arXiv:2602.00983v1 [cs.CL] 1 Feb 2026

DISPO：提升大语言模型数学推理强化学习中的训练效率与稳定性

Batuhan K. Karaman* Aditya Rawal Suhaila Shakiah 康奈尔大学
Amazon AGI Amazon AGI Mohammad Ghavamzadeh Mingyi Hong Arijit
Biswas Ruida Zhou†Amazon AGI Amazon AGI Amazon AGI Amazon AGI

2026年2月3日

摘要

可验证奖励强化学习已成为增强大语言模型推理能力的一种颇具前景的范式，尤其是在数学领域。该领域的现有方法呈现出明确的权衡：PPO类方法（例如GRPO/DAPO）能够提供训练稳定性，但由于其对策略更新施加信赖域约束，学习轨迹较为缓慢；而REINFORCE类方法（例如CISPO）展现出更高的学习效率，但由于它们在裁剪重要性采样权重的同时仍允许信任域之外存在非零梯度，因而存在性能不稳定性。为克服这些局限，我们提出DISPO，一种简单而有效的REINFORCE风格算法，它解耦了正确响应与错误响应的重要性采样权重上限裁剪与下限裁剪，从而形成四种可控的策略更新机制。通过有针对性的消融实验，我们揭示了每种机制对训练的影响：对于正确响应，权重 > 1会提高平均词元熵（即探索），而权重 < 1则会降低平均词元熵（即蒸馏）——二者均有益处，但过度时会导致性能逐渐退化。对于错误响应，过于严格的裁剪会引发性能突然崩溃，表现为重复输出（当权重 > 1时）或响应长度趋近于零（当权重 < 1时）。通过分别调节这四个裁剪参数，DISPO在维持探索-蒸馏平衡的同时避免灾难性失败，在AIME'24上达到61.04%（相比之下CISPO为55.42%，DAPO为50.21%），并在多种基准测试和模型上取得类似提升。

1 引言

近期的大语言模型（LLM），例如 DeepSeek-R1 [Guo 等， 2025], Qwen3 [Yang 等， 2025],OpenAI o1 [Jaech 等， 2024], 以及 Claude Sonnet 3.5 [Anthropic， 2024] 在推理任务上展现出了强大的性能，包括数学问题求解、逻辑演绎和科学分析。这些任务要求模型在探索多样化求解路径的同时，保持连贯的思维链（CoT）。推动这些进展的一个关键因素是可验证奖励强化学习（RLVR） [Yang 等，2025年；Guo 等，2025年；Lambert 等， 2024], 在这一过程中，模型通过强化学习目标优化输出，而这些目标直接与响应正确性相关。

GRPO [Shao 等人， 2024] 和 DAPO [Yu 等人， 2025] 等 PPO类带可验证奖励的强化学习算法，凭借类似信任域的约束带来的稳定性，在大规模部署（例如 DeepSeekMath、Qwen）中占据主导地位，但代价是学习速度较慢。近期研究重新审视了 REINFORCE [Williams， 1992], 它允许在信任域之外产生非零梯度。CISPO [Chen 等人， 2025]等 REINFORCE风格算法所需的更新次数少得多，却能超越近端策略优化方法，展现出颇具吸引力的训练效率 [Ahmadian 等人，

*在 Amazon AGI 实习期间完成的工作
†通讯作者：zruida@amazon.com

本文已被第二十九届人工智能与统计国际会议（AISTATS）2026录用。

2024年，Arnal 等人， 2025]。然而，这种效率是以牺牲稳定性为代价的：Zheng 等人 [2025] 报告称，CISPO 会突然出现性能崩溃；Arnal 等人 [2025]则记录了 REINFORCE 动态的不稳定性。图 1 展示了这种权衡——DAPO 保持稳定但速度较慢（蓝色），而 REINFORCE 初期训练效率较高，后期却会崩溃（橙色）。

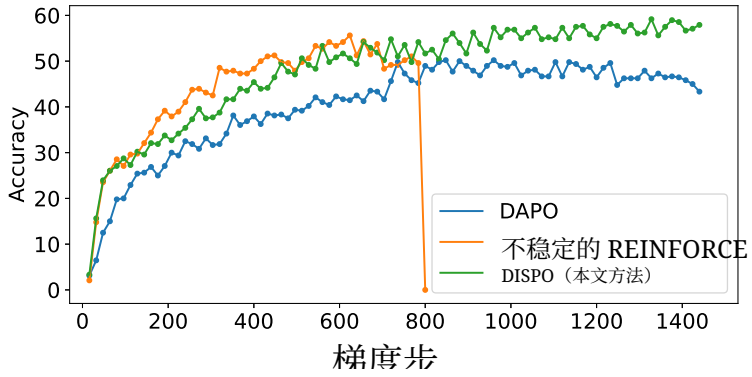


图 1：RLVR 算法的学习曲线。

在本文中，我们提出了 DISPO（**D**ecoupled **I**mportance **S**ampling-weighted**P**olicy **O**ptimization，解耦重要性采样加权策略优化），这是一种简单而有效的 REINFORCE风格 算法，实现了两全其美：在保持效率的同时确保稳定性，从而支持更长时间的训练而不会出现性能崩溃，如图 1（绿色）所示。DISPO 根据（i）奖励符号（正确与错误）以及（ii）重要性采样（IS）权重高于还是低于 1，分别设置不同的上限和下限裁剪边界，从而产生四种可控的策略更新机制，如图 2 所示。通过有针对性的消融实验，我们分离并揭示了每种机制对训练动态的独特影响。对于正确响应，重要性权重 > 1 放大词元熵以促进探索（机制1），而权重 < 1 抑制熵以进行蒸馏（机制 2）——也就是说，将概率集中到能够产生正确回答的词元上。这两种效果适度时都有益，但过度时可能导致性能逐渐退化。对于错误响应，过于严格的裁剪会触发灾难性失败：当权重 > 1 （机制 3）时，会出现重复输出；而当权重 < 1 （机制 4）时，则会出现极短的响应。我们的主要发现如下：

1. 机制1（探索）与机制 2（蒸馏）的裁剪参数对熵的影响相反，可通过联合调节来控制探索-蒸馏平衡，并缓解渐进性能退化。2. 与正确响应的渐进退化不同，当裁剪边界过于严格时，错误响应会突然崩溃：机制 3 放宽不足会导致重复输出，而机制 4 限制过严则会使响应长度趋近于零。为了实现稳定训练，不应对任何一种机制施加过多限制。3. 通过为每种机制精心调节裁剪边界，DISPO 在平衡探索与蒸馏的同时，避免了灾难性失败。DISPO 在 AIME’24 上取得了 61.04% 的成绩（而 CISPO 为 55.42%，DAPO 为 50.21%），并在各类基准测试和模型上取得了类似的提升。

提纲 第 2 节回顾相关工作，第 3 节则介绍必要的背景知识，为 DISPO 奠定基础。第 4 节介绍 DISPO，并阐述我们分析其四种策略更新机制的方法。第 5 节展示实验结果，将 DISPO 与基线进行比较，并详细考察每种策略更新机制。

2 相关工作

策略梯度方法的基础 REINFORCE 算法 [Williams，1992] 为强化学习（RL）中的策略梯度方法奠定了基础，证明了基于梯度的

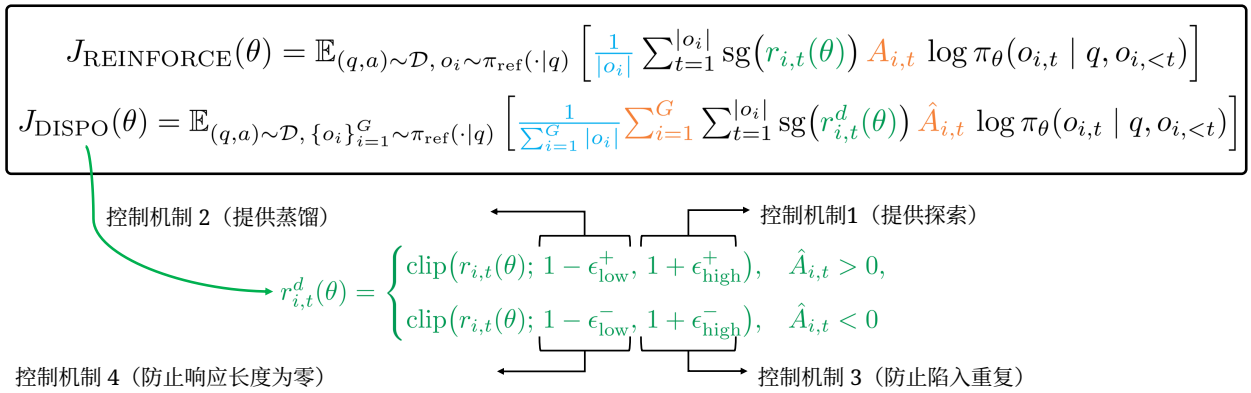


图 2：DISPO 通过以下组件扩展 REINFORCE：(i) 组相对优势估计、(ii) 令牌级归一化以及 (iii) 解耦重要性采样权重 $r^d_{i,\,t}(\theta)$。解耦重要性采样权重中的每个 ϵ 都控制着一种不同的策略更新机制。

随机策略优化方法。在此基础上，近端策略优化（PPO） [Schulman 等人，2017] 引入了裁剪代理目标来近似信任域，从而提高了稳定性。此后，PPO 已广泛应用于机器人 [Akkaya 等人、 2019], 游戏对弈 [Berner 等人、 2019], 以及连续控制/运动控制 [Heess 等人， 2017]等领域。

语言模型的强化学习 强化学习在语言模型中的应用始于基于人类反馈的强化学习（RLHF） [Christiano 等人，2017年，Stiennon 等人，2022年，Ouyang 等人，2022年，Lightman 等人， 2023], 该方法利用人类偏好来对齐模型输出，之后又扩展为 RLAIF [Lee 等人，2024年，Bai 等人， 2022] 以 AI反馈取代人类反馈。近年来，可验证奖励强化学习（RLVR）已应用于可自动验证的领域，尤其是数学推理 [Uesato 等人，2022年，Wang 等人， 2023], 从而无需人工标注。在 RLVR 中，已出现两类算法：PPO类方法（GRPO [Shao 等人， 2024], DAPO [Yu 等人， 2025] 及其他方法 [Liu 等人， 2025]）能够保持稳定，但收敛较慢；而 REINFORCE类方法则能更快地学习。CISPO [Chen 等人， 2025] 通过在 REINFORCE 中引入裁剪重要性采样，Arnal 等人 [2025] 则针对正向/负向奖励采用不同的学习率；Ahmadian 等人 [2024] 表明，经过仔细调优的 REINFORCE 可以媲美复杂算法。然而，这些方法存在训练不稳定性 [Zheng 等，2025年，Arnal 等人， 2025], 这促使我们开展本研究。

3 背景

我们考虑用于大语言模型推理的 RLVR 设置，其中 π_θ 表示一个由 θ 参数化的大语言模型策略，该策略会在词元空间 $\mathcal{T}$ 中随机预测下一个词元。给定由问答对 $(q,\ a)$ 构成的数据集 $\mathcal{D}$，其中 $q \in \mathcal{T}^*$， $a \in \mathcal{T}^*$，而 $\mathcal{T}^*$ 是词元序列空间，模型通过自回归采样词元来生成响应 o： $o = (o_1,\ o_2,\ \ldots\ldots,\ o_{|o|}) \in \mathcal{T}^{|o|}$，其中 $o_t \sim \pi_\theta(\cdot|q,\ o_{<t})$。根据真实答案 a，使用可验证奖励函数 $R(o,\ a) \in \{-1, 1\}$ 判断问题 q 的响应 o 是否正确。

