

# FIPO: 通过未来-KL 影响的策略最优化激发深层推理

 Qwen Pilot Team, Alibaba Group \*

 Project Page  GitHub  HuggingFace  ModelScope

## Abstract

我们提出 **未来-KL 影响策略最优化 (FIPO)**，这是一种强化学习算法，旨在克服大语言模型中的推理瓶颈。尽管类似 GRPO 的训练方式具有良好的可扩展性，但通常依赖于基于结果的奖励 (ORM)，该奖励将全局优势均匀分配给轨迹中每个 token。我们认为，这种 **粗粒度的信用分配** 由于无法区分关键逻辑转折点与无关紧要的 token，从而设定了性能上限。FIPO 通过在策略更新中引入 **折扣未来-KL 散度**，构建了 **稠密优势表述**，根据 token 对后续轨迹行为的影响程度重新加权。实证结果表明，FIPO 能使模型突破标准基线中存在的 **长度停滞现象**。在 Qwen2.5-32B 上评估，FIPO 将平均思维链长度从约 4,000 提升至超过 10,000 个 token，同时将 AIME 2024 Pass@1 准确率从 50.0% 提升至峰值 58.0% (收敛于约 56.0%)。该表现优于 DeepSeek-R1-Zero-Math-32B (~47.0%) 和 o1-mini (~56.0%)。我们的结果表明，建立稠密优势表述是推动基于 ORM 的算法演进、充分释放基础模型推理潜力的关键路径。我们开源了基于 verl 框架构建的训练系统。

## 1 引言

推理时的缩放策略，如 OpenAI 的 o 系列 (Jaech et al., 2024)、Gemini 系列 (Comanici et al., 2025) 和 DeepSeek 的 R 系列 (Guo et al., 2025) 所采用的策略，标志着大型语言模型进行推理方式的根本转变。通过在推理时分配更多的计算资源，这些方法支持更长的思维链和更审慎的推理，从而在竞争性数学和编程等高难度任务上取得显著提升。这一进展的大部分源于大规模强化学习与可验证奖励 (RLVR) (Guo et al., 2025; Team et al., 2025a; Yang et al., 2025; Team et al., 2025b; Zeng et al., 2025)，该方法利用任务特定验证器的反馈来微调模型的生成策略，从而激发并放大其推理能力。然而，由于具体的算法和训练配方仍 largely 未公开，目前尚不清楚强化学习如何作为主要催化剂来解锁潜在的推理深度，**有效地从初始不表现出此类倾向的基础模型中激发长思维链行为的出现**。

与此同时，开源社区已投入大量精力，在更加透明的情景下复现并缩放类似的算法 (Qin et al., 2024; Huang et al., 2024; Liu et al., 2025; Hu et al., 2025; Yu et al., 2025)。在这些努力中，DAPO (Yu et al., 2025) 提供了对干净基础模型应用 GRPO 风格训练的有前景的大规模复现。然而，我们认为 GRPO 框架内在对基于结果奖励的依赖引入了显著的结构约束。由于奖励仅在轨迹末端可进行二元验证，标准公式会向每个 token 分配相同的优劣值。这导致了 **完全粗粒度的信用分配**，即算法将关键推理步骤与无关紧要的 token **视为同等重要**。具体而言，我们观察到此类基线生成的推理轨迹往往在中间长度处趋于停滞。我们

\*Full author list available in the Contributions section.



图 1: FIPO 与基准方法在 AIME2024 上的性能对比。FIPO 表明，仅使用强化学习训练已经足以不仅超越其他纯强化学习基准方法（复现的 DAPO 和 Deepseek-R1-Zero-32B），还超过了 o1-mini。这种性能提升伴随着生成响应长度的显著增加。

认为这一局限性为标准 GRPO 设定了性能下限：由于均匀奖励无法突出驱动正确逻辑的具体 token，模型无法收敛至完成复杂任务所需的长周期、高复杂度推理路径。尽管这一局限促使近期研究 (Hu et al., 2025; Yue et al., 2025; Fan et al., 2025) 重新采用 PPO 框架以实现更精细的优劣估计，但我们认为这种细粒度的信用分配无需依赖评论员模型的复杂性即可实现。

我们引入了 Future-KL Influenced Policy Optimization (FIPO)。FIPO 通过引入 Future-KL 散度来修改策略更新，该方法基于后续轨迹的累积行为对当前 token 的优势进行重新加权。为保持训练稳定性，该目标与 **影响权重裁剪和过滤机制**相结合。我们在 Qwen2.5-32B-Base 上评估该方法，该模型未接触过长 CoT 合成数据，使用 DAPO 公开发表的训练数据集 (Yu et al., 2025) 以确保严格可控的对比。如 Figure 1 所示，FIPO 打破了标准基线的性能天花板；当 DAPO 在 AIME 2024 上达到 50.0% (Pass@1) 时，FIPO 实现了推理链的持续延展，模型从基准的 4,000 个 token 稳步扩展至超过 10,000 个 token 的深度推理模式。这一持续扩展使准确率提升至峰值 58.0%，与近期基于 PPO 的方法表现相当。**这些发现表明，建立稠密优势形式有效弥合了 GRPO 效率与 PPO 性能之间的差距，解锁了在均匀奖励方案下原本无法触及的深度推理能力。**

我们的实现基于 verl 框架 (Sheng et al., 2025) 和 DAPO 代码库。通过完全公开完整的训练代码和配置方案，我们旨在揭示大规模强化学习在大模型中的宝贵见解，以惠及更广泛的科研社区。

## 2 初步的

在本节中，我们回顾了本文工作核心的策略最优化框架：PPO 及其无价值网络变体 GRPO 与 DAPO。在本文中，令  $T$  表示轨迹的总长度， $t$  表示该轨迹中当前步骤的索引。在 GRPO 情景下，对于每个问题提

示  $q$ ，我们采样  $G$  个轨迹，得到的输出记为  $o$ 。

## 2.1 近端策略最优化

**近端策略优化 (PPO)** (Schulman et al., 2017) 为策略最优化引入了裁剪的替代目标。通过使用裁剪机制将策略更新限制在旧策略的邻近范围内，PPO 稳定了训练过程并提升了样本效率。具体而言，PPO 最大化：

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}_{(q,o) \sim \mathcal{D}, o \sim \pi_{\theta_{\text{old}}}(\cdot|q)} [\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)].$$

此处， $r_t(\theta) = \frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{\text{old}}}(o_t|q, o_{<t})}$  表示在步骤  $t$  时的 token 级别概率比， $\hat{A}_t$  是通过学成的价值函数估计的优势， $\epsilon$  是裁剪系数。关键的是，标准的 PPO 实现使用广义优势估计 (GAE) (Schulman et al., 2015) 来计算优势  $\hat{A}_t$ 。这产生了不同的、与 token 相关的优势信号，使模型能够执行时间上的信用分配。这与仅从最终结果推导优势的简化公式形成对比，后者实际上将统一的信号广播到轨迹中的所有 token。通过使用 GAE，PPO 在每一步都提供了稠密的监督，使其能够在生成过程中区分关键动作和影响较小的动作。

## 2.2 群体相对策略最优化

**分组相对策略最优化 (GRPO)** (Shao et al., 2024) 通过基于分组的采样来估计优势，避免了价值网络带来的计算负担。对于给定的查询  $q$  (以及真实值  $a$ )，从旧策略  $\pi_{\theta_{\text{old}}}$  中采样一组输出  $\{o_i\}_{i=1}^G$ 。第  $i$  个样本的序列级优势被标准化为：

$$\hat{A}_i = \frac{R_i - \mu}{\sigma}, \quad \text{with } R_i = \mathbb{I}(\text{Verify}(o_i, a)), \quad (1)$$

其中  $\mu$  和  $\sigma$  分别表示样本组内奖励的样本均值和标准差。与 PPO 类似，GRPO 采用截断目标函数，但直接在损失中添加了每 token 的 KL 罚项：

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left( \min(\rho_{i,t}(\theta)\hat{A}_i, \text{clip}(\rho_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_i) - \beta D_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}}) \right) \right]. \quad (2)$$

此处， $\rho_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$  表示概率比。按照设计，计算得到的标量  $\hat{A}_i$  在整个序列中进行广播；具体而言，对于每个 token  $t$ ，优势值均被设置为相同的值  $\hat{A}_{i,t} = \hat{A}_i$ 。与 PPO 不同，PPO 中广义优势估计 (GAE) 为每个 token 提供了不同的信号，而 GRPO 则对轨迹中的每一步分配相同的信用，无论其对最终结果的个体贡献如何。

## 2.3 解耦剪枝与动态采样策略最优化

**解耦剪裁与动态采样策略最优化 (DAPO)** (Yu et al., 2025) 通过消除显式的 KL 惩罚扩展了 GRPO 框架。相反，它在区间  $(1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}})$  内采用非对称剪裁，以增强有利动作的更新，有效缓解了 GRPO 中常见的熵崩溃问题。此外，DAPO 引入了基于 token 级别的策略梯度损失，以在长思维链强化学习训练背景下维持健康的最优化动态。同时，DAPO 实施了一种动态采样机制，确保每个组内都包含正例和负例  $\{o_i\}_{i=1}^G$ 。该机制保证了优化过程中具有非平凡梯度的有效更新。本文采用 DAPO 作为主要基准。

### 3 FIPO

在本节中，我们介绍未来 KL 诱导策略最优化（FIPO）的核心框架。我们首先讨论概率转移，这是我们的目标函数的基本构建单元。接着，我们详细阐述未来 KL 的公式化表达。最后，我们说明我们的方法如何通过聚焦局部“未来上下文”，实现一种“软衰减窗口”策略。该机制自然地优先考虑邻近信号而非遥远信号，从而将有效视野限制在最相关的后续 token 范围内。

#### 3.1 概率转移： $\Delta \log p$

我们的方法建立在我们近期对大模型（LLMs）在强化学习过程中动态行为的研究基础之上。具体而言，我们之前关于 RLVR 更新 (Huang et al., 2025) 的工作表明，概率变化的大小和方向  $\Delta \log p$  可作为推理能力提升的稳健指标。在此基础上，我们对分布偏移 (Meng et al., 2025) 的细粒度分析进一步揭示，这一生成过程通常由少数“稀疏但关键”的 token 驱动，这些 token 对后续思维链产生不成比例的影响。受此启发，我们将 token 级别的概率变化识别为信用分配机制的基本单元。形式上，我们将时间步  $t$  时的概率变化定义为当前策略与旧策略在对数空间中的差值：

$$\Delta \log p_t = \log \pi_\theta(o_t | q, o_{<t}) - \log \pi_{\theta_{\text{old}}}(o_t | q, o_{<t}). \quad (3)$$

该术语用作捕捉即时策略漂移的差分信号：

- **正向偏移** ( $\Delta \log p_t > 0$ )：表示当前策略相对于旧策略提高了 token  $o_t$  的似然。这通常表明训练目标正在强化这一特定推理步骤。
- **负向偏移** ( $\Delta \log p_t < 0$ )：表示策略正在抑制  $o_t$  的生成，表明更新后的模型相对于参考策略主动降低了该特定 token 的权重。

与传统的 KL 惩罚项将这种漂移主要视为需最小化的正则化成本不同，我们将  $\Delta \log p_t$  解释为行为调整的方向性信号，从而显式地将最优化目标与生成动态耦合起来。然而，仅依赖这一瞬时变化是不够的，因为它无法捕捉决策的长期后果。这一局限性促使我们提出了 Future-KL 机制，该机制通过聚合当前 token 的未来轨迹中的分布漂移来重新加权当前 token。

#### 3.2 未来-KL 估计

虽然  $\Delta \log p_t$  捕捉了局部分布变化，但推理本质上是一个顺序过程，当前 token 的真正意义取决于其引发的轨迹。为了捕捉这种因果影响，我们定义 Future-KL 为从当前步骤  $t$  到序列末尾  $T$  的累积有符号概率变化：

$$\text{FutureKL}_t = \sum_{k=t}^T \Delta \log p_k. \quad (4)$$

该求和在数学上等价于后续序列  $o_{t:T}$  的联合概率分布的对数似然比。因此，它可以被解释为限制在未来时间范围内的 KL 散度的基于样本的估计，衡量当前策略相对于参考策略在轨迹剩余部分的累积偏差。我们因此将此指标称为 Future-KL。从功能上看， $\text{FutureKL}_t$  作为一个前瞻性指标，量化了策略分布在未来轨迹中的累积变化。正值 ( $\text{FutureKL}_t > 0$ ) 表示更新后的策略整体上 **强化了** 由 token  $o_t$  引发的后续轨迹，表明  $o_t$  在后续推理链中起到了稳定锚点的作用。相反，负值 ( $\text{FutureKL}_t < 0$ ) 意味着策略整体上抑制了  $o_t$  之后的未来 token，表明从这一点出发的轨迹在优化过程中变得不那么受青睐。

