学习率	1×10^{-6}
KL 系数	1×10^{-4}	总训练轮数	2
下限裁剪比率（ϵ_{low}）	0.2	上限裁剪比率（ϵ_{high}）	0.28
温度	0.9	最大响应长度	756
rollout 次数	6	小批量大小	64
边际权重	0.2	形状函数	$-\log(1+x)$
推理与评估配置			
推理温度	0.5	Top-p	0.9

表 7. Qwen2.5-32B-Base 的 SetPO+GRPO 和 GRPO 的超参数与实验设置

参数	值	参数	值
基础模型	Qwen2.5-32B-Base	算法	GRPO
数据集	Math-Hard（Zeng等人，2025）	嵌入模型	Qwen3-Embedding-0.6B
训练配置			
全局批次大小	1024	学习率	1×10^{-6}
KL 系数	1×10^{-3}	总训练轮数	15
裁剪比率	0.2	小批量大小	256
温度	1	最大响应长度	4096
边际权重	0.1	形状函数	$-\log(1+x)$
展开次数	6	小批量大小	64
推理与评估配置			
推理温度	0.5	Top-p	0.9

表 8. 用于 Countdown 的 Qwen2.5-3B GRPO 实验的详细超参数设置。

参数	值	参数	值
基础模型	Qwen2.5-3B	算法	GRPO
数据集	Countdown	嵌入模型	Qwen3-Embedding-0.6B
训练配置			
全局批次大小	128	学习率	1×10^{-6}
KL 系数（β）	0.001	训练轮数	1
采样组大小（G）	7/ 3	训练样本总数	86400
提示词最大长度	512	响应最大长度	2048
边际权重	0.5	形状函数	$-\log(1+x)$
推理与评估配置			
评估响应长度	2048	评估样本数 (N)	64
温度	1.0/ 0.5	Top-p	0.9

在我们的实验中，我们将总生成预算固定为 $n = 64$，并评估 $k \in \{1, 2, 4, 8, 16, 32, 64\}$ 的 Pass@k。当 $n = k$ 时，该公式自然简化为经验准确率。此外，如果错误样本的数量少于 k（即 $n - c < k$），则该指标定义为 1，这意味着任意选取大小为 k 的子集，都必然包含一个正确解。

F. 数学推理的更多结果

F.1. 数学基准上的完整比较

从 Pass@1 到 Pass@k 的完整比较如图 6 和 7所示。我们评估了三种基线方法及其 SetPO 增强变体在不同采样预算下的表现。所有实验均使用五个随机种子进行，以减少随机性；阴影区域表示方差。我们观察到，SetPO 始终能降低基线方法的方差，并带来稳定的性能提升。此外，随着采样预算增加，收益并未减弱，这表明 SetPO 通过覆盖更多奖励模式，有效提升了结果多样性。最后，图 10进一步证实，我们的方法在扩展到更大模型时依然有效。

F.2. 当 K 足够高时取得更高性能

正如（Yue等人，2025）所指出的，Pass@1 的提升并不总能反映真正的能力进步。在某些情况下，只需增加基础模型的采样预算，就能获得相近的性能。此外，他们还表明，采用当前策略梯度方法训练的模型在 Pass@k 指标上可能逊于基础模型，这说明 RLVR 带来的表面提升，部分可能源于结果多样性的降低，而非底层能力的增强。受这一观察启发，我们评估了模型在 AIME24 上从 pass@1 到 pass@256的表现。如图 8所示，我们的方法在整个 K范围内（包括 $K = 256$）始终显著优于基础模型，这表明性能提升并非仅仅是增加采样带来的假象，而是反映了相较于基础模型的稳健进步。

F.3. 将 SetPO 添加到原始算法中的时间开销

尽管 SetPO 使用额外的嵌入模型（Qwen3-0.6B-Embedding）来计算留一边际奖励，但额外计算开销很小且易于控制。图 9 报告了基线算法及其 SetPO 增强版本的端到端墙钟时间。在 GRPO、GSPO 和 DAPO 中，加入 SetPO 后总运行时间增加不到 10%，这表明基于嵌入的边际奖励估计不会显著影响训练吞吐量。此外，这种开销并非不可避免；通过更高效地实现嵌入流水线（例如优化并行化以及改进 IO/序列化），还可以进一步降低开销，使 SetPO 成为现有强化学习流水线中实用的即插即用增强方案。