GRPO 算法 对于每个问题 q，GRPO 从模型的冻结参考快照（记为 π_{ref}）中采样 G 个响应。目标函数采用近端策略优化的裁剪代理目标，如下所示：

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\big(r_{i,t}(\theta)\hat{A}_{i,t},\ r^g_{i,t}(\theta)\hat{A}_{i,t}\big) \right], \tag{1}$$

其中，$r_{i,t}(\theta) = \frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{ref}}(o_t|q,o_{<t})}$ 是重要性采样（IS）权重，而

$$r^g_{i,t}(\theta) = \text{clip}\big(r_{i,t}(\theta);\, 1-\epsilon,\, 1+\epsilon\big) \tag{2}$$

是其裁剪后的对应项。裁剪函数定义为 $\text{clip}(x; a, b) := \min\big(\max(x, a),\ b\big)$。第 i 个响应的组相对优势估计（在 t 上为常数）为

$$\hat{A}_{i,t} = \frac{R_i - \mu_G}{\sigma_G}, \tag{3}$$

其中，$R_i \in \{-1, 1\}$ 是二元验证器奖励，$\mu_G = \frac{1}{G}\sum_{j=1}^{G} R_j$ 和 $\sigma_G = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_j - \mu_G)}$ 分别是组均值和标准差。因此，$A_{i,t} > 0$ 表示正确回答，而 $A_{i,t} < 0$ 表示错误回答。PPO 式裁剪（公式 2）与 $\min(\cdot)$ 代理界限共同在 $r_{i,t}(\theta)$ 留下 $[1-\epsilon,\, 1+\epsilon]$ 时约束有效更新，从而保持类似信任域的稳定性。为简洁起见，我们省略了公式 1 中的 KL 正则化项。

DAPO 算法 DAPO [Yu 等人，2025] 通过以下方式扩展了 GRPO：（i）采用非对称裁剪界限 $\epsilon_{\text{high}} > \epsilon_{\text{low}}$ 以促进探索；（ii）采用动态采样，在更新前过滤掉信息量不足的组（例如全部正确或全部错误的组）；（iii）通过 $\frac{1}{\sum_{i=1}^{G}|o_i|}$ 进行令牌级归一化；以及（iv）引入超长惩罚项，以抑制过长的响应。DAPO 目标函数如下：

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \min\big(r_{i,t}(\theta)\hat{A}_{i,t},\ r^{c}_{i,t}(\theta)\hat{A}_{i,t}\big) \right], \tag{4}$$

其中裁剪后的重要性采样权重 $r^{c}_{i,t}(\theta)$ 定义为

$$r^{c}_{i,t}(\theta) = \text{clip}\big(r_{i,t}(\theta);\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}}\big). \tag{5}$$

与 GRPO 类似，DAPO 中的非对称窗口（ϵ_{low},ϵ_{high}）保持了类似信任域的稳定性。

REINFORCE 算法 离策略 REINFORCE 在词元级的目标可写为

$$J_{\text{REINFORCE}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, o_i\sim\pi_{\text{ref}}(\cdot|q)} \left[\frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \text{sg}\big(r_{i,t}(\theta)\big)\, A_{i,t} \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \right], \tag{6}$$

其中 $A_{i,t}$ 为优势，$\text{sg}(\cdot)$ 表示停止梯度算子，即重要性采样权重 $r_{i,t}(\theta)$ 仍对损失加权，但不参与求导。与近端策略优化/GRPO/DAPO 不同，REINFORCE 不施加信赖域约束；因此，即使重要性采样权重 $r_{i,t}(\theta)$ 显著偏离 1，梯度仍可流动。

CISPO 算法 CISPO 在离策略 REINFORCE 的基础上进行了扩展：(i) 组大小为 G 的组采样（与 GRPO 相同）；以及 (ii–v) DAPO 风格的组件：非对称裁剪界限（ϵ_{low},ϵ_{high}）、动态采样、令牌级归一化，以及超长惩罚项。CISPO 的目标为

$$J_{\text{CISPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\, \{o_i\}_{i=1}^{G}\sim\pi_{\text{ref}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \text{sg}\big(r^{c}_{i,t}(\theta)\big)\, \hat{A}_{i,t}\, \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}) \right] \tag{7}$$

其中，裁剪后的重要性采样权重 $r^{c}_{i,\ t}(\theta)$ 的定义见公式 5。与 REINFORCE 类似，CISPO 不施加信赖域约束：梯度可以通过每个词元传播，尽管其影响会受到裁剪。

值得注意的是，CISPO 对正确回答和错误回答中的词元都采用相同的非对称裁剪窗口。我们指出，这种一视同仁的处理忽略了 REINFORCE 四种策略更新机制之间本质上不同的优化动态，从而导致训练不稳定性和探索受限，这也促使我们在 DISPO 中提出解耦裁剪策略。

4 解耦重要性采样加权策略优化（DISPO）

DISPO 通过采用组相对优势估计和令牌级归一化扩展了 REINFORCE，与 CISPO 类似。此外，我们为正确响应和错误响应中的重要性采样（IS）权重分别设置了不同的裁剪边界。形式上，DISPO 的目标函数定义为

$$J_{\mathrm{DISPO}}(\theta)=\mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\mathrm{ref}}(\cdot|q)}\left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\mathrm{sg}\big(r_{i,t}^{d}(\theta)\big)\,\hat{A}_{i,t}\,\log\pi_\theta(o_{i,t}\mid q,o_{i,<t})\right] \tag{8}$$

其中，解耦重要性采样权重 $r_{i,t}^{d}(\theta)$ 由下式给出

$$r_{i,t}^{d}(\theta)=\begin{cases}\mathrm{clip}\big(r_{i,t}(\theta);\,1-\epsilon_{\mathrm{low}}^{+},\,1+\epsilon_{\mathrm{high}}^{+}\big), & \hat{A}_{i,t}>0,\\ \mathrm{clip}\big(r_{i,t}(\theta);\,1-\epsilon_{\mathrm{low}}^{-},\,1+\epsilon_{\mathrm{high}}^{-}\big), & \hat{A}_{i,t}<0.\end{cases} \tag{9}$$

这种解耦裁剪策略对 REINFORCE 中四种不同的策略更新机制提供了细粒度控制，使我们能够同时放大或抑制正确响应与错误响应的梯度。在第 4.1 节中，我们系统地分解这些机制，并提出消融方法，以分离它们各自对训练动态的影响。我们注意到，遵循 DAPO 和 CISPO，DISPO 同样在损失计算中纳入了动态采样和超长惩罚项。

梯度权重视图 我们可视化不同目标如何根据重要性比率 $r_{i,t}(\theta)$ 调节策略梯度的幅度。对于上文讨论的所有策略目标，梯度可以ˆ写成与 $\mathrm{sg}(w_{i,t}(\theta)r_{i,t}(\theta))\ \nabla_\theta\log\ \pi_\theta(y_{i,t}\mid q,y_{i,<t})A_{i,t}$ 成正比的函数，忽略长度归一化项。梯度权重 $w_{i,t}(\theta)$ 刻画了算法特定的门控／裁剪效应，应被解释为相对缩放因子（而非完整梯度），用以展示当 $r_{i,t}(\theta)$ 偏离 1 时，各方法如何衰减或保留更新。图 3 通过展示 DAPO、CISPO 和 DISPO 作为重要性采样权重函数的梯度权重曲线，提供了直观视图。我们观察到，PPO 式裁剪（DAPO）施加硬截断，一旦$r_{i,t}(\theta)$离开信赖域，便将更新权重设为零；而 REINFORCE 式变体更像软门控。DISPO 使用ˆ ˆ 符号相关的比率控制，对$A_{i,t}>0$ 和 $A_{i,t}<0$应用不同的门控曲线，从而产生作为 $r_{i,t}(\theta)$ 函数的非对称梯度加权。

4.1 分析策略更新机制

在离策略训练中，IS 权重$r_{i,\ t}(\theta)$继承自上一步更新后的策略，并会影响当前步骤的梯度。在首次更新时，$r_{i,\ t}(\theta)=1$，因此，式 8 中 DISPO 目标函数的梯度里每个词元的贡献形式如下

$$\nabla_\theta J\propto\hat{A}_{i,t}\nabla_\theta\log\pi_\theta(o_{i,t}\mid q,o_{i,<t}). \tag{10}$$

然而，随着训练的进行，$r_{i,\ t}(\theta)$可能偏离 1，变得大于或小于 1，这反映了当前策略 π_θ相对于参考策略 π_{ref}的偏离程度。

我们从两个维度对 DISPO 中的策略更新机制进行分解：（1）响应是否正确（ˆ$A_{i,\ t}>0$）或错误（ˆ$A_{i,\ t}<0$）；（2）IS 权重是放大（$r_{i,\ t}(\theta)>1$）还是抑制（$r_{i,\ t}(\theta)<1$）梯度。由此得到四种不同的更新机制，我们将在下文详细介绍：

机制1：放大的正向更新（ˆ$A_{i,\ t}>0$，$r_{i,\ t}(\theta)>1$）该机制涵盖正确回答中相较于参考策略、在先前更新期间概率有所提高的词元，如图4所示。我们可以将这些词元的梯度贡献写为

$$\nabla_\theta J\propto\mathrm{sg}\left(r_{i,t}^{d}(\theta)\right)\hat{A}_{i,t}\nabla_\theta\log\pi_\theta(o_{i,t}\mid q,o_{i,<t}). \tag{11}$$

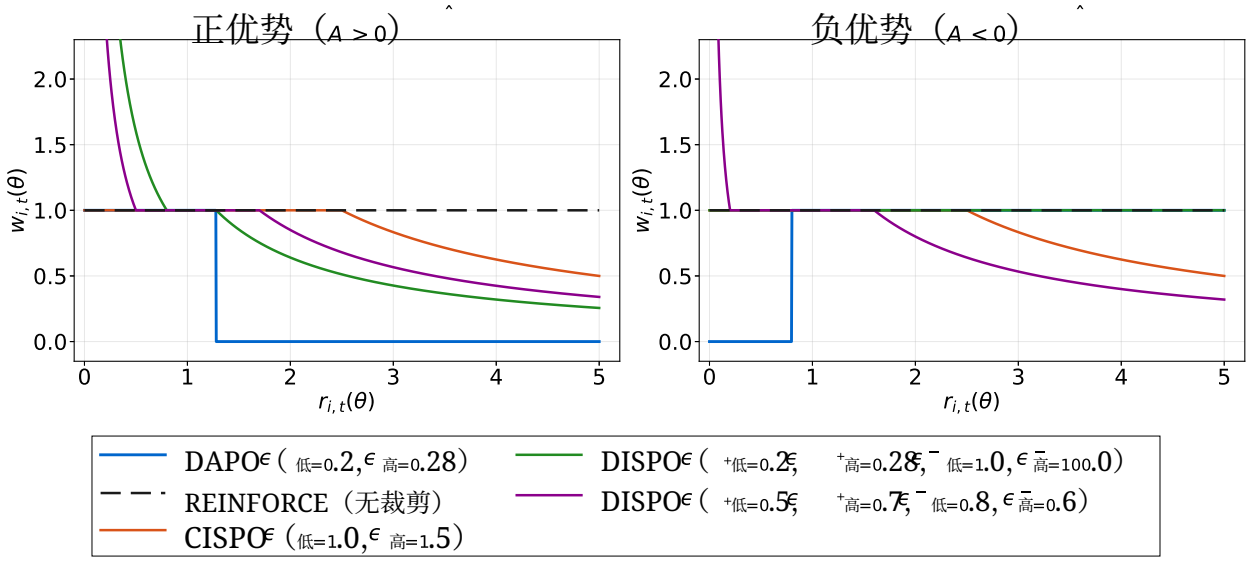


Figure 3：梯度权重 $w_{i,t}(\theta)$ 作为重要性采样权重 $r_{i,t}(\theta)$ 的函数

当 $r_{i,t}(\theta) > 1$ 时，该机制会将正向学习信号放大到超出式 10 中基线梯度的程度，从而强化模型已经学会偏好的词元。我们注意到，在式 9 中增大 $\epsilon^{+}_{\text{高}}$，可以允许 $r^{d}_{i,t}(\theta)$ 取更大的值，从而实现更大的放大；而将 $\epsilon^{+}_{\text{高}} = 0$ 设为较大值，则会将 $r^{d}_{i,t}(\theta)$ 截断为 1，并恢复为式 10 中的基线设置。我们的实验发现，机制1会提高平均词元级熵，成为训练期间推动探索的关键因素（第 5.3 节）。