然而，在实际应用中，这种公式往往加剧了由分布偏移引起的方差。由于  $\text{FutureKL}_t$  作为优势函数的权重系数（详见后续章节），未来 logits 的过度偏差（例如由训练-推理不一致导致）会不成比例地放大其规模。这使得最优化过程对噪声 token 过于敏感，而非关注推理链的内在质量。经验上，我们观察到在缺乏安全机制的情况下，训练过程容易出现严重不稳定。如 Figure 2 所示，这种崩溃现象伴随着“低裁剪比例”度量的急剧上升，该度量跟踪触发双裁剪阈值（对负例施加硬裁剪比率）的样本频率 (Ye et al., 2020)。负例上如此高的重要性比例表明存在严重错位：模型为一个实际上有害的动作分配了高概率。在我们的实验中，这一峰值（约在第 70 步）与梯度范数和策略 KL 的激增相吻合。<sup>1</sup>，表明策略分布发生了显著变化，同时响应长度立即下降。这种同步性表明，在没有调控的情况下， $\text{FutureKL}_t$  产生的累积负面信号可能达到某些极端值，从而导致训练过程不稳定。

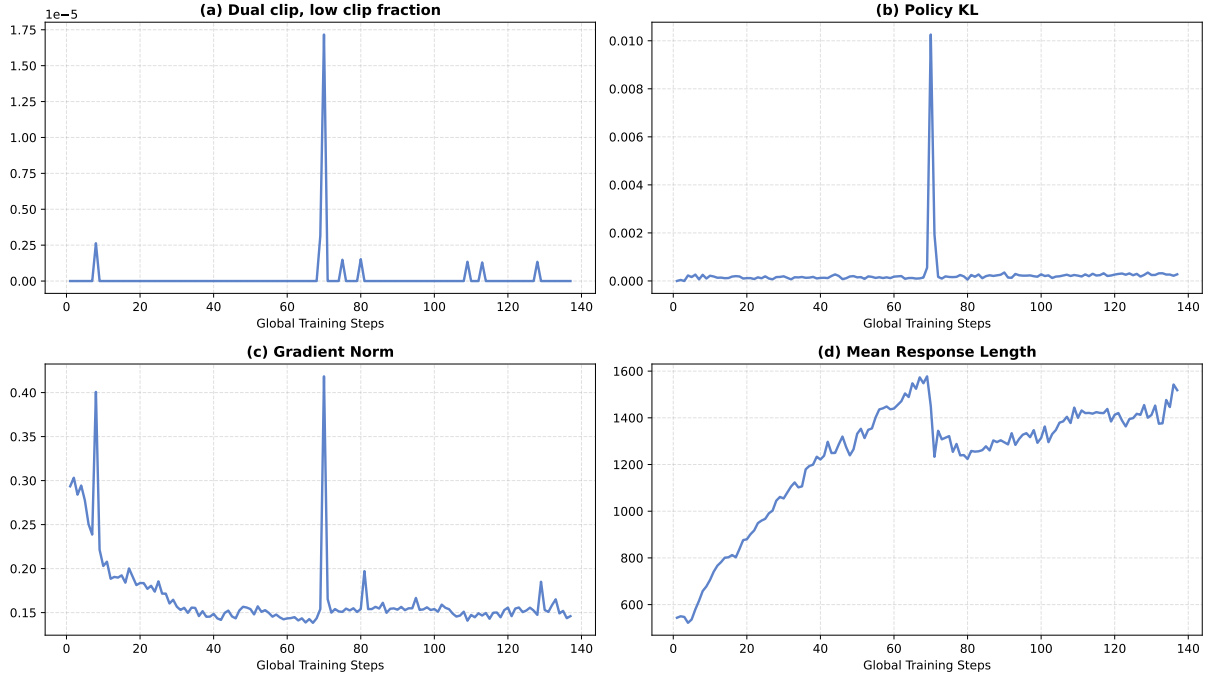


图 2: 原始 FutureKL 的训练不稳定性。对第 70 步附近观察到的不稳定运行进行分析。(a) 低裁剪比例的急剧上升（表明负例驱动策略分布发生剧烈变化）触发了 (b) 策略 KL 的突然发散。(c) 梯度范数的立即爆炸，这些内部不稳定性共同导致 (d) 响应长度的灾难性坍塌，证实未经调控的负信号会使最优化过程变得不稳定。

受这些观察的启发，我们通过显式地对超出 Dual-Clip 阈值的 token 进行掩码处理，来优化 FutureKL 计算。由于这些 token 代表了“有害”动作，其梯度已通过裁剪策略目标被裁剪，若允许它们过高的重要性比率传播到递归求和中，将引入严重的方差。通过将这些特定异常值的未来累积置零，我们消除了不稳定的主要来源。改进后的目标定义如下：

$$\text{FutureKL}_t = \sum_{k=t}^T M_k \cdot \Delta \log p_k, \quad M_k = \mathbb{I} \left( \frac{\pi_\theta(o_k|o_{<t})}{\pi_{\text{old}}(o_k|o_{<t})} \leq c \right). \quad (5)$$

此处， $M_k$  作为二值滤波器，仅当重要性比率保持在 Dual-Clip 阈值  $c$  范围内（通常为  $c \geq 10$ ）时取值为 1，否则为 0。这确保了触发硬约束的 token 从 FutureKL 计算中被有效排除，从而防止梯度爆炸，同时不改变轨迹的有效信号。

<sup>1</sup>我们计算策略的 KL 散度为负对数比的批量均值： $\text{Policy KL} = \frac{1}{B \cdot L} \sum_{i=1}^B \sum_{t=1}^L (\log \pi_{\text{old}}(o_{i,t}|o_{<t}) - \log \pi(o_{i,t}|o_{<t}))$ 。它衡量了在当前策略与梯度更新前的策略（即回放策略）之间生成序列的 KL 散度。



### 3.2.1 软衰减窗口

除了稳定性约束之外，我们还处理了长时域生成中的固有不确定性。当前动作  $o_t$  与未来 token  $o_k$  之间的因果依赖关系会随着时域  $k - t$  的增加而自然减弱。近期的后续 token 直接依赖于当前的选择，而远期 token 则受到累积随机性的影响，变得越来越不可预测。为了建模这种影响的衰减，我们引入了一个折扣因子  $\gamma \in (0, 1]$ 。将这一衰减机制融入掩码目标函数，得到了我们在实验中使用的最终公式：

$$\text{FutureKL}_t = \sum_{k=t}^T M_k \cdot \gamma^{k-t} \cdot \Delta \log p_k. \quad (6)$$

我们将衰减率参数化为  $\gamma = 2^{-\frac{1}{\tau}}$ ，其中  $\tau$  是一个超参数，用于控制未来监督的有效范围（或“半衰期”）。该公式确保信用分配集中在即时推理链上，对遥远且高度不确定的 token 分配较低权重。从功能上看， $\tau$  定义了这一软衰减窗口的孔径。与在固定步数后突然丢弃信息的硬截断不同，这种指数形式构建了一个连续滑动窗口，其中  $\tau$  表示未来信号影响减弱至一半时的距离。该机制使模型能够优先关注窗口  $\tau$  内的局部一致性，同时平滑地过滤来自遥远未来的噪声，而不会引入边界伪影。

### 3.2.2 FutureKL 重加权优势剪裁

最后，我们将软衰减窗口和掩码机制整合到策略最优化目标中。我们提出使用未来影响权重  $f_t$  来调节标准优势估计  $\hat{A}_t$ 。修改后的优势  $\tilde{A}_t$  定义为：

$$f_t = \text{clip} \left( \exp(\text{FutureKL}_t), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}} \right), \quad \tilde{A}_t = \hat{A}_t \cdot f_t. \quad (7)$$

该公式引入了两个关键操作：

1. **指数映射**：我们将对数空间中的累积标量信号变换到乘法领域。从数学上讲，未截断的项表示自然比的衰减加权乘积，作为反映策略对生成未来有效偏好的重要性权重。
2. **影响权重裁剪**：我们将乘法系数  $f_t$  限制在区间  $[1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}]$  内。此操作严格用于限制优势调制的幅度，防止指数项向梯度估计引入过大的方差。通过裁剪权重，我们确保未来轨迹对更新信号的调制处于可控范围内，避免因极端累积对数概率偏移导致的数值不稳定性。

从功能上看，这种调制根据生成的未来轨迹的强化或抑制来调整策略更新的幅度。当更新后的策略**强化**后续轨迹（即  $\text{FutureKL}_t > 0$ ）时，权重项  $f_t > 1$  会放大梯度信号。因此，正向优势被增强以鼓励当前 token 作为稳定的锚点，而负向优势则受到更严厉的惩罚，以严格纠正由此路径引发的错误。相反，当策略**抑制**未来轨迹（即  $\text{FutureKL}_t < 0$ ）时，项  $f_t < 1$  会削弱更新。这种削弱有效降低了那些偶然出现在成功序列中的局部有害 token 的奖励信号，并减轻了那些困于失败序列中的优质 token 的惩罚。在实际应用中，为确保训练稳定性并防止过度惩罚，我们对具有过大的重要性比率且关联负向优势（ $\hat{A}_t < 0$ ）的 token 重置  $f_t = 1$ 。

### 3.3 目标损失

采用 DAPO (Yu et al., 2025) 的 token 级公式, 我们最大化以下 FIPO 目标:

$$J_{\text{FIPO}}(\theta) = \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\} \sim \pi_{\theta_{\text{old}}}} \left[ \frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min(r_{i,t} f_{i,t} \hat{A}_{i,t}, \text{clip}(r_{i,t}, 1 - \epsilon, 1 + \epsilon) f_{i,t} \hat{A}_{i,t}) \right]. \quad (8)$$

此处,  $G$  表示每个查询的样本输出数量,  $r_{i,t} = \frac{\pi_{\theta}(a_{i,t}|s_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t}|s_{i,t})}$  表示当前策略与旧策略之间的重要性比率。项  $\hat{A}_{i,t}$  指的是组相对优势, 而  $f_{i,t}$  为之前引入的未来 KL 重要性权重。

## 4 实验

### 4.1 实验情景

在本工作中, 我们采用 DAPO (Yu et al., 2025) 的训练情景, 特别聚焦于数学推理任务, 以确保严格的对比。我们使用 VeRL 框架 (Sheng et al., 2025) 进行训练和基准复现。我们保持与 DAPO 一致的最优化设置, 并在公开发布的 DAPO-17K 数据集上进行训练。每个训练批量包含 512 个提示, 每个提示采样 16 个响应, 总计产生 8,192 个训练样本。在标准的 DAPO 配置中, 以小批量大小为 512 个样本 (32 个提示) 进行更新, 导致每轮训练迭代执行 16 次梯度更新。然而, 我们的实验发现表明, 更大的小批量大小可提升训练稳定性。因此, 我们采用了 1,024 个样本 (64 个提示) 的小批量大小, 使每轮迭代仅执行 8 次梯度更新。关于该增大小批量大小影响的更详细讨论见附录第 F 节。对于 Future-KL 计算, 我们将衰减率  $\tau$  的有效时域设为 32。针对 32B 模型的训练, Future-KL 权重在  $[1, 1.2]$  内被截断; 这有效地放大了与成功推理轨迹相关的 token 的奖励, 同时对导致错误结果的轨迹施加更严格的惩罚。FIPO 和 DAPO 均具有最大响应长度 20,480 token, 对超过 16,384 token 的轨迹施加过长惩罚。基准和 FIPO 的详细超参数配置见附录第 B 节。

在评估过程中, 我们采用 AIME 2024 作为主要的验证基准, 并辅以 AIME 2025, 以确保与 DAPO 基准进行严格且全面的对比。为保持结果的稳定性并考虑思维链生成中的方差, 我们遵循 DAPO 协议, 重复评估 32 次, 并报告 Pass@1 (对 32 个样本取平均)。推理超参数始终设置为温度 1.0 和 top- $p$  为 0.7。

### 4.2 主要结果

Table 1 展示了在 AIME 2024 和 AIME 2025 基准上的定量评估结果。相较于 DAPO 基准, FIPO 在两个数据集上均实现了约 6.0% 的系统性提升 (Pass@1, Avg@32)。我们优先考虑该指标, 因其是推理可靠性最稳健的指示器。尽管我们也观察到一致性方面的提升, 但在覆盖范围 (Pass@32) 上的改进相对有限, 尤其在 AIME 2025 上更为明显。我们认为这归因于仅通过强化学习扩展大模型绝对问题求解范围的内在挑战。在缺乏外部知识增强或工具集成的情况下, 强化学习主要受限于优化模型对其现有内部知识的导航方式。因此, 尽管 FIPO 显著提升了模型在其潜在能力范围内可靠求解问题的能力 (推动 Avg@32 提升), 但改变可求解问题的边界 (Pass@32) 仍具有显著难度。