G. 案例研究

G.1. 案例研究：数学推理解答的多样性

在本小节中，我们使用一道典型的奥林匹克风格数论问题，来展示我们的 SetPO+GRPO 训练所生成的推理路径的多样性。

问题陈述。设 $b \geq 2$为一个整数。若一个正整数 n *b-eautiful* 用 b 进制表示时恰有两位，且这两位数字之和为 $\sqrt{n}$，则称其为 n *b-eautiful* 数。例如，81 是 13-*eautiful* 数，因为 $81 = 6\,3_{13}$ 且 $6 + 3 = \sqrt{81}$。求最小的整数 $b \geq 2$，使得大于十个整数是 b-*eautiful* 数。

训练后的模型不会只收敛到一个证明，而是能够为这个问题生成多种解题策略，并得出相同的最终答案。下面我们展示其中两种解法：一种主要通过对数位约束进行代数变形，另一种则采用构造性视角。这些不同的方法为 SetPO+GRPO 能够鼓励多样化的解题策略提供了具体证据，也验证了我们生成的解法具有多样性。

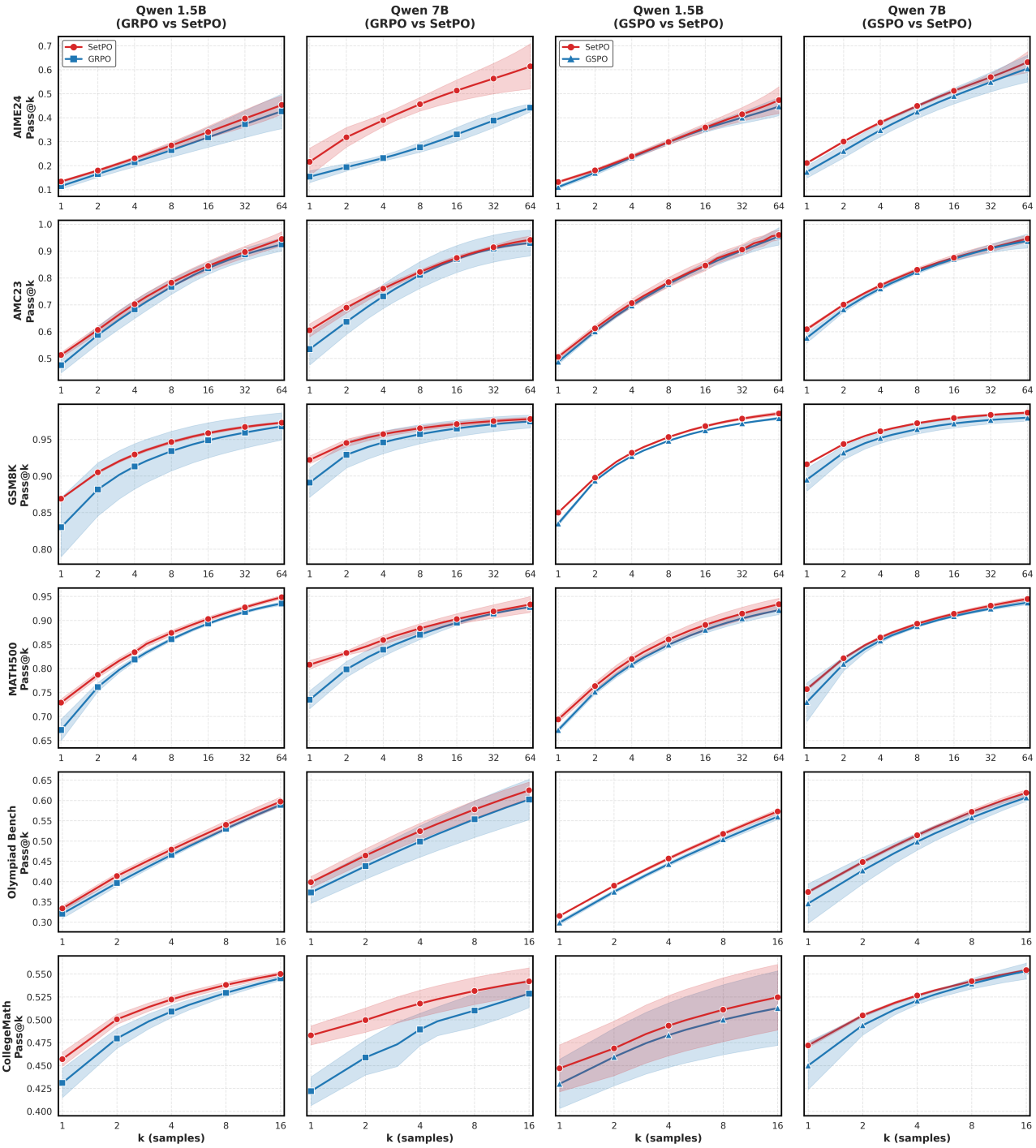

图 *6.* 在多个基准测试中，SetPO+GRPO 与 GRPO、SetPO+GSPO 与 GSPO 在 Qwen2.5-Math-1.5B 和 Qwen2.5-Math-7B 上的性能对比。SetPO 的表现始终优于相应的基线方法，并且在大多数情况下能够降低训练变异性。阴影区域表示多次运行的方差，实线表示均值。

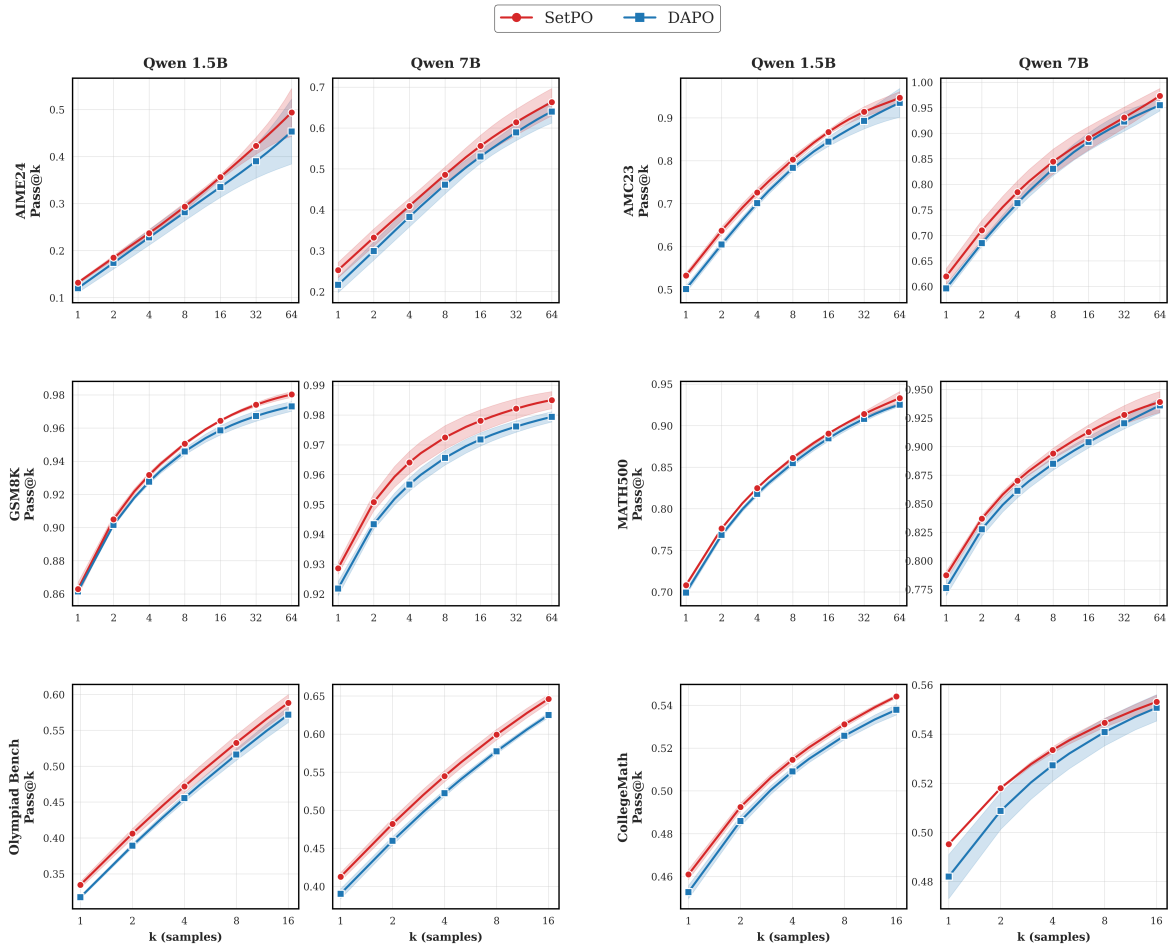

图7. 在多个基准测试上，SetPO+DAPO 与 DAPO 在 Qwen2.5-Math-1.5B 和 Qwen2.5-Math-7B 上的表现。SetPO 始终优于相应的基线方法。阴影区域表示多次运行的方差，实线表示均值。