机制 2：抑制正向更新（$\hat{A}_{i,t} > 0, r_{i,t}(\theta) < 1$）在此，正确响应中的词元相对于参考策略的概率有所下降，如图4所示。梯度表达式与式 11 相同，但由于 $r_{i,t}(\theta) < 1$，正向更新信号受到抑制。我们注意到，在式 9 中增大$\epsilon^{+}_{\text{低}}$会允许 $r^{d}_{i,t}(\theta)$ 取更小的值，从而产生更强的抑制；而将 $\epsilon^{+}_{\text{低}} = 0$ 设为该值则会把设定恢复为式 10 中的基线。我们观察到，机制 2 会降低平均词元级熵，作为一种蒸馏机制，在训练过程中巩固已学习模式（第 5.3 节）。

机制 3：放大的负向更新（$\hat{A}_{i,t} < 0$，$r_{i,t}(\theta) > 1$）此机制涵盖错误回答中那些在先前更新后相对于参考策略概率有所上升的词元，如图4所示。这些词元的梯度贡献会变得

$$\nabla_\theta J \propto \text{sg}\left(r^{d}_{i,t}(\theta)\right)\hat{A}_{i,t}\nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}). \tag{12}$$

当 $\hat{A}_{i,t} < 0$ 和 $r_{i,t}(\theta) > 1$时，此机制会放大负向学习信号，从而更有力地遗忘模型错误地学会偏好的词元。我们注意到，提高 $\epsilon^{-}_{\text{high}}$ 可以允许 $r^{d}_{i,t}(\theta)$ 取更大的值，从而实现更大幅度的放大；而将 $\epsilon^{-}_{\text{high}} = 0$ 设为高值，则会使梯度恢复为公式 10 中的基线。我们观察到，机制 3 中放大不足会导致重复诱发的崩溃，使模型无法充分遗忘错误模式（第 5.3 节）。

机制 4：抑制负向更新（$\hat{A}_{i,t} < 0$，$r_{i,t}(\theta) < 1$）最后，错误回答中概率相较于参考策略有所降低的词元属于这一机制，如图4所示。与公式 12 一样，负优势会促使模型进一步降低这些概率；但由于 $r_{i,t}(\theta) < 1$，遗忘信号会受到削弱。我们注意到，增大 $\epsilon^{-}_{\text{较低}}$，会因允许 $r^{d}_{i,t}(\theta)$ 取更小的值而加强抑制；而将 $\epsilon^{-}_{\text{较低}} = 0$，则会使梯度恢复为公式 10 中的基线。我们观察到，机制 4 中过度抑制会使响应长度趋近于零，这表明遗忘过于激进，进而破坏模型的生成能力（第 5.3 节）。

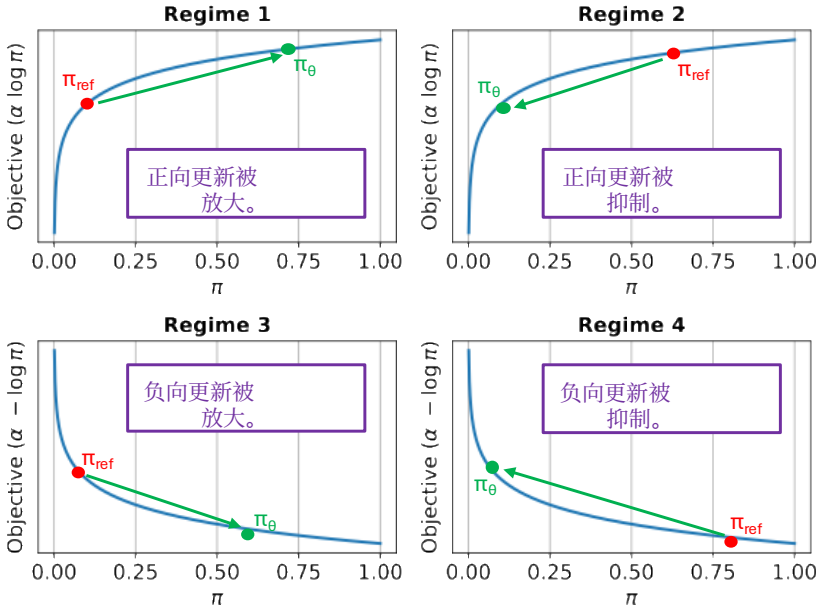


图4：DISPO 的四种策略更新机制。

为隔离各机制对训练动态的影响，我们通过改变公式 9 中的四个裁剪参数来设计受控消融实验。将任一参数设为零，即可禁用对应机制；正值则控制放大或抑制的强度。表1展示了我们的消融配置。对于机制 1 和机制 2，我们的基线将 $\epsilon^{+}_{\text{低}} = \epsilon^{+}_{\text{高}} = 0$，并将错误响应的梯度置零，这相当于在线监督微调（监督微调）。对于机制 3 和 4，我们的基线是同时启用这两种机制的 DISPO，因为它们对于稳定训练都不可或缺。为展示它们的必要性，我们分别通过设置 $\epsilon^{-}_{\text{高}} = 0$（机制 3）或 $\epsilon^{-}_{\text{低}} = 0$（机制 4）来单独禁用每种机制。

表1：用于分析离策略 REINFORCE 中更新机制的消融配置。对于机制 1–2，我们从在线监督微调基线出发，并分别单独启用每种机制。对于机制 3–4，我们从完整的 DISPO 配置出发，并分别单独禁用每种机制。

配置	回答类型	裁剪参数	启用的机制
从在线 *SFT* 基线开始（机制 *1* 和机制 *2* 的分析）			
在线监督微调基线	仅正确	$\epsilon^{+}_{\text{low}} = 0, \epsilon^{+}_{\text{高}} = 0$	无 ($\hat{r}_{i,t} = 1$)
+机制1	仅正确	$\epsilon^{+}_{\text{low}} = 0, \epsilon^{+}_{\text{高}} = 0.28$	增强正向
+机制1	仅正确	$\epsilon^{+}_{\text{low}} = 0, \epsilon^{+}_{\text{高}} = 10$	增强正向
+机制 2	仅正确	$\epsilon^{+}_{\text{低}} = 0.2, \epsilon^{+}_{\text{高}} = 0$	抑制正向
+机制 2	仅正确	$\epsilon^{+}_{\text{低}} = 1, \epsilon^{+}_{\text{高}} = 0$	抑制正向
从 *DISPO* 基线开始（机制 *3* 和 *4* 的分析）			
DISPO（完整）	正确 + 错误	$\epsilon^{-}_{\text{low}} = 1, \epsilon^{-}_{\text{high}} = 100$	所有状态
-机制 3	正确 + 错误	$\epsilon^{-}_{\text{low}} = 1, \epsilon^{-}_{\text{high}} = 0$	不含放大的负面项
-机制 4	正确 + 错误	$\epsilon^{-}_{\text{low}} = 0, \epsilon^{-}_{\text{高}} = 100$	不含抑制负例

5 结果与讨论

我们首先简要描述实验设置，随后给出将 DISPO 与基线方法进行比较的主要结果。接着，我们详细考察每一种策略更新机制，并强调这些洞见也指导了 DISPO 的设计。

5.1 实验设置

我们在多种模型规模和架构上，将 DISPO 与 PPO风格（DAPO）和 REINFORCE风格（CISPO）基线进行对比评估：Qwen3-8B-Base 和 Qwen3-14B-Base（均为稠密模型），以及 Qwen3-30B-A3B-Base（混合专家，激活参数为 3.3B）。所有模型均在 GSM8K、Math 和 Mathematics [Hendrycks 等人，2021b，Saxton等人，2019，Cobbe等人， 2021] 数据集上训练，并在五个数学推理基准上评估：AIME'24 [美国数学协会， 2025], AIME'25 [美国数学协会， 2025], AMC'23 [美国数学协会， 2024],MATH-500 [Hendrycks 等人，2021a], 以及 Minerva [Lewkowycz 等人， 2022]。我们使用 Qwen3-14B-Base 对策略更新机制开展消融研究。此外，我们在所有实验中均使用公式 3 中的优势函数形式，包括针对机制 1 和机制 2 的消融实验。有关基线方法、训练和评估的更多信息见附录B。

5.2 DISPO 与最先进方法的比较

DISPO 显著优于基线。表 2 展示了在所有受测模型上，DISPO 与基线在多个数学推理基准上的评估结果。在 AIME'24 数据集上，DISPO 实现了显著提升：Qwen3-14B 的准确率达到 61.04%，而 DAPO 为 50.21%，CISPO 为 55.42%；相比 DAPO 提高了 10.83 个百分点。在其他基准测试中也呈现出类似趋势，尤其是在竞赛级问题上，DISPO 的提升尤为明显（AIME'25：45.83%，而 DAPO 为 38.96%；AMC'23：92.03%，而 DAPO 为 87.66%）。无论是不同模型规模还是不同架构，提升都始终如一——从 8B 稠密模型到 30B 混合专家（MoE）变体均是如此。

表 2：不同模型规模和架构下 DISPO 与 DAPO、CISPO 的比较。"–"表示初始检查点的性能。每个模型和基准测试的最高值以**粗体**标出。

基准测试 (**Avg@16**)	**强化学习算法**			
	–	DAPO	CISPO	DISPO
模型：Qwen3-14B-Base				
AIME'24	3.96	50.21	55.42	**61.04**
AIME'25	1.67	38.96	40.83	**45.83**
AMC'23	22.19	87.66	89.84	**92.03**
MATH-500	73.91	91.89	93.15	**94.61**
Minerva	36.10	45.22	45.66	**46.78**
模型：Qwen3-30B-A3B-Base（混合专家）				
AIME'24	0.83	45.83	48.54	**53.75**
AIME'25	1.04	31.25	35.62	**38.12**
AMC'23	12.50	83.75	87.03	**89.38**
MATH-500	73.54	90.56	91.18	**92.59**
Minerva	30.35	43.45	44.26	**44.60**
模型：Qwen3-8B-Base				
AIME'24	4.58	38.33	43.54	**45.00**
AIME'25	2.71	27.50	30.83	**31.46**
AMC'23	24.69	80.94	80.94	**87.66**
MATH-500	71.22	91.25	90.74	**92.34**
Minerva	30.95	**44.85**	43.52	44.60

DISPO 在保持训练稳定性的同时，平衡探索与蒸馏。图5展示了 DISPO 和基线方法在 AIME'24 上的学习曲线。熵曲线（下方面板）揭示了明显的差异。我们观察到，CISPO 的熵在整个训练过程中持续下降，而 DISPO 的熵则有所上升。这一差异意义重大，因为较高的熵意味着更多的词元级探索 [Wang