## 5 分析

除了总体指标之外, 我们观察到几个我们认为支撑这些性能提升的显著现象。通过剖析训练动态和推理行为, 我们识别出 FIPO 有效性的三个关键驱动因素: 推理链中出现的基于长度的缩放, 由响应长度加

表 1: AIME 基准推理性能对比。所有结果均以百分比 (%) 报告。我们报告 32 个样本的平均 Pass@1 (Avg@32)、多数投票结果 (Cons@32) 以及至少一个正确答案的概率 (Pass@32)。为与先前基准报告保持一致并降低对数字级别生成方差的敏感性，最终数值四舍五入到最接近的整数。

Method	AIME 2024			AIME 2025		
	Avg@32	Cons@32	Pass@32	Avg@32	Cons@32	Pass@32
DAPO (Baseline)	50.0%	60.0%	80.0%	38.0%	47.0%	63.0%
FIPO (Ours)	56.0%	73.0%	83.0%	43.0%	50.0%	67.0%

权平均优势公式捕捉到的**独特的正向学习信号**，以及与标准基准相比显著提高的**最优化稳定性**。

## 5.1 长度和性能的缩放

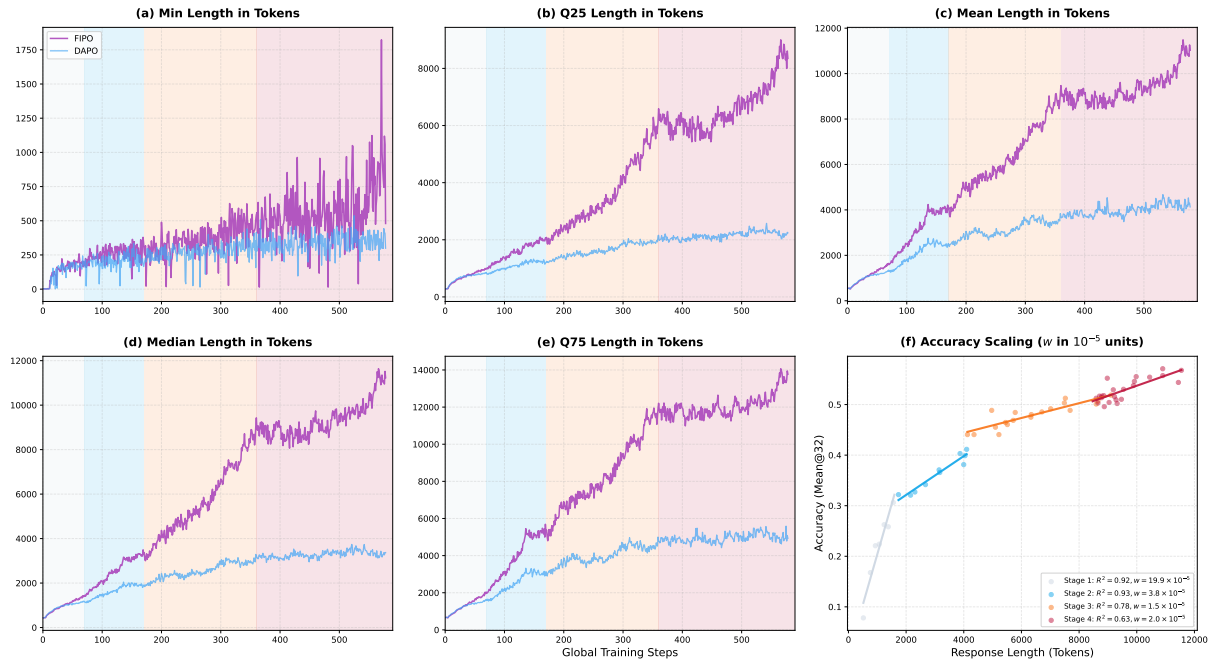


图 3: 响应长度与性能缩放在训练过程中的动态变化。子图 (a)-(e) 展示了响应长度度量指标（最小值、四分位数 25%、均值、中位数、四分位数 75%）随全局训练步数的演化过程。相较于 DAPO 基准，FIPO 显著增加了响应长度，有效激发了更广泛的思维链推理。子图 (f) 表明，这种增加的长度与准确率的提升呈强相关性，表明更长的思维链是突破性能瓶颈的关键。

FIPO 训练中的一个核心观察是，性能提升与响应长度的持续扩展紧密耦合。随着训练的推进，我们观察到 token 数量显著增加，且其增长与模型准确率同步。如Figure 3所示，DAPO 的响应长度在初始增长后逐渐进入停滞阶段，平均稳定在约 4,000 个 token。相比之下，FIPO 展现出卓越的缩放韧性。这一缩放过程通过不同的演化阶段展开（如Figure 3中彩色区域所示），标志着从初期快速探索向长期深度推理的转移。值得注意的是，尽管设置了过长惩罚以抑制冗余，FIPO 仍成功引导模型生成广泛的思维链 (CoT) 推理。附录E提供的定性分析表明，这种长度扩展源于自我反思行为的逐步涌现；模型逐渐利用扩展的序列长度重新评估其中间步骤，并探索多种方法以验证其结论。有趣的是，这种自发产生的系统性自我验证机制与先进推理模型在推理时的缩放行为相吻合（例如 OpenAI o 系列和 DeepSeek-R1）。这表明 FIPO 有效触发了**推理时推理**，优先关注**分析深度**，从而解锁更高性能。

对训练动态的进一步分析表明，这种长度的激增并非由孤立的异常值驱动，而是代表了整体分布的迁移。



如图Figure 3(a)–(e) 所示，在 FIPO 训练下，所有与长度相关的分位数（从极小点和 Q25 到中位数和 Q75）均表现出同步且稳定的上升趋势。具体而言，在这些训练阶段中，中位数 token 数量从初始的 200 稳步攀升至超过 10,000。整个分布上的这种迁移表明，FIPO 促成了模型潜在问题求解策略的根本性转变：模型从直接响应模式转向系统化、自我验证的推理过程。至关重要的是，我们发现这种向更长推理链的集体迁移正是实现实验中观察到性能突破的关键所在。如图Figure 3(f) 所示，模型准确率与响应长度在所有已识别阶段均呈现出强烈的正相关关系。尽管不同阶段的相关系数斜率（记为  $w$ ）略有差异，但其轨迹始终保持正向。当 DAPO 基准的表现因长度趋于平稳而陷入瓶颈时，FIPO 能够持续释放更多的“思考空间”，使模型得以应对日益复杂的逻辑依赖关系。这证实了 FIPO 成功将增加的序列长度转化为真正的推理深度，使模型能够在高难度推理任务上超越标准基线的性能极限。

## 5.2 优势与持续推理增长的动态

我们进一步通过比较奖励和优势的演化来研究训练动态。如 Figure 4(a) 所示，基准方法（DAPO）在训练过程中始终保持着比 FIPO 更高的平均奖励。然而，我们认为这种差异是奖励函数设计带来的数值伪影，而非性能优越的体现。由于奖励函数包含了过长的惩罚项，FIPO 构建复杂推理链的特性不可避免地导致更高的惩罚，从而抑制了其平均原始奖励。相反，基准方法较高的奖励源于其倾向于生成较短的响应。尽管该策略通过最小化惩罚最大化了即时奖励，但表明其收敛到了受限搜索空间内的局部最优解。

这一假设进一步得到了图Figure 4(b) 中 DAPO 采样批量数迅速增加的佐证。这一趋势表明，模型正在对训练集产生过拟合，不断生成缺乏区分性的样本（即全部正确或全部错误的批量），这些样本提供的梯度信息微乎其微。因此，算法被迫更加激进地采样，以获取足够的有效数据用于最优化。相比之下，FIPO 主动探索更广阔的搜索空间，优先考虑挑战性推理任务所需的结构深度，而非仅仅规避惩罚。

当从原始奖励转向优势所提供的动态激励时，这种差异变得更加显著。如 Figure 4(c) 所示，DAPO 在响应长度加权平均相对优势方面表现出下降趋势。<sup>2</sup> 在整个训练过程中，这表明正例的长度逐渐被负例所主导，导致扩展推导过程的动力不断减弱；由于更长的长度不再带来更多的正向优势，模型最终在推理能力上达到平台期。与此形成鲜明对比的是，FIPO 展现出持续上升的趋势。这表明正例正在变得比其负例显著更加充实。这种动态促进了持续增长的轨迹：随着生成更长、有效的推理链所带来的正向优势不断增大，在奖励稳步提升的推动下，模型得以保持动力，继续追求更长且更严谨的推理路径。

## 5.3 平滑策略漂移、探索与梯度更新

为了进一步刻画训练过程，我们考察了策略行为的演化和最优化稳定性。如 Figure 5(a) 所示，FIPO 展现出策略 KL 散度的稳定且有结构的上升趋势。这代表了一种渐进的策略迁移，模型持续地远离其 **先前的策略状态**，以朝着更专门的推理模式演进。这一趋势与我们的滚动观察结果定性一致：自我反思段落的长度是逐步增加而非突然跃升，反映了推理过程的渐进扩展（详见附录 E 中的示例）。

最优化特性在梯度尺度方面也存在显著差异。如图Figure 5(b) 所示，FIPO 的梯度范数在整个训练过程中保持较低且稳定，体现出基于细粒度更新的演化过程。相比之下，基准方法（DAPO）表现出高度波动的起伏，其梯度范数频繁出现剧烈峰值。这些波动的更新表明，DAPO 的搜索过程存在突发性转变和潜在的不稳定性风险。

<sup>2</sup>我们将响应长度加权的平均优势定义为： $\bar{A} = \frac{\sum_{i=1}^B \sum_{t=1}^{L_i} A_{i,t}}{\sum_{i=1}^B L_i}$ ，其中  $B$  为批量大小， $L_i$  为第  $i$  个样本的响应长度， $A_{i,t}$  表示 token 级别的组相对优势。

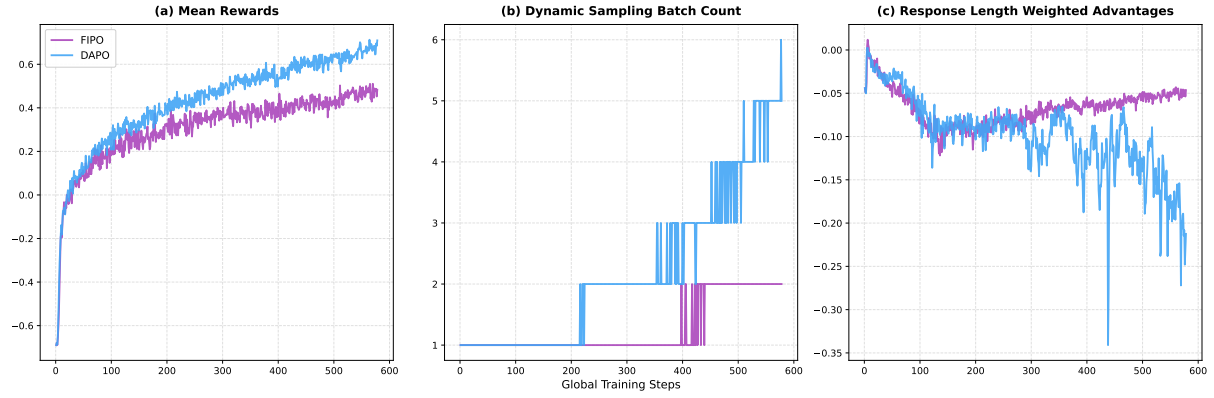


图 4: 训练奖励与长度加权优势的分析。(a) 平均训练奖励。DAPO 获得了更高的原始得分，这是预期的结果，因为两种方法都引入了过长惩罚，抑制了较长响应的奖励。(b) 采样批量数量。该指标反映了维持足够数量的有效批量所需的采样冗余度。较高的采样需求表明模型在训练集上频繁生成无信息量的轨迹，这可能是过拟合的一个潜在指示。(c) 响应长度加权的平均优势。FIPO 表现出持续上升的趋势，建立了正向强化循环，即更长的响应逐渐产生更大的正向优势。相比之下，DAPO 呈现下降趋势，表明其未能将长度有效转化为推理收益，最终限制了其性能。