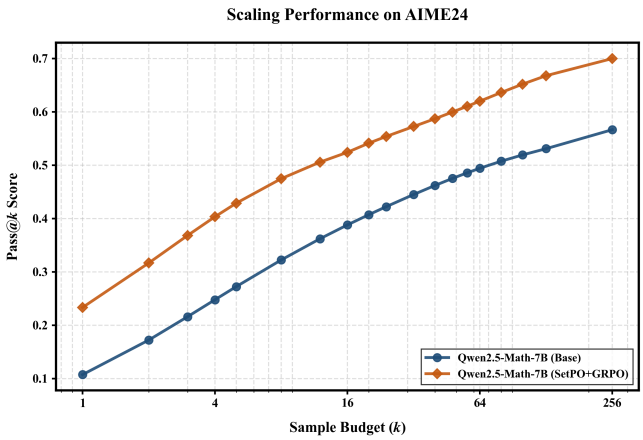

图8. 在不同采样预算 k 下，AIME24基准上的 Pass@k 表现，对比指令微调基础模型与强化学习微调模型（SetPO+GRPO）。值得注意的是，SetPO+GRPO 在较大的 k 下仍保持出色表现，即使在高通过率评估中也超过了基础模型。

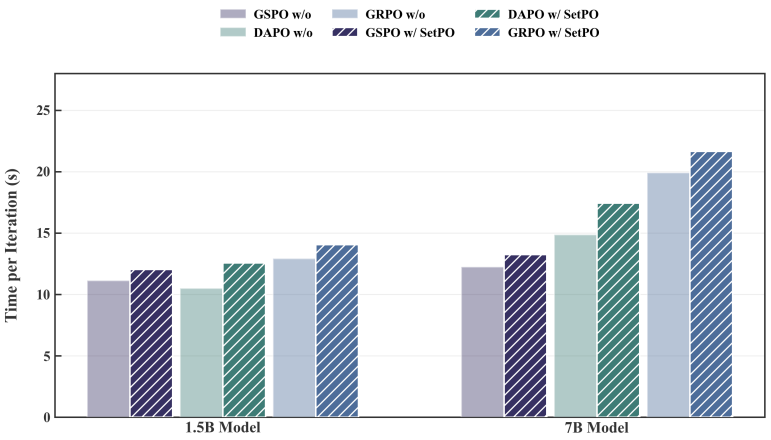

图*9*。 基线算法与采用 SetPO 增强的变体之间的墙钟时间对比。对于 GRPO、GSPO 和 DAPO，加入 SetPO 带来的额外计算开销不到 10%，通过更高效的实现还可以进一步降低这一开销。