等， 2025], 这也解释了 DISPO 为何表现更优。DAPO 的熵也呈现类似的上升趋势，但其性能明显较低，因为它采用的是词元截断的近端策略优化，而非 REINFORCE。具体而言，它会丢弃参考似然较低、但作为熵变化关键驱动因素的词元，例如“but”“aha”和“since”；先前研究已指出了这些词元的重要性 [Chen 等，2025，Wang 等， 2025]。附录 C 展示了其他模型（Qwen3-30B-A3B-Base 和 Qwen3-8B-Base）的学习曲线，它们呈现出类似的模式。Qwen3-14B-Base 的响应长度曲线也见附录 C。

关于裁剪边界，我们采用 DAPO 的默认值： $\epsilon_{\text{low}} = 0.2$ 和 $\epsilon_{\text{high}} = 0.28$。由于原始 CISPO 的裁剪边界未公开，我们使用 $\epsilon_{\text{low}} = 1$ 和 $\epsilon_{\text{high}} = 100$实现了该方法，并且在所有评估模型上都证明了其稳定性。在 CISPO 中使用 $\epsilon_{\text{low}} < 1$ 或 $\epsilon_{\text{high}} < 100$ 都会导致性能突然崩溃。请注意， $\epsilon = 100$ 实际上禁用了裁剪，同时仍可防止重要性权重变为无穷大。对于 DISPO，我们设置 $\epsilon^{+}_{\text{低}} = 0.2$、 $\epsilon^{+}_{\text{高}} = 10$、 $\epsilon^{-}_{\text{低}} = 1$以及 $\epsilon^{-}_{\text{高}} = 100$。在第 5.3 节中，我们将通过相应的策略更新机制，考察这四个 ϵ 参数如何影响训练动态，说明我们选择 DISPO 超参数的依据，并分析 Zheng 等人此前报告的 CISPO 崩溃现象 [Zheng et al., 2025]。

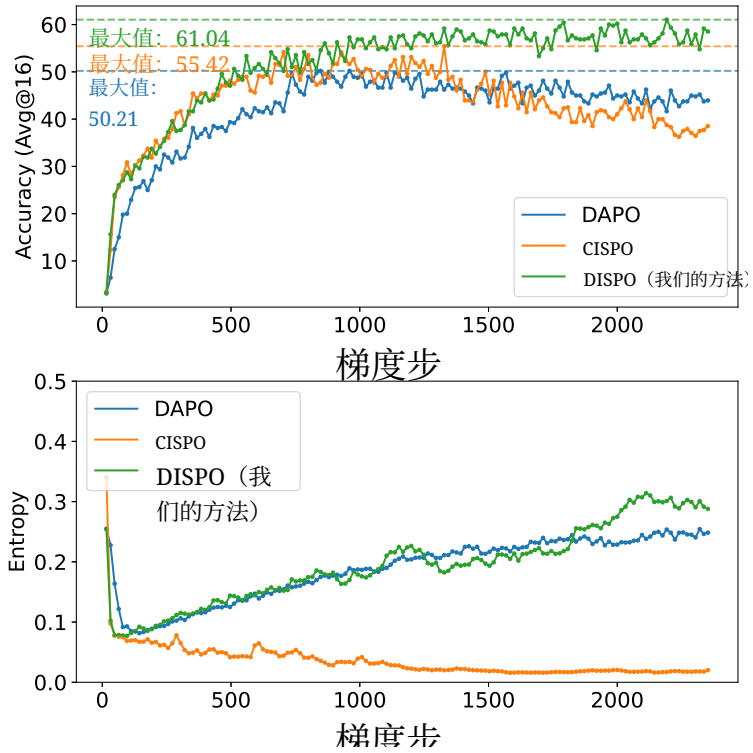


图5：DAPO、CISPO 和 DISPO 的准确率与熵曲线。

5.3 离策略 REINFORCE 中的策略更新机制分析

提高 ϵ^{+} 机制1中的高值能够增强探索。图6展示了在保持 $\epsilon^{+}_{\text{低}} = 0$的同时，通过调整不同的 $\epsilon^{+}_{\text{高}}$值来允许$r_{i,t} > 1$所产生的影响。在线监督微调基线（蓝色）表现出稳定但缓慢的准确率提升，而采用 $\epsilon^{+}_{\text{高}} > 0$ 的两个模型则展现出更快的提升，表明训练效率有所提高。熵曲线揭示了其背后的机制：基线在整个训练过程中熵保持不变，而采用 $\epsilon^{+}_{\text{高}} > 0$ 的配置则表现出熵不断增加，反映出探索逐步增强。正如我们在第5.2节中讨论的，在正确响应中放大低参考似然词元的梯度，可以更有效地利用词元并增强探索。

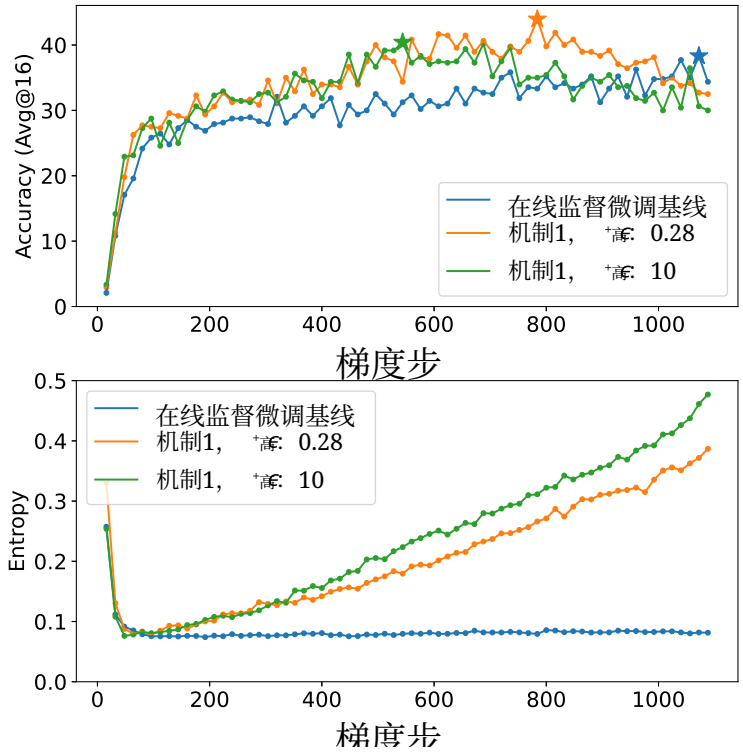


图 6：机制1运行的准确率和熵曲线。 ⋆ 标记最大值

e.

过度探索会导致性能逐渐退化。随着训练阶段推进，探索的增加最终会产生负面影响。采用 $\epsilon^{+}_{\text{高}} > 0$ 的两个模型在准确率达到峰值后均出现下降，因为过度探索会导致模型采样越来越不可能出现的词元，从而损害推理连贯性。这揭示了一个关键的权衡：放大通过有益的探索提高学习效率，但过度放大会导致失控探索，并使训练后期的性能逐步下降。与采用 $\epsilon^{+}_{\text{高}} = 10$的模型相比，采用 $\epsilon^{+}_{\text{高}} = 0.28$ 的模型能够进行更可控的探索——熵增长得更为平缓，因此可以在性能退化前训练更长时间，并取得更高的峰值准确率。这表明，谨慎调整 $\epsilon^{+}_{\text{高}}$能够在保留效率优势的同时，降低失控探索的风险。我们注意到，图5中的DISPO在延长训练后也可能出现准确率下降，这是由机制1导致的，不过计算约束使我们无法验证这一点。值得注意的是，在机制1中，不稳定性表现为逐渐退化，而非突然崩溃。

提高 $\epsilon^{+}_{\text{机制 2 中的低}}$值可改善蒸馏效果。图 7 展示了在保持 $\epsilon_{\text{high}} = 0$ 不变的情况下改变 $\epsilon^{+}_{\text{低}}$值的影响。我们发现，允许 $r_{i,\,t} < 1$ 与在线监督微调基线相比可带来效率增益。如图 7 所示，机制 2 会降低熵——这与机制 1 的探索行为相反——表明模型会主动剪枝其词元词表。这种熵的降低体现了一种蒸馏机制：通过削弱概率降低的词元（$r_{i,\,t} < 1$）所接收的学习信号，模型甚至会在正确响应中滤除可靠性较低的词元。尽管某些被抑制的词元可能包含有用模式，但其中很可能也包含噪声或次优求解路径。这种选择性强化通过聚焦于构成核心解题策略的高概率词元，而不是不加区分地强化正确解中的所有词元，加速了学习。比较 $\epsilon^{-}_{\text{低}} = 0.2$ 值与 1。0，前者通过逐步降低熵实现了更可控的蒸馏，并取得了略高的峰值准确率。

蒸馏单独使用的益处有限，但能有效抵消过度探索。虽然机制 2 能提升效率（图 7），但与机制1由探索驱动的性能（图 6）相比，这些提升仍然有限。机制 2 中熵的下降表明

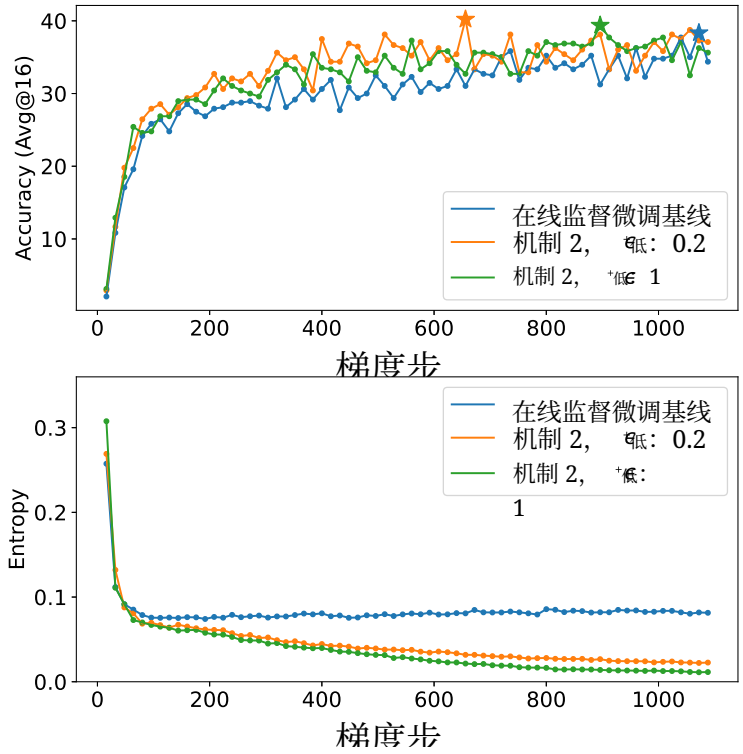


图 7：机制 2 各次运行的准确率与熵曲线。 ⋆ 标记最大 valu e.