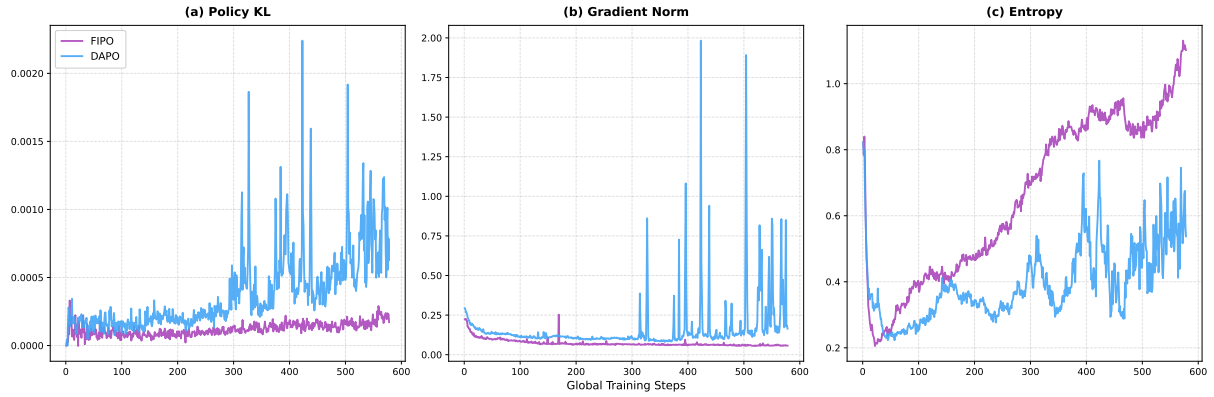


图 5: 策略演化与最优化动态。(a) 策略 KL 散度。(b) 策略熵。(c) 梯度范数。相较于 DAPO，FIPO 展现出更受控的策略漂移和更平滑的更新梯度。值得注意的是，FIPO 熵值的上升，结合图 4(b) 中的加权优势趋势，表明模型正在积极探索更广泛的推理空间，其中更长的思维链路径越来越多地对应正确解。

这种稳定性差异在策略熵 (Figure 5c) 中进一步体现。尽管 FIPO 保持了平稳且持续上升的熵值，表明对推理空间的持续而稳定的探索，但 DAPO 在整个训练过程中表现出噪声强烈的振荡。相比之下，随着训练的进行，DAPO 的熵值也呈现出明显的噪声振荡。综上所述，这些指标共同描绘了 FIPO 模型在实现复杂推理方向上的显著且有目的性的策略演化的同时，确保最优化过程在数值上保持良好行为。

## 6 结论

在本文中，我们提出了 Future-KL Influenced Policy Optimization (FIPO)，一种旨在解决标准 GRPO 中粗粒度贡献度分配问题的强化学习方法。通过将折扣化的 Future-KL 散度引入策略更新，FIPO 将基于稀疏结果的奖励转化为稠密的、token 级别的监督信号。我们的实证分析揭示并解决了现有基线中存在的关键“长度-性能平台”现象，证明了标准的均匀奖励无法维持长链推理。在 Qwen2.5-32B-Base 上进行验证后，FIPO 成功打破了这一瓶颈：其在 AIME 2024 上的表现从基线的 50.0% 提升至峰值 58.0%（收敛于 56.0%），并将平均思维链长度从 4,000 扩展至超过 10,000 个 token。至关重要的是，这些发现挑战了当前认为复杂评论员模型是实现细粒度贡献度分配必要条件的普遍假设，证明了稠密监督可以在更高

---

效的 GRPO 框架内有效实现。为促进未来研究，我们开源了完整的训练代码与训练方案，为社区提供了一条可扩展且易于访问的路径，以推动大规模推理模型的发展。

## 7 局限性和未来工作

尽管有效，FIPO 仍存在一些局限性：

**成本与效率。**一个主要的约束是扩展推理序列所带来的计算成本增加。由于 FIPO 成功地实现了超过 10,000 token 的思维链长度，训练和推理的开销显著增长，这对资源受限的部署带来了挑战。我们认为，高级推理的发展应是一个逐步进行的过程：首先激发长且高质量的推理能力，随后对这些能力进行效率优化。尽管本文专注于第一阶段，即突破长度停滞以实现更优性能，但将这些长推理路径转化为更简洁高效形式的任务是接下来的关键步骤。我们将留待未来探索。

**任务泛化。**另一个局限性在于，我们的评估主要在数学推理基准上进行。然而，我们认为数学可以作为深度推理的严格且具有代表性的代理；其对客观、可验证的真实值以及高密度逻辑一致性的要求，使其成为对我们算法最具挑战性的测试平台。在证明了 FIPO 能够在这一具有挑战性的领域中克服长度停滞问题后，我们将在未来的工作中探索并验证这些激发行为在其他开放性或结构更不明确的领域中的表现。

**训练数据范围。**为确保与基准方法进行严格且公平的比较，我们仅将训练限定在 DAPO 所使用的数据集上。因此，我们尚未探索 FIPO 在更大规模或更高质量数据集上的可扩展性。尽管这种受控情景有助于隔离我们方法的算法贡献，但 FIPO 在更广泛或更多样化数据分布上训练时的潜力仍待探索。此外，虽然 FIPO 在数学基准测试中表现优于 o1-mini，但这一优势本质上是领域特定的。由于我们的训练严格局限于数学数据集，我们不预期这些优势能在非数学领域（如编程或符号逻辑）中泛化，而 o1-mini 则得益于大规模、多阶段强化学习。因此，我们将未来工作重点放在探索 FIPO 在更广泛数据情景下的泛化能力及其基本缩放特性上。

**模型范围有限。**本研究的核心目标是探究从无任何长链思维（Long-CoT）合成数据先验影响的纯净基础模型出发，由强化学习驱动的推理过程。这一对实验纯度的严格要求显著限制了合适骨干模型的选择。当前大多数针对推理优化的开源模型均已通过大量监督微调（SFT）或从长序列推理迹中进行蒸馏训练。我们认为，直接从原始基础模型中激发推理的潜在训练动态，与进一步优化已内化蒸馏推理模式的模型存在根本性差异。因此，我们的模型选择被限制在少数高质量的原始基础模型上，例如 Qwen2.5 系列，以确保研究结果能够具体刻画内在推理潜能的涌现，而非预蒸馏的链式思维行为的优化。在未来工作中，我们计划研究该算法应用于此类预蒸馏的长链思维模型时的有效性及其机制行为，探索稠密优势表述是否能够进一步优化或与已存在的蒸馏推理能力产生协同效应。

**性能差距与蒸馏。**尽管基于强化学习的自演化显著提升了推理能力，但其本质上仍是一种“发现式”过程，效率远低于直接蒸馏。更大的教师模型提供了更为稠密的监督信号和更优的启发式信息（Logit），这些对于较小模型仅通过稀疏奖励难以自行推导，导致自训练版本与蒸馏版本之间存在持续的性能差距。

---

## 8 贡献

### 核心贡献者

马驰宇<sup>1,5</sup>，杨硕<sup>2,5</sup>

### 贡献者

黄可欣<sup>5</sup>，陆金达<sup>5</sup>，孟浩明<sup>3,5</sup>，王尚尚<sup>4,5</sup>

### 监督

丁波林<sup>6</sup>，沃苏吉·沃索吉<sup>1</sup>，王国银<sup>5</sup>，周景仁<sup>6</sup>

### 单位

<sup>1</sup> 达特茅斯学院

<sup>2</sup> 北京大学

<sup>3</sup> 多伦多大学

<sup>4</sup> 南加州大学

<sup>5</sup> Qwen Pilot 团队

<sup>6</sup> 阿里巴巴

---

## 参考文献

- Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. arXiv preprint arXiv:2505.15134, 2025.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261, 2025.
- Tiantian Fan, Lingjun Liu, Yu Yue, Jiaze Chen, Chengyi Wang, Qiying Yu, Chi Zhang, Zhiqi Lin, Ruofei Zhu, Yufeng Yuan, et al. Truncated proximal policy optimization. arXiv preprint arXiv:2506.15050, 2025.
- 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, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature, 645(8081):633–638, 2025.
- Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. Skywork open reasoner 1 technical report. arXiv preprint arXiv:2505.22312, 2025.
- Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290, 2025.
- Kexin Huang, Haoming Meng, Jukang Wu, Jinda Lu, Chiyu Ma, Ziqian Chen, Xue Wang, Bolin Ding, Jiancan Wu, Xiang Wang, Xiangnan He, Guoyin Wang, and Jingren Zhou. On the direction of RLVR updates for LLM reasoning: Identification and exploitation. In The Fourteenth International Conference on Learning Representations, 2025. URL <https://openreview.net/forum?id=r6Pw3RiMYL>.
- Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? arXiv preprint arXiv:2411.16489, 2024.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Hel-yar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv preprint arXiv:2412.16720, 2024.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124, 2024.



- 
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding rl-zero-like training: A critical perspective. arXiv preprint arXiv:2503.20783, 2025.
- Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in RLVR fine-tuning of LLMs. In The Fourteenth International Conference on Learning Representations, 2025. URL <https://openreview.net/forum?id=8vWIXno8LW>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. arXiv preprint arXiv:2410.18982, 2024.
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
- Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, et al. Spurious rewards: Rethinking training signals in rlvr. arXiv preprint arXiv:2506.10947, 2025.
- 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. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. arXiv preprint arXiv:2501.12599, 2025a.
- Ling Team, Anqi Shen, Baihui Li, Bin Hu, Bin Jing, Cai Chen, Chao Huang, Chao Zhang, Chaokun Yang, Cheng Lin, et al. Every step evolves: Scaling reinforcement learning for trillion-scale thinking model. arXiv preprint arXiv:2510.18855, 2025b.

- 
- Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation for entropy-safe reasoning. arXiv preprint arXiv:2509.22611, 2025a.
- Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, et al. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. arXiv preprint arXiv:2507.10532, 2025b.
- Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, et al. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. arXiv preprint arXiv:2510.18927, 2025.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024.
- 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.
- Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pp. 6672–6679, 2020.
- Qiyong 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.
- Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What’s behind ppo’s collapse in long-cot? value optimization holds the secret. arXiv preprint arXiv:2503.01491, 2025.
- Yu Yue, Yufeng Yuan, Qiyong Yu, Xiaochen Zuo, Ruofei Zhu, Wenyan 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, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. arXiv preprint arXiv:2508.06471, 2025.
- Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. arXiv preprint arXiv:2505.19590, 2025.
- 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.
- Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. arXiv preprint arXiv:2504.16084, 2025.

---

## 附录目录

A 相关工作	17
B 参数设置	17
B.1 Qwen2.5 32B 基础版 . . . . .	17
B.2 通义千问 2.5 7B 数学 . . . . .	17
C Qwen2.5 7B 数学结果	19
C.1 表现 . . . . .	19
C.2 结果分析 . . . . .	19
D 消融研究	20
D.1 夹具高度、最大长度和响应长度 . . . . .	21
D.2 极值滤波 . . . . .	22
D.3 影响权重裁剪 . . . . .	23
D.4 衰减率的有效范围 . . . . .	23
E 案例研究	24
F 关于失效试验及重现的讨论	25
G 训练成本的更多详情	26

## A 相关工作

**大模型的强化学习.** 强化学习 (RL) 是大语言模型后训练流水线的核心组成部分。尽管早期工作主要利用人类反馈的强化学习 (RLHF) 来对齐模型行为与人类偏好 (Stiennon et al., 2020; Ouyang et al., 2022), 但近期进展已转向通过强化学习提升模型的推理能力。典型代表包括 OpenAI o 系列 (Jaech et al., 2024), 其开创了以推理为中心的方法, 以及 DeepSeek-R1 (Guo et al., 2025), 其提出了一个全面的 RLVR (Lambert et al., 2024) 框架, 通过 GRPO 算法 (Shao et al., 2024) 开发推理模型。这些突破进一步激发了一系列行业领先的研究工作, 如 Kimi (Team et al., 2025a)、Qwen3 (Yang et al., 2025) 和 Gemini 2.5 (Comanici et al., 2025)。

**大规模开源强化学习配方.** 与推理模型的专有进展并行, 开源社区在实现大规模强化学习训练的民主化方面取得了显著进展。这些努力旨在弥合高级算法概念与可高效扩展的稳定实际实现之间的差距, 同时持续改进训练流水线。值得注意的是, GSPO (Zheng et al., 2025)、BAPO (Xi et al., 2025)、SAPO (Gao et al., 2025) 和 OR1 (He et al., 2025) 主要基于已具备长思维链能力的模型开发其强化学习算法。其他工作则投入大量精力从更干净的基础模型出发, 激励复杂的推理能力, 特别是 Qwen2.5-32B-Base。在这些努力中, Open-Reasoner-Zero (Hu et al., 2025)、VC-PPO (Yuan et al., 2025)、VAPO (Yue et al., 2025) 和 T-PPO (Fan et al., 2025) 均在 PPO 框架 (Schulman et al., 2017) 的基础上构建其算法, 而 DAPO (Yu et al., 2025) 则作为 GRPO 的一种改进版本开发。