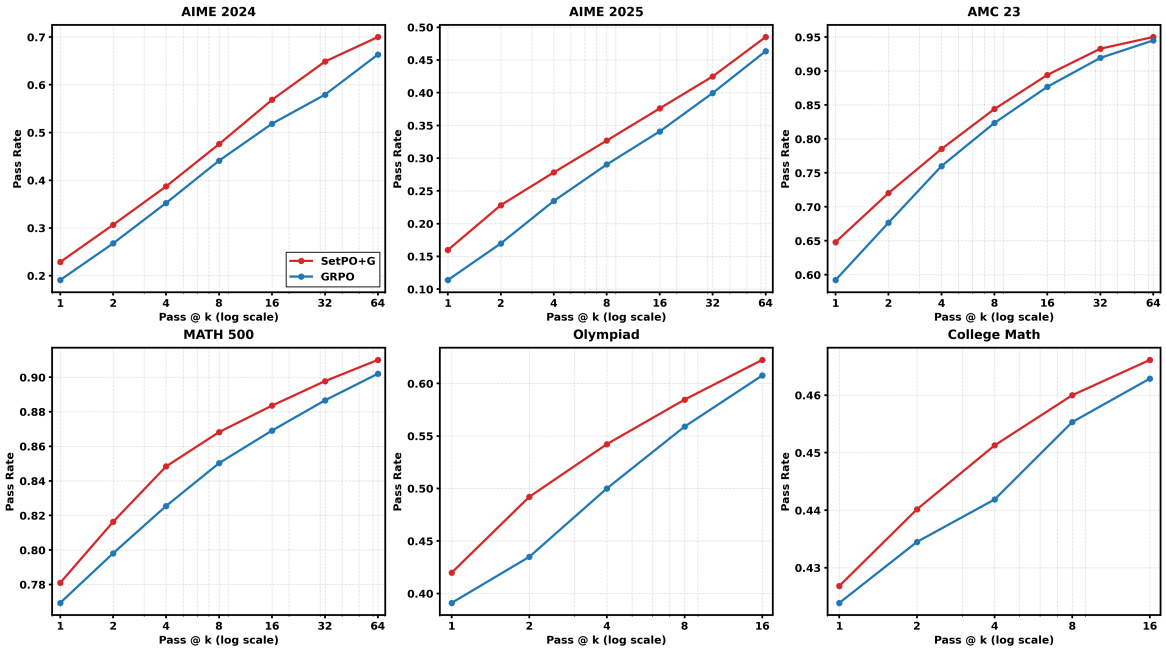

图*10*。 Qwen2.5-32B 上 SetPO+GRPO 与 GRPO 的性能对比。

解法一

要解决这个问题，我们需要找到最小的进制 $b \geq 2$，使得存在十个以上的 b-美丽整数。如果一个数 n 在进制 b 下恰好有两个数字，且这些数字之和 $\sqrt{n}$等于 b，那么它就是 b-美丽数。让我们逐步分析这个问题：

1. 进制 b 下的两位数可以写成 $n = ab + c$，其中 a 和 c 是数字，$1 \leq a \leq b-1$，且$0 \leq c \leq b-1$。

2. 要使 n 成为 b-美丽数，条件是 $a + c = \sqrt{n}$。由于 $n = ab + c$，我们有 $\sqrt{ab + c} = \overline{a+c}$。两边平方后，得到 $ab + c = (a + c)^2 = a^2 + 2ac + c^2$。

3. 对方程重新整理，得到 $ab + c = a^2 + 2ac + c^2$，可化简为 $ab = a^2 + 2ac + c^2 - c$。还可以进一步化简为 $ab = a^2 + c(2a + c - 1)$。

我们需要找出满足该方程且解的数量超过十个的最小进制 b。我们将从$b = 2$ 开始，逐步增加 b，直到找到一种进制，使得其中超过十个整数是 b美丽数。使 $b \geq 2$ 进制下 b美丽整数超过十个的最小整数是 211。

解法二

让我们逐步分析这个问题，并用来找到答案。

1. **理解 b-美丽数**：如果一个数 n 在 b 进制下恰有两位数字，且这些数字之和为 $\sqrt{n}$，那么它就是 b美丽数。如果 n 在 b进制下有两位数字，就可以表示为 $n = d_1 \cdot b + d_2$，其中 d_1和 d_2 是数字，且 $1 \leq d_1 \leq b-1$ 和 $0 \leq d_2 \leq b-1$。n 是 b美丽数的条件是$d_1 + d_2 = \sqrt{n}$。

2. **的范围**：由于 n 在 b进制下有两位数字，因此 $b \leq n < b^2$。所以，$b \leq n < b^2$ 和 $\sqrt{n}$ 必须是整数。这意味着 n 必须是一个完全平方数，记作 $n = k^2$，其中 $b \leq k^2 < b^2$。因此，k 必须是满足$\sqrt{b} \leq k < b$的整数。

3. **寻找 b-美丽数**：对于范围 $\lceil \sqrt{b} \rceil \leq k < b$ 内的每个 k，我们需要检查 k^2 是否可以表示为 $d_1 \cdot b + d_2$，且满足 $d_1 + d_2 = k$。这意味着 $d_1 = \left\lfloor \frac{k^2}{b} \right\rfloor$ 和 $d_2 = k^2 \bmod b$。我们需要检查 $d_1 + d_2 = k$ 是否成立。