模型过早收敛到有限的解题策略，以牺牲探索能力换取一致性，并可能错过需要低概率词元的创新方法。这解释了为何纯蒸馏会导致较低的峰值准确率——要取得更优的推理性能，探索比巩固更为关键。两种机制的响应长度曲线见附录 C。值得注意的是，当机制 1 和机制 2 同时运行时（附录 D），二者相互竞争的熵效应形成平衡的探索—蒸馏动态，其准确率高于任一机制单独运行时的准确率，这表明蒸馏的主要价值在于调节探索，而非作为独立策略发挥作用。

在机制 3 中，设置 $\epsilon^{-}_{\text{高}} > 0$ 是防止重复崩溃的必要条件。图 8 显示，与完整 DISPO（蓝色）相比，禁用机制 3（$\epsilon^{-}_{\text{高}} = 0$，橙色）会在训练早期导致准确率突然崩溃和响应长度激增。机制 3 针对概率超过参考策略的错误词元（$r_{i,t} > 1$）。若没有机制 3 的放大，这些词元就得不到足以实现遗忘的负梯度——因为 $|\nabla_\theta \log \pi_\theta|$ 会随着 $\pi_\theta \to 1$ 自然减小，高概率词元的梯度本就较弱。本质上，模型失去了“遗忘”其错误的能力，因为遗忘信号变得过弱，无法克服先前更新所带来的强化。将 $\epsilon^{-}_{\text{高}} = 0$ 会使 $r_{i,t}$ 的上限为 1，进一步削弱这些梯度，并妨碍充分惩罚。因此，模型会反复生成这些错误的高概率词元，从而造成图 8（下）中重复导致的长度激增。附录 C 中的图 13 展示了一个代表性示例。我们推测，这是此前报告的 CISPO 崩溃的原因之一 [Zheng 等人， 2025]——我们的实验证实，$\epsilon^{-}_{\text{高}} < 1$ 会使 CISPO 在我们的实验设置中以类似的重复模式失败。

在机制 4 中，设置 $\epsilon^{-}_{\text{较低的}} > 0$ 可以防止因过度遗忘而导致的长度崩溃。图 8 显示，禁用机制 4（$\epsilon^{-}_{\text{较低的}} = 0$，绿色）时，准确率和响应长度都会迅速恶化。机制 4 对概率低于参考策略的错误词元进行处理（$r_{i,t} < 1$）。这些低概率词元自然会获得较强的负梯度，因为随着 $\pi_\theta \to 0$ 增大，$|\nabla_\theta \log \pi_\theta|$ 也会增大。设置 $\epsilon^{-}_{\text{较低的}} = 0$ 会移除

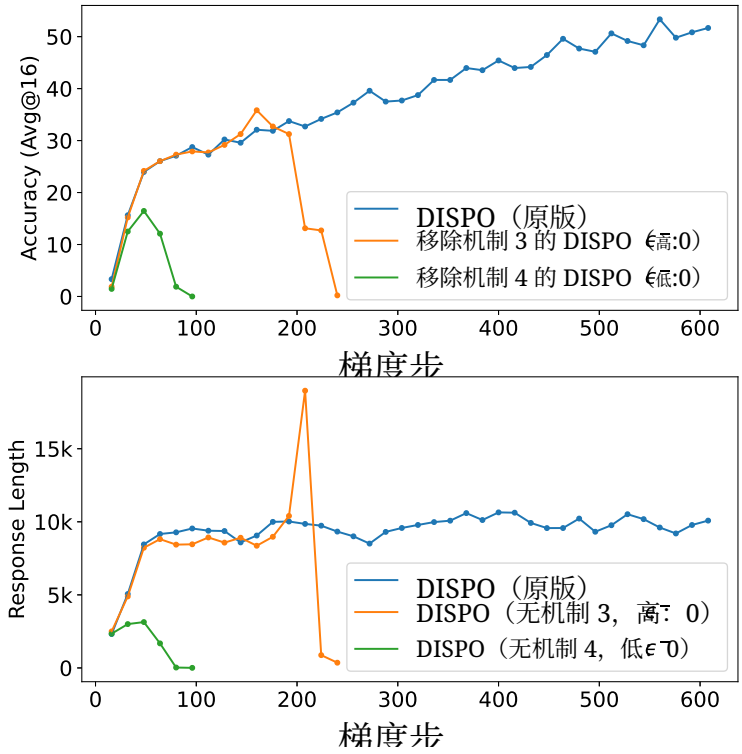


图 8 ：禁用机制 3 和 4 的运行结果中的准确率和响应长度曲线。

抑制——允许$r_{i,\ t}$ 低于 1 会放大这些本已很强的梯度，导致过度惩罚，进而使响应长度趋近于零（图 8 底部）。这种过度惩罚实际上教会模型在生成时“放弃”，而不是学习正确的模式。因此，机制 4 的梯度抑制可以防止对本已不太可能出现的词元进行过度纠正。我们推测，这也是 CISPO 崩溃的另一个原因 [Zheng 等人， 2025], 因为我们的实验显示， ϵ^{-} 较低的 < 100 会导致 CISPO 出现类似的长度崩溃。

6 结论

我们提出了 DISPO，这是一种简单而有效的离策略 REINFORCE 改进方法，通过对正确响应和错误响应的重要性采样权重分别进行裁剪，实现了二者的解耦。通过对由此形成的四种策略更新机制进行系统性消融实验，我们发现了离策略 REINFORCE 风格方法中各不相同的失效模式。结果表明，恰当平衡这些机制有助于促进探索-蒸馏平衡，同时保持训练稳定性，并取得优于现有方法的性能。

参考文献

A. Ahmadian、C. Cremer、M. Gallé、M. Fadaee、J. Kreutzer、O. Pietquin、A. Üstün 和 S. Hooker。回归基础：重新审视用于从人类反馈中学习的大语言模型 REINFORCE 风格优化，2024年。网址：https://arxiv.org/abs/2402.14740。

I. Akkaya、M. Andrychowicz、M. Chociej、M. Litwin、B. McGrew、A. Petron、A. Paino、M. Plappert、G. Powell、R. Ribas、J. Schneider、N. Tezak、J. Tworek、P. Welinder、L. Weng、Q. Yuan、W. Zaremba 和 L. Zhang。用机器人手解魔方，2019。网址：https://arxiv.org/abs/1910.07113。

Anthropic。Claude 3.7 Sonnet，2024年。URL https://www.anthropic.com/claude/sonnet .

C. Arnal、G. Narozniak、V. Cabannes、Y. Tang、J. Kempe 和 R. Munos。面向离策略强化学习的非对称强化：平衡正向和负向奖励，2025年。URL https://arxiv.org/abs/ 2506.20520。Y. Bai、S. Kadavath、S. Kundu、A. Askell、J. Kernion、A. Jones、A. Chen、A. Goldie、A. Mirhoseini、C. Mckinnon、C. Chen、C. Olsson、C. Olah、D. Hernandez、D. Drain、D. Ganguli、D. Li、E. Tran-Johnson、E. Perez、J. Kerr、J. Mueller、J. Ladish、J. Landau、K. Ndousse、K. Lukosuite、L. Lovitt、M. Sellitto、N. Elhage、N. Schiefer、N. Mercado、N. Dassarma、R. Lasenby、R. Larson、S. Ringer、S. Johnston、S. Kravec、E. Showk、S. Fort、T. Lanham、T. Telleen-Lawton、T. Conerly、T. Henighan、T. Hume、S. Bowman、Z. Hatfield-Dodds、B. Mann、D. A modei、N. Joseph、S. Mccandlish、T. Brown 和 J. Kaplan。宪法式人工智能：来自 AI反馈的无害性，2022年。URL https://arxiv.org/pdf/2212.08073。C. Berner、G. Brockman、B. Chan、V. Cheung、P. Dębiak、C. Dennison、D. Farhi、Q. Fischer、S. Hashme、C. Hesse、R. Józefowicz、S. Gray、C. Olsson、J. Pachocki、M. Petrov、H. P. d. O. Pinto、J. Raiman、T. Salimans、J. Schlatter、J. Schneider、S. Sidor、I. Sutskever、J. Tang、F. Wolski 和 S. Zhang。借助大规模深度强化学习玩 Dota 2。*arXiv:1912*。*06680* [计算机科学，统计学]，12 2019。URL https: //arxiv.org/abs/1912.06680。A. Chen、A. Li、B. Gong、B. Jiang、B. Fei、B. Yang、B. Shan、C. Yu、C. Wang、C. Zhu、C. Xiao、C. Du、C. Zhang、C. Qiao、C. Zhang、C. Du、C. Guo、D. Chen、D. Ding、D. Sun、D. Li、E. Jiao、H. Zhou、H. Zhang、H. Ding、H. Sun、H. Feng、H. Cai、H. Zhu、J. Sun、J. Zhuang、J. Cai、J. Song、J. Zhu、J. Li、J. Tian、J. Liu、J. Xu、J. Yan、J. Liu、J. He、K. Feng、K. Yang、K. Xiao、L. Han、L. Wang、L. Yu、L. Feng、L. Li、L. Zheng、L. Du、L. Yang、L. Zeng、M. Yu、M. Tao、M. Chi、M. Zhang、M. Lin、N. Hu、N. Di、P. Gao、P. Li、P. Zhao、Q. Ren、Q. Xu、Q. Li、Q. Wang、R. Tian、R. Leng、S. Chen、S. Chen、S. Shi、S. Weng、S. Guan、S. Yu、S. Li、S. Zhu、T. Li、T. Cai、T. Liang、W. Cheng、W. Kong、W. Li、X. Chen、X. Song、X. Luo、X. Su、X. Li、X. Han、X. Hou、X. Lu、X. Zou、X. Shen、Y. Gong、Y. Ma、Y. Wang、Y. Shi、Y. Zhong、Y. Duan、Y. Fu、Y. Hu、Y. Gao、Y. Fan、Y. Yang、Y. Li、Y. Hu、Y. Huang、Y. Li、Y. Xu、Y. Mao、Y. Shi、Y. Wenren、Z. Li、Z. Li、Z. Tian、Z. Zhu、Z. Fan、Z. Wu、Z. Xu、Z. Yu、Z. Lyu、Z. Jiang、Z. Gao、Z. Wu、Z. Song 和 Z. Sun。Minimax-M1：利用闪电注意力高效扩展测试时计算，2025年。URL https://arxiv.org/abs/2506.13585。P. Christiano、J. Leike、T. B. Brown、M. Martic、S. Legg 和 D. Amodei。基于人类偏好的深度强化学习，2017年。URL https://arxiv.org/abs/1706.03741。K. Cobbe、V. Kosaraju、M. Bavarian、M. Chen、H. Jun、L. Kaiser、M. Plappert、J. Tworek、J. Hilton、R. Nakano、C. Hesse 和 J. Schulman。训练验证器来解决数学应用题。*arXiv* 预印本 *arXiv:2110.14168*，2021。D. Guo、D. Yang、H. Zhang、J. Song、R. Zhang、R. Xu、Q. Zhu、S. Ma、P. Wang、X. Bi、X. Zhang、X. Yu、Y. Wu、W. Z. F、Z. Gou、Z. Shao、Z. Li、Z. Gao、A. Liu、B. Xue、B. Wang、B. Wu、B. Feng、C. Lu、C. Zhao、C. Deng、C. Zhang、C. Ruan、D. Dai、D. Chen、D. Ji、E. Li、F. Lin、F. Dai、F. Luo、G. Hao、G. Chen、G. Li、H. Zhang、H. Bao、H. Xu、H. Wang、H. Ding、H. Xin、H. Gao、H. Qu、H. Li、J. Guo、J. Li、J. Wang、J. Chen、J. Yuan、J. Qiu、J. Li、C. J. L、J. Ni、J. Liang、J. Chen、K. Dong、K. Hu、K. Gao、K. Guan、K. Huang、K. Yu、L. Wang、L. Zhang、L. Zhao、L. Wang、L. Zhang、L. Xu、L. Xia、M. Zhang、M. Zhang、M. Tang、M. Li、M. Wang、M. Li、N. Tian、P. Huang、P. Zhang、Q. Wang、Q. Chen、Q. Du、R. Ge、R. Zhang、R. Pan、R. Wang、C. R. J、J. R. L、R. Chen、S. Lu、S. Zhou、S. Chen、S. Ye、S. Wang、S. Yu、S. Zhou、S. Pan、L. S. S、S. Zhou、S. Wu、S. Ye、T. Yun、T. Pei、T. Sun、T. Wang、W. Zeng、W. Zhao、W. Liu、W. Liang、W. Gao、W. Yu、W. Zhang、X. W. L、W. An、X. Liu、X. Wang、X. Chen、X. Nie、X. Cheng、X. Liu、X. Xie、X. Liu、X. Yang、X. Li、X. S u、X. Lin、L. X. Q、X. Jin、X. Shen、X. Chen、X. Sun、X. Wang、X. Song、X. Zhou、X. Wang、X. Shan、L. Y. K、W. Y. Q、W. Y. X、Y. Zhang、Y. Xu、Y. Li、Y. Zhao、Y. Sun、Y. Wang、Y. Yu、Y. Zhang、Y. Shi、Y. Xiong、Y. He、Y. Piao、Y. Wang、Y. Tan、Y. Ma、Y. Liu、Y. Guo、Y. Ou、Y. Wang、Y. Gong、Y. Zou、Y. He、Y. Xiong、Y. Luo、Y. You、Y. Liu、Y. Zhou、Z. Y. X、Y. Xu、Y. Huang、Y. Li、Y. Zheng、Y. Zhu、Y. Ma、Y. Tang、Y. Zha、Y. Yan、R. Z. Z、Z. Ren、Z. Sha、Z. Fu、Z. Xu、Z. Xie、Z. Zhang、Z. Hao、Z. Ma、Z. Yan、Z. Wu、Z. Gu、Z. Zhu、Z. Liu、Z. Li、Z. Xie、Z. Song、Z. Pan、Z. Huang、Z. Xu、Z. Zhang 和 Z. Zhang。DeepSeek-R1：通过强化学习激励大语言模型的推理能力，2025年。URL https://arxiv.org/abs/2501.12948。