为确保评估的严谨性, 我们采用 Qwen2.5-32B-Base 作为骨干模型, 并使用 DAPO 作为主要基准。尽管 Open-Reasoner-Zero 回退至 PPO 以避免原始 GRPO 中稀疏优势信号的问题, 我们则通过直接优化 GRPO 框架来解决这一挑战。值得注意的是, 由于 Open-Reasoner-Zero 不依赖辅助价值模型, 其性能最终仍不及 DAPO。相比之下, 其他方法如 VC-PPO、VAPO 和 T-PPO 高度依赖于由已通过 Long-CoT 数据监督微调 (SFT) 的模型预训练的价值模型。我们认为, 这种方法通过价值模型引入了外部知识先验, 可能在评估中造成混淆因素。这使得难以判断性能提升究竟是源于策略最优化算法本身, 还是仅仅继承自预训练的价值模型。通过摒弃对价值模型的需求, 并从原始基础模型出发, FIPO 实现了与这些基于预训练价值模型的方法相当, 甚至在某些情况下更优的性能。**这表明, 建立稠密优势表述是一种有前景的方向, 可推动基于 ORM 的 GRPO 算法演进, 从而释放基础模型固有的推理潜力。**

## B 参数设置

我们详细说明了微调 Qwen2.5-32B-Base 和 Qwen2.5-7B-Math 所采用的具体超参数配置。我们从公开发布的 DAPO 训练脚本出发, 为确保完全可复现性, 我们将发布包含实验中使用的所有训练脚本的完整代码库。

### B.1 Qwen2.5 32B 基础版

我们使用全局批量大小为 512, 每提示词生成  $G = 16$  个样本进行训练。模型采用学习率  $1 \times 10^{-6}$  和权重衰减 0.1 进行优化。对于 FIPO 特有的参数, 我们将未来 KL 衰减率设置为 32.0, 并采用安全阈值 10.0 来过滤极端影响权重。策略更新受到非对称裁剪比率 [0.2, 0.28] 的约束。为支持广泛的推理链, 最大响应长度设置为 20,480 个 token。我们方法与 DAPO 之间的超参数详细对比可参见 Table 2。

表 2: Qwen2.5-32B-Base **实验的超参数设置**。我们将所提出的 FIPO 配置与 DAPO 基准进行对比。大部分基础设施和最优化设置保持一致，以确保比较的公平性。

Hyperparameter	DAPO (Baseline)	FIPO (Ours)
Shared Optimization Settings		
Base Model	Qwen2.5-32B-Base	
Global Batch Size	512	
Group Size (G)	16	
Learning Rate	1e-6	
LR Scheduler	Constant with 10 Warmup Steps	
Weight Decay	0.1	
Gradient Clipping	1.0	
Max Prompt Length	2,048	
Max Response Length	20,480	
Overlong Buffer	4096	
Sampling Temp / Top-p	1.0 / 1.0	
Dual Clip Ratio	10.0	
Policy Clip Ratio	[0.2, 0.28] (Asymmetric)	
KL Penalty Coef	0.0	
Method-Specific Settings		
Mini-Batch Size	32	64
Loss Function	DAPO	Future-KL
Future-KL Decay Rate	-	32.0
Future-KL Clip Ratio	-	[1.0, 1.2]
Safety Threshold	-	10.0

## B.2 通义千问 2.5 7B 数学

我们最初在 Qwen2.5-7B-Math 上进行了初步研究，以验证方法的有效性，随后扩展到 32B 参数规模。在这一初步阶段，我们观察到训练性能起初波动较大，且独立运行之间的推理增益可复现性不一致。为解决这些稳定性问题，我们将组大小增加至  $G = 32$ ，以提供更稳定的最优化信号，并将优势裁剪阈值严格设定为 3.0，以过滤掉导致不稳定的更新。**这些调整成功稳定了训练轨迹，使结果更加可靠，同时保持了相同的性能水平。**模型使用学习率  $1 \times 10^{-6}$  和权重衰减 0.1 进行优化，未来 KL 损失衰减率设置为 32.0。在初步实验的稳定性得到保障后，我们将经过验证的框架扩展至 32B 模型。7B 模型的超参数设置总结见 Table 3，7B MATH 实验的详细结果详见第 C 节。

## C Qwen2.5 7B 数学结果

本节记录了在 Qwen-2.5 7B MATH 模型上的实验结果。由于 32B 模型的计算成本过高，我们早期阶段的探索和消融实验主要使用该变体进行。如第 3.2 节所述，7B 模型在不同训练配置下表现出明显的性能敏感性。为解决此问题，我们进行了针对性的超参数调整，以稳定训练动态并确保所观察趋势的可靠性。注意，我们按照 DAPO verl 脚本的说明，将上下文窗口从 4096 扩展至 32768。

### C.1 表现

Table 4 展示了不同基于强化学习的方法在 AIME 2024 与 AIME 2025 基准上的 Pass@1 性能对比分析。我们提出的 FIPO (7B) 在 AIME 2024 上取得了 40.0% 的显著性能，显著优于 GRPO (7B) 基线 (22.0%) 和 DAPO (7B) 方法 (36.0%)。然而，我们在 AIME 2025 基准上观察到整体性能压缩现象。尽管 FIPO 仍保持领先，达到 19.0%，但相较于 AIME 2024 结果，三种方法之间的性能差距明显缩小。这一现象主要归因于 AIME 2025 问题具有更高的内在难度以及“实时”特性，其复杂度已接近未借助外部引导的 7B



表 3: Qwen2.5-7B-MATH 实验的超参数设置。我们将所提出的 FIPO 与 DAPO 基准进行对比。大部分基础设施和最优化设置保持相同，以确保比较的公平性。

Hyperparameter	DAPO (Baseline)	FIPO (Ours)
Shared Optimization Settings		
Base Model	Qwen2.5-7B-MATH	
Global Batch Size	512	
Learning Rate	1e-6	
LR Scheduler	Constant with 10 Warmup Steps	
Weight Decay	0.1	
Gradient Clipping	1.0	
Max Prompt Length	2,048	
Sampling Temp / Top-p	1.0 / 1.0	
Dual Clip Ratio	10.0	
Policy Clip Ratio	[0.2, 0.28] (Asymmetric)	
KL Penalty Coef	0.0	
Method-Specific Settings		
Mini-Batch Size	32	64
Loss Function	DAPO	Future-KL
Future-KL Decay Rate	-	32.0
Future-KL Clip Ratio	-	[0.8, 1.2]
Safety Threshold	-	3.0
Group Size (G)	16	32
Max Response Length	8192	10240
Overlong Buffer	4096	2048

参数模型的推理上限。在此规模下，尤其是在缺乏更先进思维链结构预训练的情况下，AIME 2025 所需推理链的复杂性增加，使得不同最优化算法之间的微小提升难以区分，因为大多数模型均面临类似的结构性瓶颈。

表 4: Qwen 2.5-7B-MATH 在 AIME2024 与 AIME2025 上的 Pass@1 性能对比。所有结果均以百分比 (%) 报告。我们报告了 32 个样本下的峰值平均 Pass@1 (Avg@32)。为降低对数字级别生成方差的敏感性，最终数值四舍五入到最接近的整数。

Method	AIME 2024 (Pass@1)	AIME 2025 (Pass@1)
GRPO (Guo et al., 2025)	22.0%	18.0%
DAPO	36.0%	18.0%
FIPO (Ours)	40.0%	19.0%

## C.2 结果分析

如图Figure 6所示，两种算法均保持了稳定的长度加权优势，且平均响应长度在 1200 个 token 左右波动。值得注意的是，这与 32B 模型训练形成了鲜明对比，在后者中，FIPO 通常会引发响应长度的持续增长。我们推测，这种长度停滞并非优化失败，而是 7B 模型固有容量限制和训练先验的体现。具体而言，Qwen-2.5 7B MATH 基础模型在预训练时使用了受限的 4K 上下文窗口 (Yang et al., 2024)，这很可能在其推理深度上设定了物理上限，除非有外部引导。此外，该模型对基于代码的推理具有强烈的初始偏好 (Shao et al., 2025)，这种偏好倾向于逻辑稠密、确定性的路径而非冗长的探索，再加上潜在的 AIME24 数据泄露 (Wu et al., 2025b)，为模型提供了”高置信度捷径”，使其能够快速获得简洁解，而非探索更长、迭代式的路径。

这种“高效但受限”的推理现象在 Figure 7(c) 所示的熵动态中得到了进一步证实。尽管更大的模型通常依赖持续增长的熵来探索复杂的推理空间，但 7B MATH 模型在 FIPO 下通过收敛到显著更低熵的策略

实现了最优性能。这种发散表明了缩放行为的根本差异：虽然 32B 模型受益于广泛的探索，但 7B 模型似乎更注重对特定、高置信度推理流形的精炼。

在该规模下，低熵状态对性能至关重要的假设，进一步得到了我们关于未来 KL 影响权重裁剪的消融研究的支持。如Appendix D所述，7B 和 32B 模型对未来 KL 裁剪值域的敏感性存在差异。尽管 32B 模型在影响权重裁剪范围为 1.0 至 1.2 时仍保持优异性能，但将相同范围应用于 7B 模型时，熵持续增加，其行为与 32B 模型类似，却导致性能显著下降。相反，我们发现 7B 模型在另一种配置下表现最优，即采用 0.8 至 1.2 的裁剪范围。这一观察表明，7B 模型可能缺乏足够的内在自探索容量，无法从高熵状态中获益。在此配置下，维持较高熵似乎引入了更多有害噪声而非有益发现。因此，在此规模下，模型最有可能通过收敛至特定的、低熵推理迹线来实现卓越性能。这一现象从根本上符合熵最小化 (Agarwal et al., 2025)、熵调控 (Wu et al., 2025a)、自引导 (Zuo et al., 2025) 以及自确定性最优化 (Zhao et al., 2025) 的原则，这些原则在相似规模的模型中通常被证明是有效的。

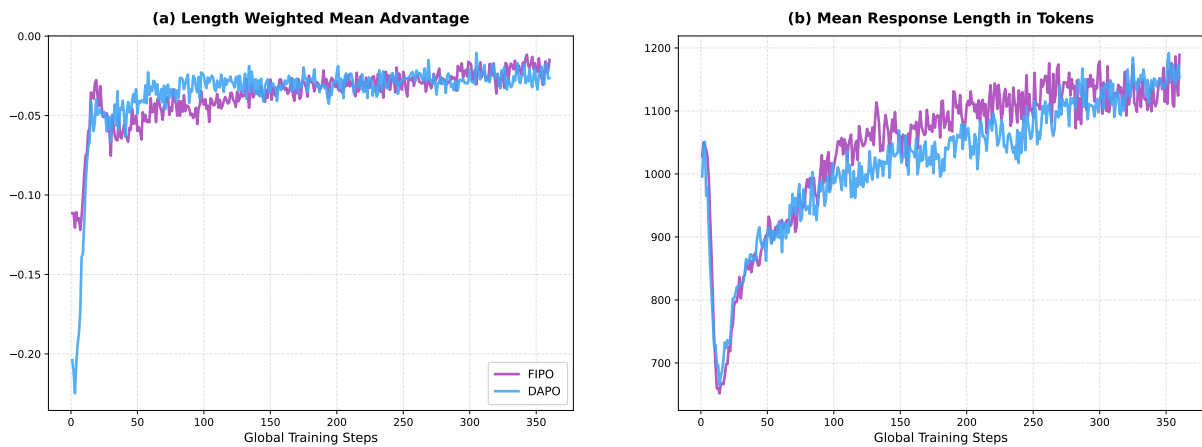


图 6: Qwen2.5-7B-MATH 的长度加权优势与平均响应长度分析。(a) 子图 (a) 显示 FIPO 保持了稳定的长度加权平均优势。(b) 子图 (b) 显示平均响应长度稳定在约 1200 个 token 左右。这种长度增长的抑制很可能是由于基础模型的 4K 上下文窗口限制，以及其固有的对简洁代码推理的偏好，从而限制了生成深度。