4. **计数 b-美丽数**：我们需要找到最小的 b，使得 b-美丽数的数量超过十个。我们将从 2 开始遍历 b，并计算每个 b 对应的 b-美丽数的数量。让我们在此实现代码以找到答案。大于十个 b-美丽整数所对应的最小整数 $b \geq 2$ 是 211。

G.2. 案例研究：倒计时任务解答的多样性

为了展示生成输出的多样性，我们分析 Countdown 任务的一个具体实例，其目标是使用操作数集合 $\{54, 24, 21, 3\}$ 构造目标整数 63。经过 SetPO+GRPO 训练的模型成功找到了多条有效的计算路径。模型没有收敛到单一启发式方法，而是生成了七种不同的表达式。这些解法在两个维度上体现出多样性：代数策略（使用不同的主要运算符）和结构变化（利用交换律和运算符优先级）。

生成的解法如下：

$$\begin{cases} 54+(24-21)\times 3 & (22) \\ (24-21)\times 3+54 & \\ 54-(21-24)\times 3 & (23) \\ (54/3)+24+21 & (24) \\ 24+54/3+21 & \\ 24+21+54/3 & \\ 21+24+54/3 & \end{cases}$$

公式（22）至（23）体现了一种基于差分放大的策略，具体而言是计算 $3\times 3+54$。值得注意的是，公式（23）通过将加法转换为减去负差值，展现了模型对有符号算术的隐式理解。相较之下，公式（24）采用基于除法的策略，对各个部分进行求和（$18+45$）。后面这些公式之间的差异证实，模型在解空间中掌握了交换性质。这一输出表明，模型并非只是回忆单一的解题模式，而是主动探索算术领域，提供多样且有效的构造方式。

H. 本文使用的提示

用于数学推理的提示以下是GSM8K数据集的一个典型案例。

一个GSM8K问题示例

朱莉正在读一本120页的书。昨天，她读了12页，今天她读的页数是昨天的两倍。如果她想在明天读完剩余页数的一半，她应该读多少页？请逐步推理，并将最终答案放在 \boxed{}中。

用于Countdown的提示 以下是倒计时数据集的一个典型案例。

Countdown题目示例

用户与助手之间的对话。用户提出问题，助手解答。助手首先在心中思考推理过程，然后向用户给出答案。用户：使用数字 [35, 51, 46],组成一个等于40的算式。你可以使用基本算术运算（+、-、*、/），且每个数字只能使用一次。请在 <think> </think> 标签中展示解题过程，并在 <answer> </answer> 标签中返回最终答案，例如 <answer>(1 + 2) / 3 </answer>。助手：我来一步一步地解答。<think>

用于多样性评估的提示我们的提示用于多样性评估，内容如下。我们使用 Gemini-2.5-Flash。

提示：AIME解题多样性评估器

你是一位数学竞赛教练专家，负责评估 AIME 解法。你的任务是评估 16 个生成尝试中解题策略的多样性。

输入数据

问题：
{问题}
16 次解题尝试：{响应}

评估标准

1. **识别策略簇**：根据解法所依据的基本数学路径进行分组。 2. **计算／穷举**：使用 Python 脚本或通过反复试错且没有证明的解法，归为同一个簇。 3. **忽略表面差异**：变量名称不同或代数步骤略有差异，都属于同一种策略。

评分标准（1-5）

得分1（最低）

- **13+ 响应** 依赖完全相同的定理或计算路径。
- 没有对替代性数学性质进行任何有意义的探索。

评分 2（低）

- **10-12 个响应**遵循标准的教科书方法。
- 只有 1 或 2 次尝试展现了不同的视角。

评分 3（中等）

- **7-9 个响应**采用主导策略。
- 其中包含 2-3 种明显不同的数学框架。

得分4（高）

- **4-6 个响应**使用最常见的方法。
- 这组内容展示了3至4种各具特色且发展完善的策略。

得分5（最高）

- **3 或更少的响应**使用最常见的方法。
- 这些解法涵盖了广泛的技术。
- 至少成功采用了 4+ 种本质上不同的数学路径。

输出格式

首先在双括号内输出分数，后接简短理由。
格式示例：[[3]]这些解法采用了两种主要方法：坐标几何和相似三角形.....

你的响应：