N. Heess、D. TB、S. Sriram、J. Lemmon、J. Merel、G. Wayne、Y. Tassa、T. Erez、Z. Wang、S. M. A. Eslami、M. Riedmiller 和 D. Silver。《丰富环境中的运动行为涌现》。*arXiv:1707.02286[*计算机科学*]*, 07 2017。网址：https://arxiv.org/abs/1707.02286。D. Hendrycks、C. Burns、S. Kadavath、A. Arora、S. Basart、E. Tang、D. Song 和 J. Steinhardt。《使用 Math 数据集衡量数学问题求解能力》。*arXiv:2103.03874 [*计算机科学*]*, 11 2021a。网址：https: //arxiv.org/abs/2103.03874。D. Hendrycks、C. Burns、S. Kadavath、A. Arora、S. Basart、E. Tang、D. Song 和 J. Steinhardt。《使用 Math 数据集衡量数学问题求解能力》。神经信息处理系统大会，2021b。A. Jaech、A. Kalai、A. Lerer、A. Richardson、A. El-Kishky、A. Low、A. Helyar、A. Madry、A. Beutel、A. Carney、A. Iftimie、A. Karpenko、A. T. Passos、A. Neitz、A. Prokofiev、A. Wei、A. Tam、A. Bennett、A. Kumar、A. Saraiva、A. Vallone、A. Duberstein、A. Kondrich、A. Mishchenko、A. Applebaum、A. Jiang、A. Nair、B. Zoph、B. Ghorbani、B. Rossen、B. Sokolowsky、B. Barak、B. McGrew、B. Minaiev、B. Hao、B. Baker、B. Houghton、B. McKinzie、B. Eastman、C. Lugaresi、C. Bassin、C. Hudson、C. M. Li、d. Bourcy、C. Voss、C. Shen、C. Zhang、C. Koch、C. Orsinger、C. Hesse、C. Fischer、C. Chan、D. Roberts、D. Kappler、D. Levy、D. Selsam、D. Dohan、D. Farhi、D. Mely、D. Robinson、D. Tsipras、D. Li、D. Oprica、E. Free man、E. Zhang、E. Wong、E. Proehl、E. Cheung、E. Mitchell、E. Wallace、E. Ritter、E. Mays、F. Wang、F. P. Such、F. Raso、F. Leoni、F. Tsimpourlas、F. Song、v. Lohmann、F. Sulit、G. Salmon、G. Parascandolo、G. Chabot、G. Zhao、G. Brockman、G. Leclerc、H. Salman、H. Bao、H. Sheng、H. Andrin、H. Bagherinezhad、H. Ren、H. Lightman、H. W. Chung、I. Kivlichan、I. O'Connell、I. Osband、I. C. Gilaberte、I. Akkaya、I. Kostrikov、I. Sutskever、I. Kofman、J. Pachocki、J. Lennon、J. Wei、J. Harb、J. Twore、J. Feng、J. Yu、J. Weng、J. Tang、J. Yu、J. Q. Candela、J. Palermo、J. Parish、J. Heidecke、J. Hallman、J. Rizzo、J. Gordon、J. Uesato、J. Ward、J. Huizinga、J. Wang、K. Chen、K. Xiao、K. Singhal、K. Nguyen、K. Cobbe、K. Shi、K. Wood、K. Rimbach、K. Gu-Lemberg、K. Liu、K. Lu、K. Stone、K. Yu、L. Ahmad、L. Yang、L. Liu、L. Maksin、L. Ho、L. Fedus、L. Weng、L. Li、L. McCallum、L. Held、L. Kuhn、L. Kondraciuk、L. Kaiser、L. Metz、M. Boyd、M. Trebacz、M. Joglekar、M. Chen、M. Tintor、M. Meyer、M. Jones、M. Kaufer、M. Schwarzer、M. Shah、M. Yatbaz、M. Y. Guan、M. Xu、M. Yan、M. Glaese、M. Chen、M. Lampe、M. Malek、M. Wang、M. Fradin、M. McClay、M. Pavlov、M. Wang、M. Wang、M. Murati、M. Bavarian、M. Rohaninejad、N. McAleese、N. Chowdhury、N. Chowdhury、N. Ryder、N. Tezak、N. Brown、O. Nachum、O. Boiko、O. Murk、O. Watkins、P. Chao、P. Ashbourne、P. Izmailov、P. Zhokhov、R. Dias、R. Arora、R. Lin、R. G. Lopes、R. Gaon、R. Miyara、R. Leike、R. Hwang、R. Garg、R. Brown、R. James、R. Shu、R. Cheu、R. Greene、S. Jain、S. Altman、S. Toizer、S. Toyer、S. Miserendino、S. Agarwal、S. Hernandez、S. Baker、S. McKinney、S. Yan、S. Zhao、S. Hu、S. Santurkar、S. R. Chaudhuri、S. Zhang、S. Fu、S. Papay、S. Lin、S. Balaji、S. Sanjeev、S. Sidor、T. Broda、A. Clark、T. Wang、T. Gordon、T. Sanders、T. Patwardhan、T. Sottiaux、T. Degry、T. Dimson、T. Zheng、T. Garipov、T. Stasi、T. Bansal、T. Creech、T. Peterson、T. Eloundou、V. Qi、V. Kosaraju、V. Monaco、V. Pong、V. Fomenko、W. Zheng、W. Zhou、W. McCabe、W. Zaremba、Y. Dubois、Y. Lu、Y. Chen、Y. Cha、Y. Bai、Y. He、Y. Zhang、Y. Wang、Z. Shao 和 Z. Li。《OpenAI o1 系统卡》，2024年。网址：https://arxiv.org/abs/2412.16720。

N. Lambert、J. Morrison、V. Pyatkin、S. Huang、H. Ivison、F. Brahman、M. Lester、A. Liu、N. Dziri、S. Lyu、Y. Gu、S. Malik、V. Graf、J. D. Hwang、J. Yang、R. L. Bras、O. Tafjord、C. Wilhelm、L. Soldaini、N. A. Smith、Y. Wang、P. Dasigi 和 H. Hajishirzi。《Tulu 3：推动开放语言模型后训练的前沿》，2024年。网址：https://arxiv.org/abs/2411.15124。

H. Lee、S. Phatale、H. Mansoor、T. Mesnard、J. Ferret、K. Lu、C. Bishop、E. Hall、V. Carbune、A. Rastogi、S. Prakash 和 G. Research。《RLAIF：利用 AI反馈扩展基于人类反馈的强化学习》，2024年。网址：https://arxiv.org/pdf/2309.00267。

A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari 和 V. Misra。使用语言模型解决定量推理问题。*arXiv:2206.14858 [*计算机科学*]*, 06 2022。网址：https://arxiv.org/abs/2206.14858。

H. Lightman、V. Kosaraju、Y. Burda、H. Edwards、B. Baker、T. Lee、J. Leike、J. Schulman、I. Sutskever 和 K. Cobbe。《让我们逐步验证》，2023年5月。网址：https://arxiv.org/abs/2305.20050。

Z. Liu、C. Chen、W. Li、P. Qi、T. Pang、C. Du、W. S. Lee 和 M. Lin。《理解 r1-zero 类训练：批判性视角》，2025年。网址：https://arxiv.org/abs/2503.20783。美国数学协会。美国数学竞赛——美国数学协会，2024年。网址：https: //maa.org/student-programs/amc/。美国数学协会。美国数学协会邀请赛——美国数学协会，2025年。网址：https://maa.org/ maa-invitational-competitions/。L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. L. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、K. Slama、A. Ray、J. Schulman、J. Hilton、F. Kelton、L. Miller、M. Simens、A. Askell、P. Welinder、P. Christiano、J. Leike 和 R. Lowe。《通过人类反馈训练语言模型遵循指令》。*arXiv:2203.02155 [计算机科学],*03 2022年。网址：https://arxiv.org/abs/2203.02155。D. Saxton、E. Grefenstette、F. Hill 和 P. Kohli。《神经模型数学推理能力分析》。*arXiv* 预印本 *arXiv:1904.01557*，2019年。J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov。《近端策略优化算法》，2017年8月。网址：https://arxiv.org/abs/1707.06347。Z. Shao、P. Wang、Q. Zhu、R. Xu、J. Song、X. Bi、H. Zhang、M. Zhang、L. Y. K、Y. Wu 和 D. Guo。《DeepSeekMath：推动开放语言模型数学推理能力的极限》，2024年。网址：https://arxiv.org/abs/2402.03300。N. Stiennon、L. Ouyang、J. Wu、D. M. Ziegler、R. Lowe、C. Voss、A. Radford、D. Amodei 和 P. Christiano。《通过人类反馈学习摘要》。*arXiv:2009.01325 [计算机科学],* 02 2022。网址：https://arxiv.org/ abs/2009.01325。J. Uesato、N. Kushman、R. Kumar、F. Song、N. Siegel、L. Wang、A. Creswell、G. Irving 和 I. Higgins。《利用基于过程和结果的反馈解决数学应用题》，2022年11月。网址：https://arxiv. org/abs/2211.14275。P. Wang、L. Li、Z. Shao、X. R. X、D. Dai、Y. Li、D. Chen、Y. Wu 和 Z. Sui。《Math-Shepherd：无需人工标注，逐步验证并强化大语言模型》，2023年。网址：https://arxiv.org/abs/2312.08935。S. Wang、L. Yu、C. Gao、C. Zheng、S. Liu、R. Lu、K. Dang、X. Chen、J. Yang、Z. Zhang、Y. Liu、A. Yang、A. Zhao、Y. Yue、S. Song、B. Yu、G. Huang 和 J. Lin。《超越二八定律：高熵少数词元推动大语言模型推理的有效强化学习》，2025年。网址：https://arxiv.org/abs/2506.01939。R. J. Williams。《用于连接主义强化学习的简单统计梯度跟随算法》。机器学习，8：229–256，1992年5月。doi：10.1007/bf00992696。A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv、C. Zheng、D. Liu、F. Zhou、F. Huang、F. Hu、H. Ge、H. Wei、H. Lin、J. Tang、J. Yang、J. Tu、J. Zhang、J. Yang、J. Yang、J. Zhou、J. Zhou、J. Lin、K. Dang、K. Bao、K. Yang、L. Yu、L. Deng、M. Li、M. Xue、M. Li、P. Zhang、P. Wang、Q. Zhu、R. Men、R. Gao、S. Liu、S. Luo、T. Li、T. Tang、W. Yin、X. Ren、X. Wang、X. Zhang、X. Ren、Y. Fan、Y. Su、Y. Zhang、Y. Zhang、Y. Wan、Y. Liu、Z. Wang、Z. Cui、Z. Zhang、Z. Zhou 和 Z. Qiu。《Qwen3 技术报告》，2025年。网址：https://arxiv.org/abs/2505.09388。Q. Yu、Z. Zhang、R. Zhu、Y. Yuan、X. Zuo、Y. Yue、W. Dai、T. Fan、G. Liu、L. Liu、X. Liu、H. Lin、Z. Lin、B. Ma、G. Sheng、Y. Tong、C. Zhang、M. Zhang、W. Zhang、H. Zhu、J. Zhu、J. Chen、J. Chen、C. Wang、H. Yu、Y. Song、X. Wei、H. Zhou、J. Liu、W.-Y. Ma、Y.-Q. Zhang、L. Yan、M. Qiao、Y. Wu 和 M. Wang。《DAPO：可扩展的开源大语言模型强化学习系统》，2025年。网址：https://arxiv.org/abs/2503.14476。C. Zheng、S. Liu、M. Li、X.-H. Chen、B. Yu、C. Gao、K. Dang、Y. Liu、R. Men、A. Yang、J. Zhou 和 J. Lin。《组序列策略优化》，2025年。网址：https://www.arxiv.org/abs/2507.18071。