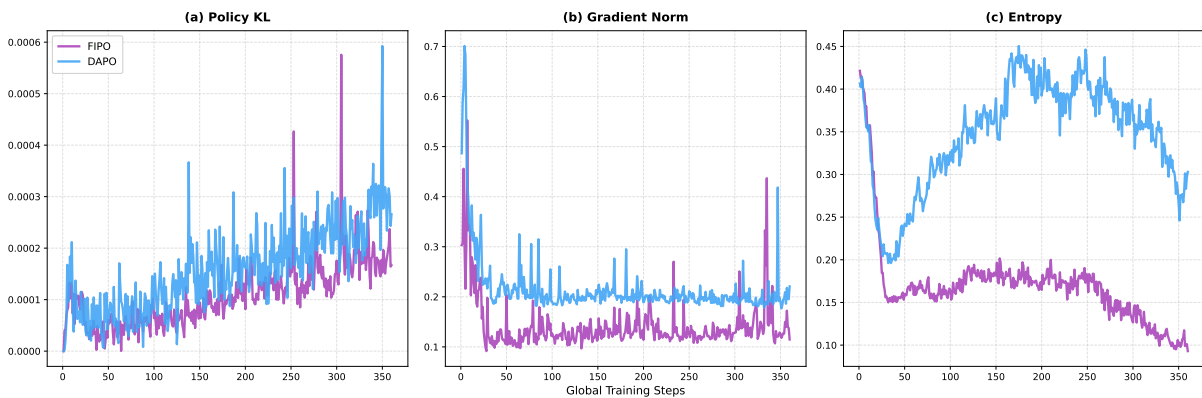


图 7: Qwen2.5-7B-MATH 的策略演化与最优化动力学。(a) 策略 KL 和 (b) 梯度范数展示了最优化路径。(c) 显著的是，FIPO 的熵低于 DAPO，表明在强化学习过程中，模型收敛于更具确定性的推理迹。

## D 消融研究

在本节中，我们对为验证我们的方法而开展的消融实验进行了全面分析。首先，我们考察了高价值截断以及最大响应长度扩展对使用 FIPO 进行 32B 模型训练时平均响应长度的影响。接着，我们展示了 7B 模

型规模下的权重过滤结果，这些结果补充了我们在 32B 实验中观察到的稳定性相关发现。此外，我们研究了不同自适应截断配置的含义及其对性能的影响，特别强调了这些选择如何导致策略熵出现发散行为。最后，我们评估了优化过程对不同衰减率选项的敏感性及其对收敛动态的潜在影响。

## D.1 夹具高度、最大长度和响应长度

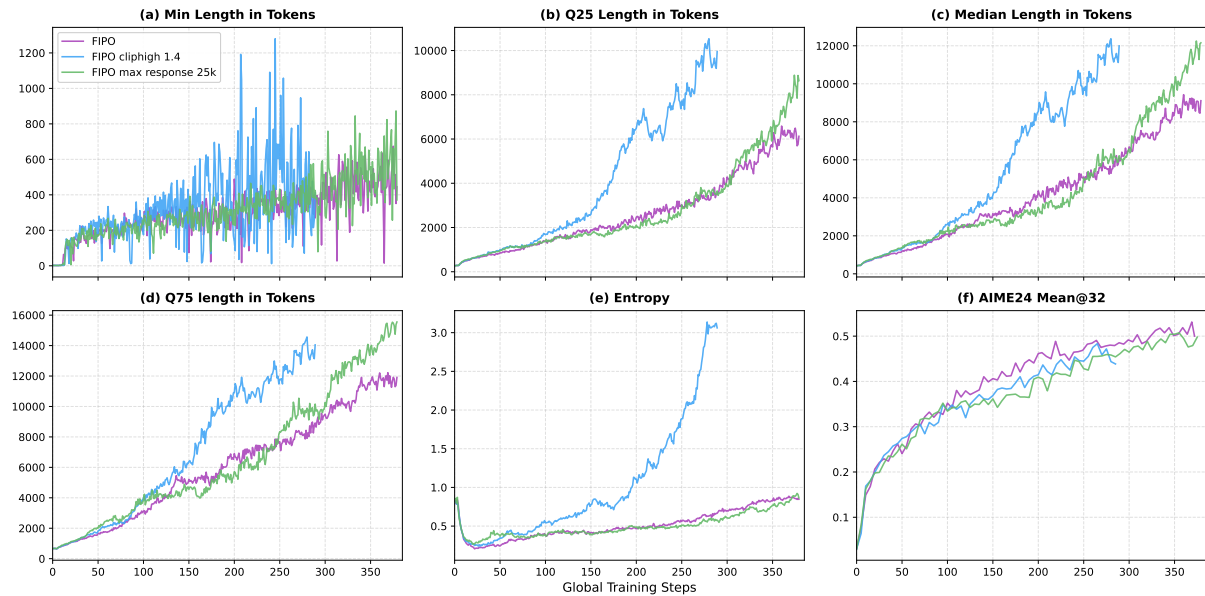


图 8: Qwen2.5-32B-Base 上响应长度控制的消融实验。我们将基准 FIPO 与裁剪值为 1.4 的 FIPO 及最大响应长度为 25K 的 FIPO 进行对比。(a-d) 提高最大长度限制会引发响应长度的显著早期激增，但导致训练效率降低且性能提升微弱。(e) 更高的裁剪比例会引起策略熵急剧上升，表明存在潜在的训练不稳定性。(f) 总体而言，尽管两种修改均促使生成更长的响应，但在提升 AIME24 性能方面不如平衡的 FIPO 配置有效。

我们观察到的一个最有趣的现象是在早期阶段，当提高 PPO 目标中的 clip-high 比率  $\epsilon_{high}$  时，响应长度出现意外的激增。在本研究中，我们设置  $\epsilon_{high} = 1.4$ ，使得在优势为正时，策略更新拥有更大的信任区域。对于最大长度试验，我们将限制延长至 25K token（从默认的 20K 扩展），同时保持过长缓冲区比例恒定在 20%。所有其他变量保持一致，以确保公平比较。如图 Figure 8(a-d) 所示，更高的  $\epsilon_{high}$  允许策略在接收到正向优势时更激进地偏离旧分布，从而激励模型几乎立即“过度探索”更长的推理路径。然而，这种激增是以训练稳定性为代价的；如图 Figure 8(e) 所示，较高的  $\epsilon_{high}$  会导致策略熵爆炸，表明当剪裁约束过于宽松时，优化地形变得极为不稳定。

当增加最大响应长度限制时，观察到一种极为相似的失败模式。尽管这同样会引发 token 数量的迅速增长，但推理效率却呈现出明显的递减回报。两种情况均暴露出表面化推理行为的出现，例如重复内容、与任务无关的 LaTeX 格式化以及过早的自我反思。我们认为，在模型尚未建立稳固的一步推理基础之前引入自我反思可能是适得其反的。在这种情况下，模型缺乏可验证的内部逻辑来评估自身状态，从而导致内部一致性冲突——即模型在相互矛盾的步骤间来回震荡，无法朝着正确解前进。**这些观察表明，在训练过程中，高阶推理模式（如自我反思）的出现并非“越早越好”。**相反，它应是一个渐进成熟的过程，复杂的策略需建立在坚实的基础逻辑之上。因此，相较于平衡的 FIPO 基准，模型在 AIME24 (Figure 8f) 上的性能提升微乎其微。**这些结果突显了对响应长度进行可控增长的重要性，确保扩展的计算被分配给真正的逻辑深度，而非冗余的自我修正循环。**

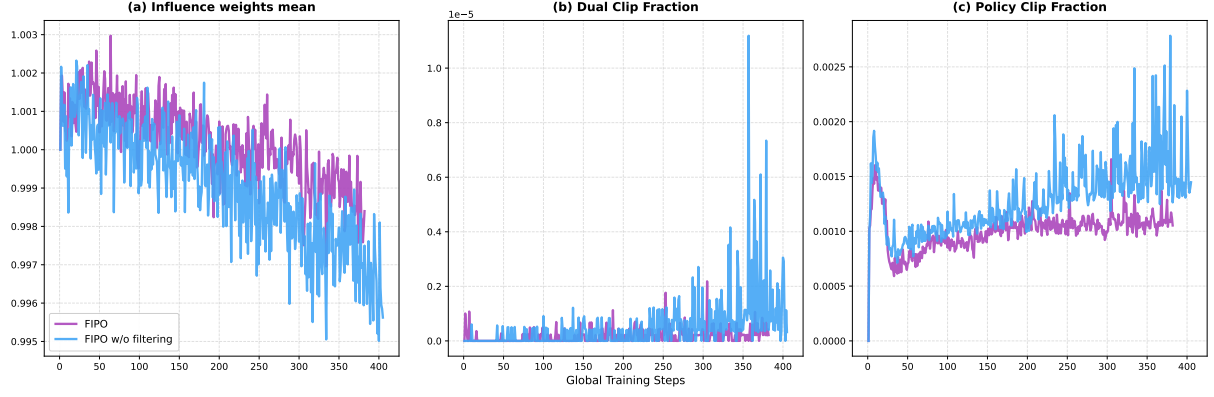


图 9: Qwen2.5-7B-MATH 上 Future-KL 过滤机制的消融实验。我们对比了 FIPO 在有无过滤极端重要性采样 (IS) 比率情况下的训练稳定性。(a) Future-KL 影响权重: 移除极端 IS 比率后, 权重值域显著更加受限且紧凑, 从而带来更稳定的策略更新。(b-c) 截断统计量: 极端 IS 比率会带来不成比例的影响权重, 导致更新更加激进。这体现在双截断和策略截断的比例上升, 当未启用过滤时, 更多 token 被推离信任区域。

## D.2 极值滤波

正如我们在 subsection 3.2 中指出的, 由于 Future-KL 的计算高度依赖于每次更新之间的 logits 值, 其一个缺点是对重要性比率波动的敏感性。如果重要性比率的变化变得显著, Future-KL 权重也会随之产生剧烈波动。与其为策略更新提供稳定的指导信号, 重要性比率中的过度噪声反而导致 Future-KL 项失去有效性。在这种情况下, 引入未来 KL 信息可能导致更新过程更加不稳定, 模型性能往往退回到接近标准基准的次优水平, 未能展现出预期的改进。这一点在 7B 模型中尤为明显, 这些模型表现出频繁的重要性比率波动。

Table 5 展示了在 7B 模型上的消融实验结果。我们观察到, 如果不过滤极端的 IS 比率, 相较于应用过滤机制的配置, 其在 AIME2024 上的表现提升并不理想。尽管在某一次实例中, 未过滤版本在 AIME2025 上取得了略高的得分, 但整体可靠性以及在 AIME2024 上的峰值性能 (40.0%) 结合 32B 训练中的观察结果, 证实了过滤机制对于更可靠训练的必要性。

此外, 如图 Figure 9(a) 所示, 去除极端 IS 比率后, 权重的值域显著变得更加受限和紧凑, 为优势估计量提供了更稳定的信号。相比之下, 缺乏过滤会导致影响权重不成比例地增大。这导致了更频繁的激进策略更新, 如图 Figure 9(b) 和 (c) 中的截断比例所证实。当缺少过滤时, 更大比例的 token 被推至指定的信任区域之外, 从而导致更高的双重截断和策略截断比例。这表明模型经常尝试进行超出 PPO 目标约束的更新, 从而解释了当 IS 比率波动较大时性能提升不理想的原因。

表 5: 极端值过滤与影响权重裁剪的消融实验: 在 Qwen2.5-7B-MATH 上, AIME2024 和 AIME2025 的 Pass@1 性能对比。所有结果均以百分比 (%) 表示。我们报告 32 个样本下的峰值平均 Pass@1 (Avg@32)。

Method	AIME 2024 (Pass@1)	AIME 2025 (Pass@1)
FIPO ( $\epsilon_{low} = 1.0, \epsilon_{high} = 1.2$ )	36.0%	19.0%
FIPO (w/o filtering, $\epsilon_{low} = 0.8, \epsilon_{high} = 1.2$ )	38.0 %	21.0%
FIPO ( $\epsilon_{low} = 0.8, \epsilon_{high} = 1.2$ )	40.0%	19.0%



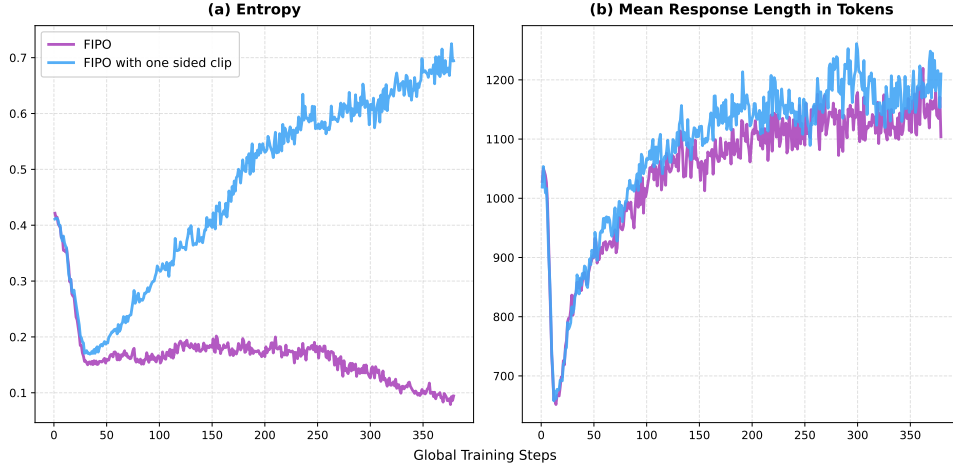


图 10: Qwen-2.5-7B-MATH 上影响权重裁剪的消融实验。(a) 与 32B 类似, 一个值域为  $[1.0, 1.2]$  的影响权重导致熵持续增长, 从而鼓励更多探索。(b) 然而, 正如前一节所讨论的, 我们并未观察到响应长度的激增。