A 局限性与未来工作

尽管 DISPO 相较于现有方法取得了显著改进，但仍有若干局限性值得进一步研究。首先，我们的实验主要聚焦于数学推理任务，这类任务的二元奖励信号易于获取。将 DISPO 扩展至奖励结构更为细致的领域（如代码生成或开放式对话），仍有待探索。其次，尽管我们为设置四个裁剪参数（ϵ^+ 高、 ϵ^+ 低、 ϵ^- 高、 ϵ^- 低）提供了经验性指导，但确定最优取值仍需经过一定的试错。未来工作可以探索自适应或可学习的裁剪调度，根据训练动态自动调整这些参数。此外，我们的分析表明，在机制 1 和机制 2 中过度探索可能导致训练后期性能逐渐退化。虽然我们通过谨慎调参缓解了这些问题，但若能开发出有原则的方法来检测并防止此类退化（例如通过熵正则化或动态探索调度），将具有重要价值。最后，受计算约束所限，我们的实验仅涵盖参数量不超过 30B 的模型。随着该领域不断迈向更大型的推理系统，未来工作可以研究 DISPO 在更大模型上的可扩展性。

B 实验细节

B.1 RLVR 基线

我们将 DISPO 与 PPO 和 REINFORCE 两类算法的代表方法进行比较。我们选择 DAPO 作为 PPO 风格基线，并采用自行实现的 CISPO 作为 REINFORCE 风格基线。由于原始 CISPO 的裁剪参数尚未公开，我们将 CISPO 配置为$\epsilon_{\text{low}} = 1$ 和 $\epsilon_{\text{high}} = 100$，这些设置在所有测试模型上均能保持稳定。值得注意的是，降低这些数值（设置为 $\epsilon_{\text{low}} < 1$ 或 $\epsilon_{\text{high}} < 100$）会导致 CISPO 训练崩溃。对于 DAPO，我们保留标准裁剪参数：$\epsilon_{\text{low}} = 0.2$ 和 $\epsilon_{\text{high}} = 0.28$。鉴于先前工作表明 CISPO 优于多种 PPO 方法（包括 DAPO 和 GRPO [Chen 等， 2025]），我们仅关注一个 PPO 风格基线。

B.2 训练

训练时，我们使用图9所示的聊天模板。按照 DAPO 配方 [Yu 等， 2025]，我们在所有训练运行中采用动态采样。我们将小批次大小设为 512，微批次大小设为 32，即每个小批次进行 16 次梯度更新。我们将组大小设为 $G = 16$。损失计算中不包含 KL 散度正则化项。最大响应长度为 20,480 个词元。所有实验均采用与 DAPO 配方相同的超长惩罚。所有模型均使用 AdamW 优化器，学习率为 1×10^{-6}。按照 CISPO 配方 [Chen 等， 2025]，在 AdamW 中，我们设置 $(\beta_1, \beta_2) = (0.9, 0.95)$、$\epsilon = 1\text{e}{-15}$，并将权重衰减设为 0.1。我们采用 1.0 的梯度范数裁剪，并使用 CISPO 配方中提出的结合重复检测的提前截断。所有模型均在 H200 GPU 上使用 Verl 代码库和 vLLM 进行 rollout 训练，每个模型约需 20,000 GPU小时。训练期间，温度设为 1.0，top-p 设为 1.0；推理时，所有模型的温度设为 =1.0，top-p 设为 = 0.7。对于 AIME’24、AIME’25 和 AMC’23 的推理，我们使用图9所示的系统提示词；对于 MATH-500 和 Minerva，则使用图10所示的 Qwen3 风格系统提示词，以确保解析方式一致。

B.3 评估

我们使用五个广受认可的数学推理基准来评估模型，即 **AIME’24** [美国数学协会,2025], **AIME’25** [美国数学协会，2025], **AMC’23** [美国数学协会，2024], **MATH-**Hendrycks 等人, 2021a], 以及 **Min-erva** Lewkowycz 等人，2022]。这些基准用于评估模型在多样领域和不同难度级别下解决复杂问题的能力。每道题都要求生成最终答案，通常是一个数字、一个简化表达式（例如， $p - q$）或一个简洁文本回答（例如， $even$）：

请逐步解答以下数学问题。回答的最后一行应采用以下格式：Answer: $Answer（不带引号），其中 $Answer 是该问题的答案。

{question}

请将答案另起一行，写在“答案：”之后。

图9：DAPO 风格的聊天模板。

{question}

请逐步推理，并将最终答案放在 \boxed{} 中。

图 10：Qwen3 风格的聊天模板。

▷ **AIME’24**：来自 2024 [MAA（美国数学协会）的美国数学邀请赛中的 30 道题目 2025]。每道题通常都需要经过多个步骤的复杂推理，答案为 0 到 999 之间的整数。

▷ **AIME’25：** 一套即将推出的题目，共 30 道，选自 [MAA（美国数学协会）的 2025年 AIME 考试 2025], ，涵盖代数、几何、数论和组合学等类似范围的主题。

▷ **AMC’23：** 来自 2023 [MAA 的美国数学竞赛中的 40 道题目， 2024]。该考试是 AIME 的前置考试，包含旨在考查代数、几何、数论和概率方面创造性解题能力的题目。

▷ **MATH-500：** 从 MATH 数据集 [Hendrycks 等人，2021a],中精选出的 500 道题目子集，由 Lightman 等人选定 [2023]。这些题目涵盖七个主题：初等代数、代数、数论、计数与概率、几何、中级代数和微积分预备。

▷ **Minerva：** 由 Lewkowycz 等人引入的基准测试 [2022], ，包含 272 道高级定量推理问题，题目来源多样，包括研究级数学和科学考试。

为评估模型输出的正确性，我们遵循数学大语言模型评估中的标准做法。我们使用正则表达式解析每个生成的解答，以提取最终答案，并将其与真实答案进行比较。对于评估集中的每道题目，我们生成 16 个响应，并报告平均准确率（记为 Avg@16）。在训练过程中，我们跟踪模型在 AIME’24 上的 Avg@16 准确率，并选择准确率最高的检查点，用于在所有基准测试上进行最终评估。此外，我们还使用同一组 16 个生成的补全文本计算辅助指标，包括词元级平均熵和响应长度。

C 更多实验结果

图11和图12分别展示了 Qwen3-30B-A3B-Base 和 Qwen3-8B-Base 的所有学习曲线。我们可以看到，所有曲线都呈现出与图5相似的特征，表明 DISPO 在不同模型规模和架构下都具有稳健性。

图 13 展示了一个在推理时间重复出现的词元示例，正如第 5.3 节关于机制 3 的讨论中所提到的。

图 14 展示了 Qwen3-14B-Base 的响应长度曲线。机制 1 和机制 2 运行的响应长度曲线分别如图15和图16所示。

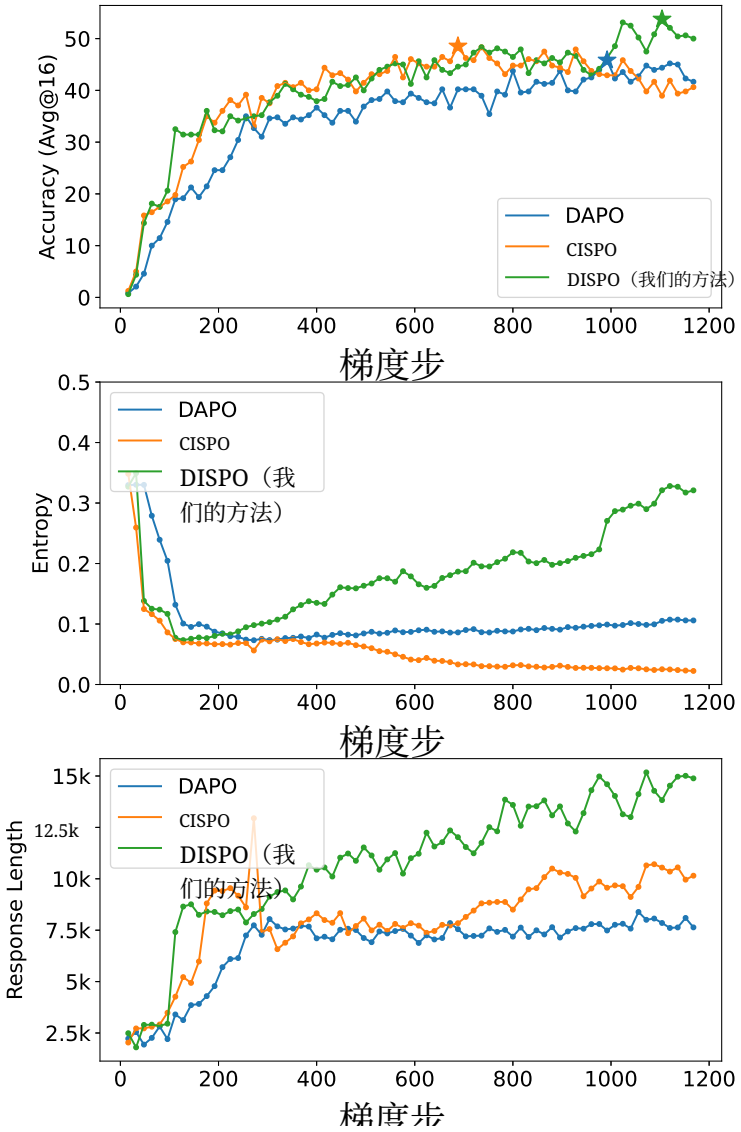


图 11: Qwen3-30B-A3B-Base 的训练运行学习曲线。 * 表示最高准确率 y.

D 关于机制 1 和 2 的补充讨论

图 17 展示了同时启用 $r_{i,t} > 1$ 和 $r_{i,t} < 1$ 的结果，方法是将 ϵ^+ low 和 ϵ^+ high 都设为非零。此配置使模型能够同时放大学习信号（机制 1）并减少恢复信号（机制 2），从而在探索与蒸馏机制之间形成相互作用，并带来更高的准确率。熵动态清晰地揭示了机制 1 和机制 2 之间的竞争。与橙色曲线（仅启用机制 1，且 ϵ^+ low $= 0.28$）相比，图 17 中的绿色曲线（ϵ^+ low $= 0.28, \epsilon^+$ high $= 1$）呈现出更慢的熵增长，说明机制 2 的同时启用抵消了机制 1 的熵增效应。

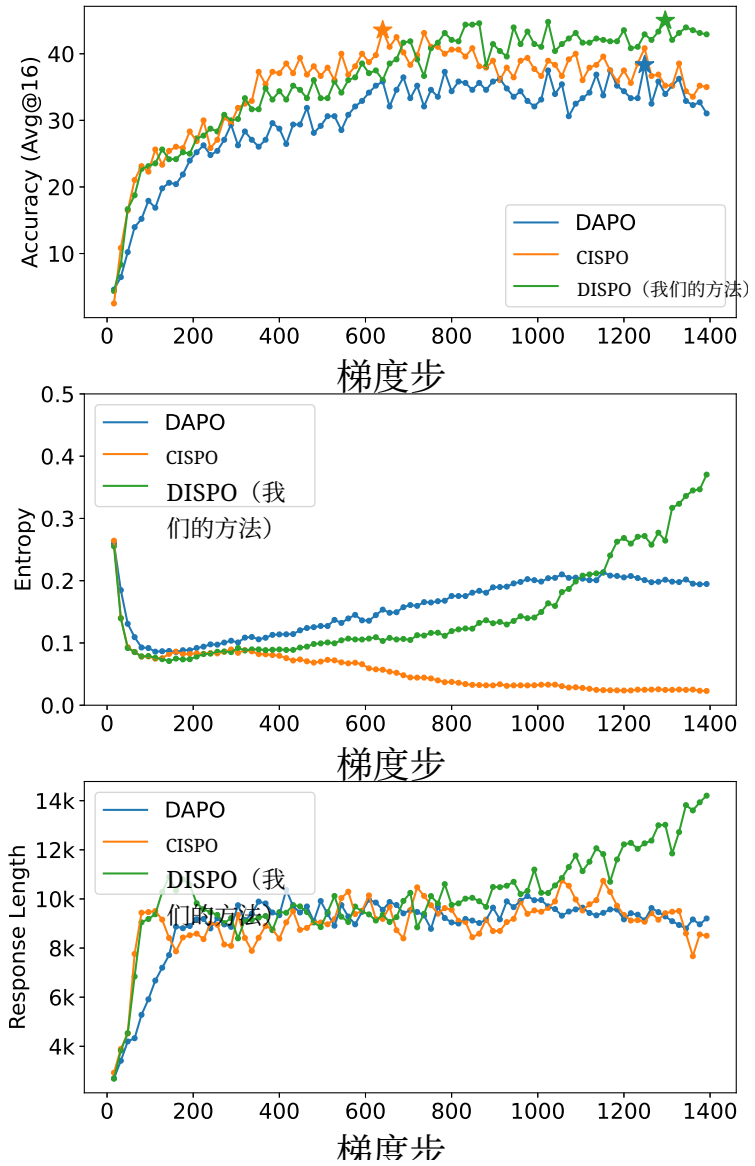


图 12：Qwen3-8B-Base 运行的学习曲线。 ⋆ 表示最高准确率。

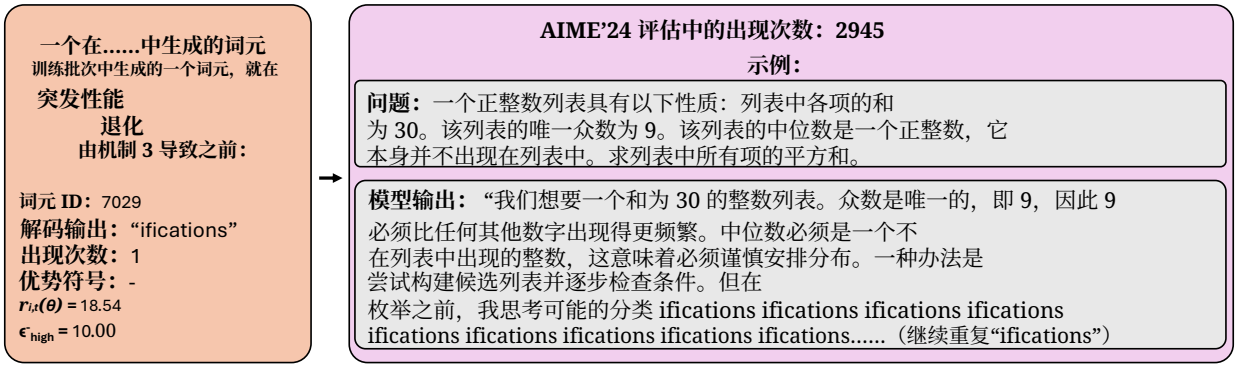


图 13：机制 3 退化示例：当 $r_{i,\ t}(\theta) > \epsilon^{-}$ 高时，机制 3 更新会受到抑制，导致词元被困在高概率区域，并在评估过程中反复生成。

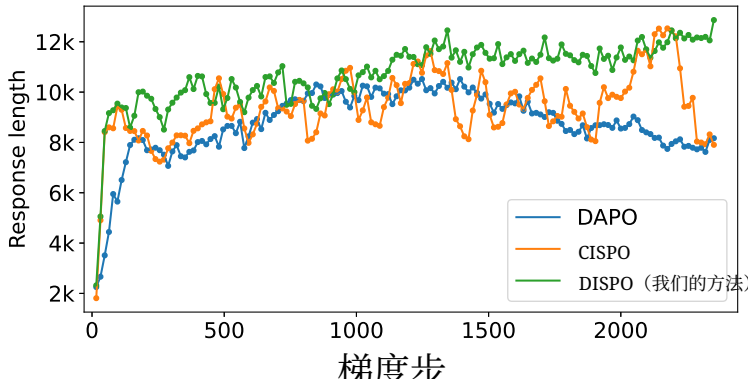


图 14：Qwen3-14B-Base 运行的响应长度曲线。

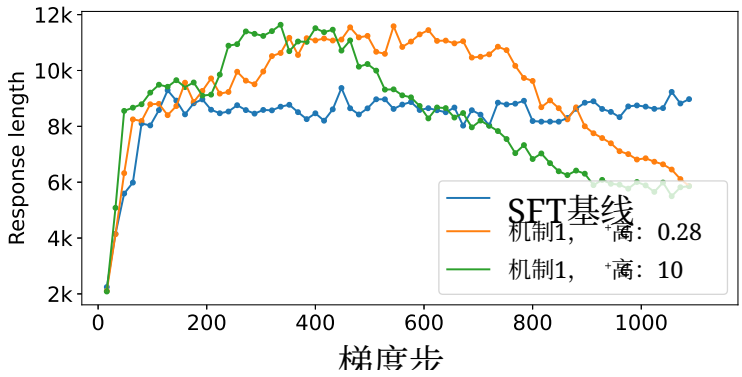


图 15：机制1各次运行的响应长度曲线。

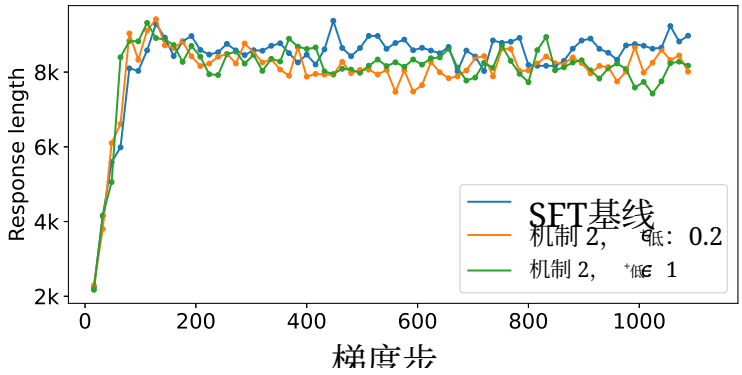


图 16：机制 2各次运行的响应长度曲线。

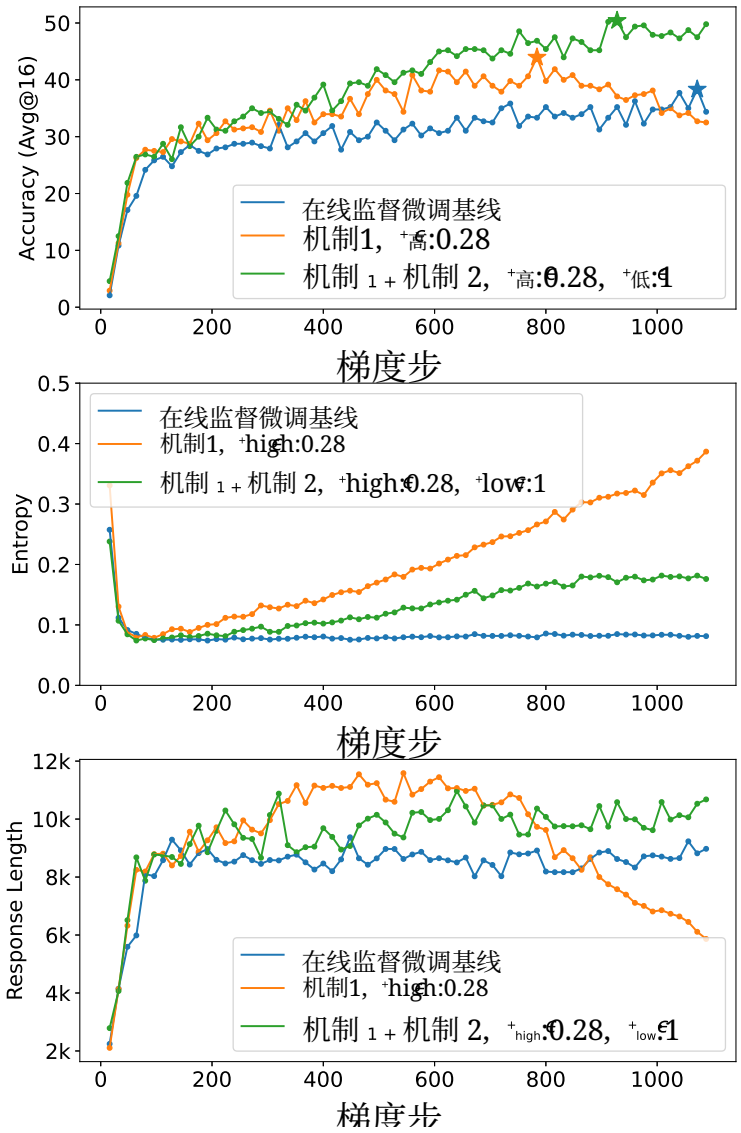


图 17：机制 1 + 机制 2 运行的学习曲线。准确率面板中的 ⋆ 表示最大值。