### D.3 影响权重裁剪

Table 5 展示了对影响权重裁剪的消融实验结果。我们主要针对 1 修改了裁剪值域。一个值域为  $[1.0, 1.2]$  的影响权重, 本质上对负例提供更大的惩罚, 同时给予正例更多的奖励。值域为  $[0.8, 1.2]$  则继承了这些特性, 同时提供了更平衡的影响; 它在正例中的某个 token 与后续的负向行为相关联时进一步降低奖励, 在负例中的某个 token 与后续的正向行为相关联时减少惩罚。这有助于实现更受控的探索, 从而带来了 Table 5 中所示的性能提升。

具体而言,  $[0.8, 1.2]$  配置在 AIME 2024 上取得了 40.0% 的表现, 超越了 36.0% 的  $[1.0, 1.2]$  情景。尽管后者鼓励更激进的探索, 如 Figure 10(a) 所示政策熵持续增长所证明, 但我们的结果表明, 7B 模型对这种过度探索压力较为敏感。这与我们对缩放行为的总体观察一致: 与能够从广泛探索中获益的大模型不同, 7B 模型似乎更注重特定高置信度推理流形的精炼。在此背景下,  $[1.0, 1.2]$  值域引发的更高熵带来了更多有害噪声, 而非有效发现。相比之下,  $[0.8, 1.2]$  的均衡影响促进了向低熵状态的收敛, 优化的是自我确信心, 而非随机搜索。

### D.4 衰减率的有效范围

为了补充我们的分析, 我们进一步对衰减率的有效时域进行了消融研究, 具体测试了  $\tau \in \{8, 32, 128, 256\}$  这些取值。这些取值定义了影响权重衰减的半衰期, 决定了未来 KL 信息仍具有显著意义的 token 距离。Table 6 展示了这些配置下的性能表现, 而 Figure 11 则描绘了相应的训练动态。如 Figure 11(a) 所示, 随着衰减时域  $\tau$  的延长, 均值影响权重的波动幅度增大。对于最大的时域 ( $\tau = 256$ ), 权重均值与 1.0 的偏离最为显著, 这可能会在策略更新中引入不稳定性。相比之下, 当时域非常小时 ( $\tau = 8$ ), 影响权重始终接近 1.0, 导致模型始终贴近基准, 相对未能充分利用未来的 KL 信息。

Figure 11(c) 中的熵动力学进一步证实了我们关于 7B 模型探索能力的早期观察。尽管所有配置均呈现出先下降、后上升、最终回落的典型趋势, 但较大衰减率下的整体熵值明显更高。在训练的大部分过程中,  $\tau = 256$  情景始终保持最高的熵值, 表明其由长程未来信号驱动的激进探索行为。正如我们在限制  $[1.0, 1.2]$  影响权重值域时所观察到的那样, 这种持续的高熵状态引入了过度的波动性, 7B 模型难以适应。



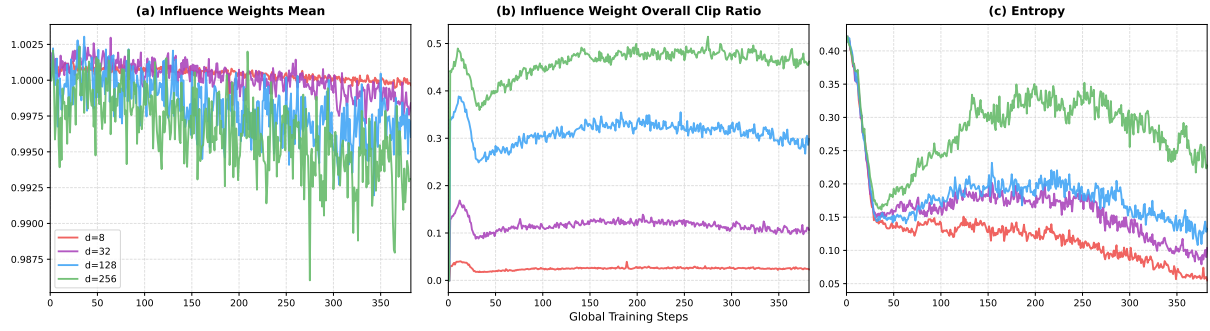


图 11: Qwen2.5-7B-MATH 上衰减率范围 ( $\tau$ ) 的消融实验。(a) 平均影响权重: 延长衰减范围 (例如,  $\tau = 256$ ) 会显著增加权重波动, 可能造成策略更新的不稳定性。相反, 过为严格的范围 (例如,  $\tau = 8$ ) 会使权重过于贴近基准。(b) 影响权重裁剪比率: 更大的衰减范围导致更高的裁剪频率。这反映了原始影响权重的剧烈变化, 直接验证了 (a) 中观察到的趋势。(c) 熵: 适中的范围 ( $\tau = 32$ ) 达到了关键平衡。它避免了由较大衰减率引发的过度波动, 也规避了由较短范围带来的短视引导所引起的熵过早坍塌, 使 7B 模型能够保持受控且有效的探索。

相反,  $\tau = 8$  情景中观察到的熵值快速下降代表了完全不同的失败模式: 由于其影响权重仅发生微小波动, 模型接收到的是高度短视的指导。缺乏足够的长期远见, 策略过早坍塌至次优的低熵状态。而中间时域  $\tau = 32$  则达到了关键的平衡。类似于  $[0.8, 1.2]$  截断值域的稳定作用,  $\tau = 32$  提供了足够的局部未来信号, 使策略能够安全地进行更新, 避免过早停滞, 从而最终收敛至特定的高质量推理流形。这一跨不同超参数的一致性进一步强化了我们的假设: 在较小规模 (如 7B) 下, 通过受控的、适度的探索来优化自我确定性, 比强制进行广泛且高熵的搜索更为有效。

表 6: 衰减率有效时域的消融实验: Qwen2.5-7B-MATH 在 AIME2024 和 AIME2025 上 Pass@1 性能对比。所有结果均以百分比 (%) 表示。我们报告 32 个样本的峰值平均 Pass@1 (Avg@32)。

Method	AIME 2024 (Pass@1)	AIME 2025 (Pass@1)
FIPO ( $\tau=8$ )	40.0%	17.0%
FIPO ( $\tau=32$ )	40.0%	19.0%
FIPO ( $\tau=128$ )	39.0%	21.0%
FIPO ( $\tau=256$ )	42.0%	16.0%

## E 案例研究

在本节中, 我们通过分析模型在 2024 年 AIME 竞赛中的具体输出, 对模型推理演化进行了定性分析。我们从分组样本中随机选取了部分响应。从 DAPO 的长度停滞到 FIPO 的持续缩放, 并非仅仅是 token 数量的量变, 更是模型如何利用其“思考”预算的质变。

**阶段 1: 表面规划 (初始步骤)。**如 Figure 13 所示, 模型在初始步骤中表现出“表面规划”行为。它生成了解题步骤的模板式大纲, 但未能执行实际的数学推导。这导致响应内容简短, 缺乏逻辑实质, 通常会立即产生幻觉结论。

**阶段 2: 线性执行 (DAPO 收敛与早期 FIPO)。**在其整个训练过程中, DAPO 始终处于此阶段 (Figure 14), 逐渐演变为一个“线性执行器”, 能够准确遵循标准的思维链 (CoT) 以达到真实值。然而, 其推理本质上仅限于“单次遍历”逻辑, 一旦找到首个结果便立即终止生成。这解释了在 Figure 3 中观察到的 DAPO 长度停滞现象。

**阶段 3：涌现的自我反思（中级阶段 FIPO）。**随着 FIPO 训练进入中级阶段 (Figure 15)，模型的行为出现显著转变。模型开始利用不断增长的 token 预算（随响应长度增加）进行自发的自我反思。在得出初步结果后，模型主动启动验证阶段，探索替代的方法论，例如从代数运算切换到几何解释，以交叉验证其结论。

**阶段 4：系统化深度推理（后期 FIPO）。**在训练的后期阶段 (Figure 16)，模型逐渐发展为一种“计算密集型”策略，优先考虑分析的深度。推理轨迹不再局限于简单的反思，而是扩展至系统化的审计过程。模型会进行多次符号重推导和细致的算术验证（例如，手动逐步展开大的平方和平方根）。这种自发出现的自我验证机制与先进推理模型在推理时表现出的缩放行为一致，即模型将长度视为关键资源，以确保更优的表现。

## F 关于失效试验及重现的讨论



图 12: Qwen2.5-32B-Base 上不同小批量大小的训练动态与稳定性。(a) **响应长度均值**: 虽然小批量大小为 32 时偶尔能取得优异性能，但其存在严重的可复现性问题，并且频繁遭遇长度停滞。值得注意的是，仅移除过长惩罚 (w/o Overlong) 无法解决这一瓶颈。(b) **策略裁剪比例**: 在成功的 32 小批量试验中，被裁剪的 token 数量少于失败的情况，这一现象与我们在 clip\_high 消融研究中的发现一致。由于 64 的小批量大小更接近同策略，且受重要性采样 (IS) 权重波动的影响较小，因此自然导致较少的 token 被裁剪，从而缓解了响应长度和性能的缩放难题。(c) **梯度范数**: 同样，成功的 32 小批量运行保持了明显更稳定且略低的梯度范数，这是由 64 小批量基准所固有实现的关键稳定特性。(d) **策略熵**: 成功案例在整个训练过程中维持更高的熵，表明具备更健康的探索能力。(e-f) **AIME24 性能**: 由于优化过程稳定且探索持续，成功的 32 小批量案例达到了约 60% 的峰值 Mean@32 和 70% 的 Cons@32，与我们 64 小批量基准的最终性能高度相当。

如前所述，由于相对稳定的特性，我们在 32B 模型训练及后续实验中采用 64 的批量大小而非 32。为了更好地理解这一参数选择，我们在此详细说明采用 32 批量大小时遇到的具体训练结果与问题。如 Figure 12(a) 所示，尽管 32 的批量大小偶尔能产生较强性能，但其存在严重的可复现性问题。最显著的失败模式是长度增长严重减速，模型难以有效拓展至更长的推理轨迹。值得注意的是，仅移除过长惩罚并不能解决这一瓶颈，表明问题源于更深层次的优化动态，而非简单的奖励惩罚机制。

我们进一步通过剪裁高消融研究中识别出的潜在原理来解释这些发现（参见 subsection D.1）。尽管该消融研究关注的是剪裁阈值而非小批量大小，但它揭示了一个核心最优化动态：当更多 token 能够在不受剪裁边界限制的情况下有效参与策略更新时，模型会获得显著动量以扩展其响应长度。观察策略剪裁比例 (Figure 12(b))，我们发现成功的批量大小 32 试验相比失败的试验剪裁了较少的 token。我们推断，当重要性采样 (IS) 权重波动减小时，相对更多的 token 保持未被剪裁，并被完全纳入训练信号。这种有效包含对于克服前述长度增长减速现象极为有利。由于批量大小为 64 的试验在更广泛的样本集上计算梯度，其内在更接近同策略，且受 IS 方差影响较小。因此，它自然避免了过度的 token 剪裁，提供了持续最优化动量，从而能够平滑地扩展推理长度与性能。

优化稳定性上的差异在梯度范数和策略熵中也十分明显。如 Figure 12(c) 所示，成功的批量大小为 32 的运行保持了明显更稳定且略低的梯度范数，相较于失败的运行。此外，它在整个训练过程中维持了更高的熵值 (Figure 12(d))，表明具备更健康、更有效的探索能力，且无剧烈发散现象。至关重要的是，这些稳定特性——受控的梯度更新和持续活跃的探索——本质上由批量大小为 64 的基准方法所实现，从而无需依赖较小批量大小下的随机成功。最终，得益于稳定的最优化和一致的探索，成功的批量大小为 32 的案例达到了约 60% 的峰值 Mean@32 和 70% 的 Cons@32 (Figure 12(e-f))。这些指标与我们批量大小为 64 的基准方法的稳定表现高度相当。因此，我们采用 64 的批量大小并非出于更高的理论性能上限，而是因其能够可靠地引导策略进入高质量的推理流形，避免陷入优化陷阱。

这一现象与现有的自适应截断技术存在本质区别 (Xi et al., 2025; Gao et al., 2025)。虽然自适应截断通过动态调整阈值主要旨在维持信任区域的稳定性并防止策略坍塌，但我们的研究揭示了一个根本不同的机制：有效的截断比例实际上充当了长度扩展的结构性阀门。通过采用 64 的批量大小以自然地稳定 IS 权重，我们允许更大比例的有效 token 安全地通过固定的边界。此外，正是在这些稳定化的优化条件下，我们提出的未来 KL 机制才能充分发挥其作为推理训练主要驱动力的预期作用。根据设计，未来 KL 引入了前瞻信号来指导当前 token 的更新，从而自然地鼓励对更长推理轨迹的探索。然而，该机制本质上依赖于长序列中梯度的连续且完整的流动。稳定的 IS 权重为此提供了精确的基础：**通过防止过度的 token 截断，它确保了未来 KL 产生的长期探索信号能够成功且平滑地向后传播，而不会被人中断。**在这一无阻碍优化动量的支持下，未来 KL 明确推动策略向更深层次、高质量的推理流形扩展，有效防止模型回退至过早的短形式响应。

## G 训练成本的更多详情

一种原始的 Future KL 实现需要计算一个稠密的  $(L, L)$  时间衰减矩阵（其中  $L$  为响应长度），导致  $\mathcal{O}(L^2)$  的内存占用，这在长轨迹推理训练过程中极易引发内存不足 (OOM) 错误。为缓解此问题，我们实现了一种基于分块的内存高效算法，如 Listing 1 所示。通过将响应序列划分为固定大小块 ( $K$ )，我们逐步计算距离掩码和衰减权重。块级贡献通过形状为  $(B, K) \times (K, L)$  的并行矩阵乘法计算。该算法在严格限制峰值内存复杂度至  $\mathcal{O}(B \cdot L + L \cdot K)$  的同时，保持了 Future KL 公式的确切解析结果。尽管时间复杂度仍为  $\mathcal{O}(B \cdot L^2)$ ，但张量化的分块操作在现代 GPU 上高度优化，有效消除了扩展推理长度时的内存瓶颈。

虽然标准 GRPO 使用严格逐元素运算计算策略目标，具有  $\mathcal{O}(B \cdot L)$  的时间复杂度，但我们引入的未来 KL 机制自然地引入了一个  $\mathcal{O}(B \cdot L^2)$  的时序聚合过程。因此，这在演员更新阶段带来了一定程度的计算

---

开销。然而，我们的分块矩阵乘法实现有效地向量化了这些操作，充分利用了现代 GPU 高度优化的稠密矩阵乘法能力。经验上，GRPO 训练迭代期间的墙钟时间 slowdown 相对较小且完全可以接受。最重要的是，我们认为这种适度的训练速度损失是值得的：**它提供了密集的、长时程的信用分配，这是扩展复杂推理所必需的，而标准  $\mathcal{O}(B \cdot L)$  GRPO 本质上难以克服这一根本瓶颈。**自然地，尽管我们的分块实现有效解决了内存限制，更复杂的计算优化仍有可能实现。

---

Listing 1: Python code for computing memory-efficient chunked Future KL

---

```
import torch

def compute_future_kl_chunked(D, M, K, tau):
    # Calculate the decay factor
    gamma = 2 ** (-1 / tau)

    # Mask out invalid tokens and filtered anomalies
    D = D * M

    # Initialize the Future KL accumulator
    F = torch.zeros_like(D)
    L = D.shape[1]

    # Column vector of query positions (L x 1)
    i = torch.arange(L, device=D.device).unsqueeze(1)

    for j_start in range(0, L, K):
        j_end = min(j_start + K, L)

        # Row vector of chunk positions (1 x K_cur)
        j = torch.arange(j_start, j_end, device=D.device).unsqueeze(0)

        # Broadcasted distance matrix (L x K_cur)
        Delta = j - i

        # Decay weight block
        W = (gamma ** torch.clamp(Delta, min=0)) * (Delta >= 0).float()

        # Extract KL values for the current chunk
        V = D[:, j_start:j_end]

        # Parallel matrix multiplication update
        F += torch.matmul(V, W.T)

    return F
```

---

---

**Question:**

Find the largest possible real part of  $[(75+117i)z+\frac{96+144i}{z}]$  where  $z$  is a complex number with  $|z|=4$ .

**Ground Truth:** 540

---

**Response:**

Step 1: The problem requires finding the largest possible real part of a complex expression involving a complex number  $z$  with a given magnitude.

Step 2: First, we need to express  $z$  in polar form as  $z = 4e^{i\theta}$ , where  $\theta$  is the argument of  $z$ .

Step 3: Then, we can rewrite the given expression in terms of  $\theta$  and find the real part.

Step 4: To find the maximum real part, we need to analyze the expression and identify the conditions that maximize the real part.

Step 5: We can use trigonometric identities and properties of the real and imaginary parts of complex numbers to simplify the expression and find the maximum real part.

Step 6: By analyzing the expression and using the given conditions, we can determine the value of  $\theta$  that maximizes the real part and calculate the maximum real part.

Step 7: Finally, we can provide the answer as the largest possible real part of the given expression.

**Answer:** 1176

图 13: 第一阶段: 表面规划 (初始步骤)。Qwen 2.5 32B 的基准响应。该模型生成了解决方案步骤的“模板式”大纲, 但未能执行任何实际的数学推导, 导致得出错误结论。



---

**Question:**

Find the largest possible real part of  $(75+117i)z + \frac{96+144i}{z}$  where  $z$  is a complex number with  $|z|=4$ .

**Ground Truth:** 540

---

**Response:** Let's break down and solve the problem step-by-step.

Let  $z = 4e^{i\theta} = 4(\cos \theta + i \sin \theta)$ . Substituting this into the expression:  $(75 + 117i) \cdot 4(\cos \theta + i \sin \theta) + \frac{96 + 144i}{4(\cos \theta + i \sin \theta)}$

**1. Simplifying the first term:**

$$[4(75 + 117i)(\cos \theta + i \sin \theta)] = 4[(75 \cos \theta - 117 \sin \theta) + i(75 \sin \theta + 117 \cos \theta)] = 4(75 \cos \theta - 117 \sin \theta) + i \cdot 4(75 \sin \theta + 117 \cos \theta)$$

**2. Simplifying the second term:**

$$\left[ \frac{96 + 144i}{4} \cdot (\cos \theta - i \sin \theta) \right] = 24(1 + 1.5i)(\cos \theta - i \sin \theta) = 24[(\cos \theta + 1.5 \sin \theta) + i(1.5 \cos \theta - \sin \theta)]$$

**3. Combining the real parts:**

$$[\text{Re}] = 4(75 \cos \theta - 117 \sin \theta) + 24(\cos \theta + 1.5 \sin \theta) \quad [\text{Re}] = 300 \cos \theta - 468 \sin \theta + 24 \cos \theta + 36 \sin \theta \quad [\text{Re}] = 324 \cos \theta - 432 \sin \theta$$

**4. Finding the maximum value:**

The maximum value of  $a \cos \theta + b \sin \theta$  is  $\sqrt{a^2 + b^2}$ . For  $a = 324$  and  $b = -432$ :  $[\text{Max}] = \sqrt{324^2 + (-432)^2} = \sqrt{104976 + 186624} = \sqrt{291600} = 540$

**Answer:** 540

图 14: 第二阶段: 线性执行 (DAPO 收敛, FIPO 初期阶段)。当模型正确执行标准的思维链 (CoT) 以达到真实值时, 其推理过程是严格线性的, 一旦找到首个结果便立即终止, 缺乏任何自我验证机制。

---

**Question:**

Find the largest possible real part of  $[(75+117i)z+\frac{96+144i}{z}]$  where  $z$  is a complex number with  $|z|=4$ .

**Ground Truth:** 540

---

**Response:**

**Let's solve the problem step by step.**

We want to find the largest possible real part of the expression. Let  $z = 4e^{i\theta}$ . Since  $|z|=4$ ,  $\frac{1}{z} = \frac{\bar{z}}{16}$ .

The expression becomes:  $[(75 + 117i)z + (6 + 9i)\overline{z}]$ .

Let us write  $z$  in its standard form  $z = x + yi$ . Expanding the terms, the real part of the expression is  $81x - 108y$ .

We need to maximize this subject to the constraint  $x^2 + y^2 = 16$ .

This is a classic problem for Lagrange multipliers. We want to find  $\lambda$  such that  $\nabla f = \lambda \nabla g$ .

...

Solving the system yielded  $x=\frac{12}{5}$  and  $y=-\frac{16}{5}$ , giving a value of **540**.

It seems that at the point  $(\frac{12}{5}, -\frac{16}{5})$  we have a value of 540. Hence, the maximum value of the expression appears to be 540.

**Let's confirm if this is indeed the highest value by looking at the geometry of the situation or another approach.**

Let us represent the problem in another manner. Let us write  $z = 4e^{i\theta}$ .

Then the full expression is expanded in terms of  $\sin$  and  $\cos$ .

The real part is given by:  $[324 \cos \theta - 432 \sin \theta]$

To maximize this, we can write it in the form  $R \cos(\theta + \phi)$  where  $R^2 = 324^2 + 432^2$ .

Let's compute this value step by step.  $324^2 = 104976$   $432^2 = 186624$   $R^2 = 291600 \implies R = \sqrt{291600} = 540$ .

Thus, the amplitude  $R$  is 540.

This confirms our previous result. Therefore, the largest possible real part of  $(75 + 117i)z + \frac{96 + 144i}{z}$  is **indeed** 540.

**Answer:** 540

图 15: 阶段 3: 涌现的自我反思 (FIPO 中间阶段)。随着 FIPO 训练的推进, 模型的行为发生显著转变。模型开始利用扩展的 token 预算进行自发的自我反思, 通过其他数学路径对初始结论进行交叉验证。

Question:

Find the largest possible real part of  $[(75+117i)z + \frac{96+144i}{z}]$  where  $z$  is a complex number with  $|z|=4$ .

Ground Truth: 540

Response:

Let's solve the given problem step by step.

Let us denote  $z$  as a complex number with  $|z| = 4$ . It implies we can write  $z = 4e^{i\theta}$ . Then  $\frac{1}{z} = \frac{1}{4}e^{-i\theta}$ .

Substituting into the expression:

$[(75+117i) \cdot 4e^{i\theta} + \frac{96+144i}{4}e^{-i\theta}] = (300+468i)e^{i\theta} + (24+36i)e^{-i\theta}$  Using Euler's formula  $e^{i\theta} = \cos\theta + i\sin\theta$ , we find the real part:

- From first term:  $300\cos\theta - 468\sin\theta$ .
- From second term:  $24\cos\theta + 36\sin\theta$ .

Total Real Part =  $(300+24)\cos\theta + (-468+36)\sin\theta = 324\cos\theta - 432\sin\theta$ .

The maximum value of  $A\cos\theta + B\sin\theta$  is  $\sqrt{A^2 + B^2}$ .

$R = \sqrt{324^2 + (-432)^2} = \sqrt{(108 \cdot 3)^2 + (108 \cdot 4)^2} = 108\sqrt{3^2 + 4^2} = 108 \cdot 5 = 540$ .

To verify this reasoning: Let's look at the expression more directly...

let us represent it in the standard form  $R\cos(\theta + \alpha)$ ...

[Model re-derives the amplitude formula] ...

This confirms that  $R = 540$ .

To ensure there is no mistake, let me double check my simplification process carefully:

Recall  $R = \sqrt{324^2 + 432^2}$ . Let's recompute it step by step:

- $324^2 = 104976$ .
- $432^2 = 186624$ .
- $324^2 + 432^2 = 104976 + 186624 = 291600$ .
- $R = \sqrt{291600} = \sqrt{(540)^2} = 540$ .

[Model then performs a third full-scale re-expansion of the entire complex expression to confirm the coefficients 324 and -432 are correct.] ...

summing up the real components obtained above:  $(300\cos\theta - 468\sin\theta) + (24\cos\theta + 36\sin\theta) = 324\cos\theta - 432\sin\theta$ .

The result is consistent throughout. Thus, the largest possible real part of  $(75 + 117i)z + \frac{96 + 144i}{z}$  with  $|z|=4$  must indeed then equal 540.

Answer: 540

图 16: 阶段 4: 系统化深度推理 (FIPO 后期阶段)。在训练的后期阶段, 模型收敛到一种 “计算密集型” 策略。它超越了简单的反思, 转而进行严格的 多轮审计——包括符号重推导和细致的算术验证——以确保在复杂推理任务中表现更佳。