12.24 组会分享

Soft Adaptive Policy Optimization (SAPO)

Qwen 2025-12-02

Soft Adaptive Policy Optimization

Chang Gao* Chujie Zheng Xiong-Hui Chen Kai Dang Shixuan Liu
Bowen Yu* An Yang* Shuai Bai Jingren Zhou Junyang Lin
Qwen Team, Alibaba Inc.

😟 提出了一种**基于Sigmoid的软门控机制（Soft Gating）**替代传统 GRPO、GSPO 中的 Hard Clipping，并**引入非对称温度（对正负样本施加不同的梯度衰减速率）**，解决了 LLM RL 训练中的不稳定问题以及与学习效率失衡问题。

1. Problem

- LLM 强化学习中，Importance Ratios ($r_t(\theta)$) 在 Token 级别经常出现高方差，尤其在 MoE 模型中（不同的路由和长 Response 会进一步放大偏差）。对此**现有方法的硬裁剪机制（Hard Clipping）难以平衡稳定性与有效学习信号保留**：
  - GRPO 使用 `clip(r, 1-ε, 1+ε)` 进行截断，如果 Clip 太紧（ε小），浪费大量有效样本，**Sample Efficiency 低**；如果Clip太松（ε大），引入来自 Off-policy 样本的噪声梯度，导致**训练不稳定**。
  - GSPO 在 Sequence 级别进行截断，如果序列中仅有少数几个 Token 严重 Off-policy，GSPO 也会抑制整个序列的梯度，降低 **Sample Efficiency** 。

---

2. Design

1. 与其使用非黑即白的 Hard Clipping，不如设计一个**平滑的软门控机制**。当Ratio偏离1时，梯度权重平滑衰减，既保留了适度的探索信号，又抑制了极端的Off-policy更新

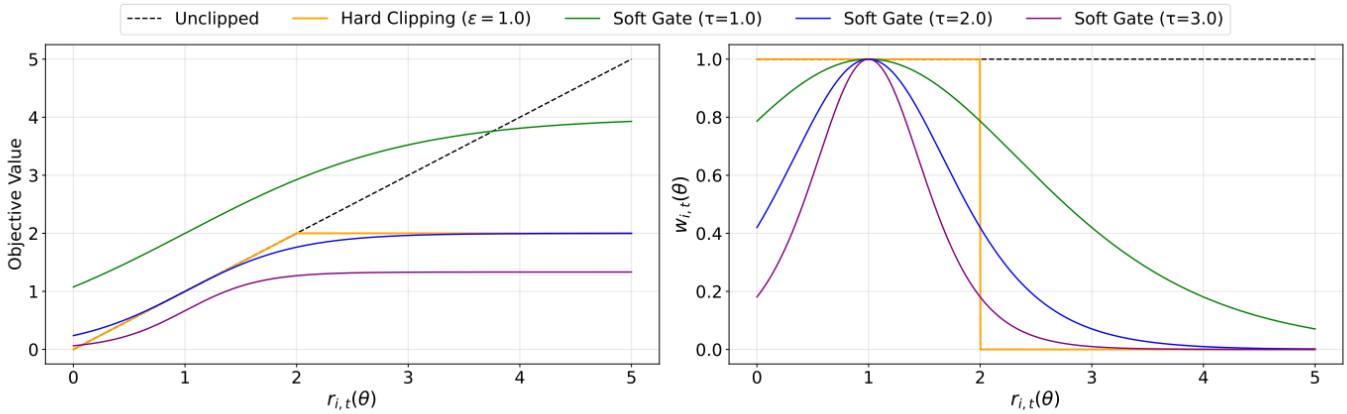


Figure 1: Comparison of policy-update objectives under positive advantage. The left panel shows the surrogate objective value; the right panel shows the corresponding gradient weight $w_{i,t}(\theta)$ as a function of the policy ratio $r_{i,t}(\theta)$.

2. 需要在保持**序列级连贯性**（像GSPO）的同时（paper中有理论证明 SAPO 在温和假设下，SAPO 的 token 级软门控可近似为序列级软门控），具备**Token级的适应性**（像GRPO但更Soft）。即：只降低那些坏掉的Token的权重，保留同一序列中好Token的权重
3. **正样本**提升当前Token的Logit；而**负样本**会压低当前Token，提升 Vocabulary 中其余所有未采样 Token 的 Logit。这虽然能增加熵（探索性），但也更容易**引入不稳定性**，所以对负样本施加更强的约束（也就是更高的温度系数 τ ）。这也在后续温度的消融实验中得到证明

$$\begin{aligned}\frac{\partial \log \pi_\theta(y_{i,t} \mid q, y_{i,<t})\, \widehat{A}_{i,t}}{\partial z_v} &= \frac{\partial \pi_\theta(y_{i,t} \mid q, y_{i,<t})}{\partial z_v} \cdot \frac{\widehat{A}_{i,t}}{\pi_\theta(y_{i,t} \mid q, y_{i,<t})} \\ &= \frac{\mathbb{1}(v = y_{i,t}) \exp(z_{y_{i,t}}) \sum_{v' \in \mathcal{V}} \exp(z_{v'}) - \exp(z_{y_{i,t}}) \exp(z_v)}{\left(\sum_{v' \in \mathcal{V}} \exp(z_{v'})\right)^2} \cdot \frac{\widehat{A}_{i,t}}{\pi_\theta(y_{i,t} \mid q, y_{i,<t})} \\ &= \begin{cases} \left(1 - \pi_\theta(y_{i,t} \mid q, y_{i,<t})\right) \cdot \widehat{A}_{i,t} & \text{if } v = y_{i,t} \quad \text{(sampled token)} \\ -\pi_\theta(v \mid q, y_{i,<t}) \cdot \widehat{A}_{i,t} & \text{otherwise} \quad \text{(unsampled token)} \end{cases}\end{aligned} \tag{9}$$

---

3. 具体 Method

3.1 SAPO 核心公式

- SAPO 将 Hard Clip 替换为加权的 Soft Gate:

$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} f_{i,t}(r_{i,t}(\theta)) \widehat{A}_{i,t} \right], \tag{5}$$

where

$$f_{i,t}(x) = \sigma\left(\tau_{i,t}\,(x-1)\right) \cdot \frac{4}{\tau_{i,t}}, \quad \tau_{i,t} = \begin{cases} \tau_{\text{pos}}, & \text{if } \widehat{A}_{i,t} > 0 \\ \tau_{\text{neg}}, & \text{otherwise} \end{cases}, \tag{6}$$

- 其中 σ 是 Sigmoid 函数。当重要性采样 r=1 时，梯度与无Clip时一致；当 r 偏离时，梯度按 Sigmoid导数形式衰减

$$f_{i,t}^{\mathrm{SAPO}'}(r_{i,t}(\theta)) = 4\,\sigma\big(\tau_i\,(r_{i,t}(\theta)-1)\big)\Big(1-\sigma\big(\tau_i\,(r_{i,t}(\theta)-1)\big)\Big) = \mathrm{sech}^2\Big(\frac{\tau_i}{2}\,(r_{i,t}(\theta)-1)\Big), \qquad (16)$$

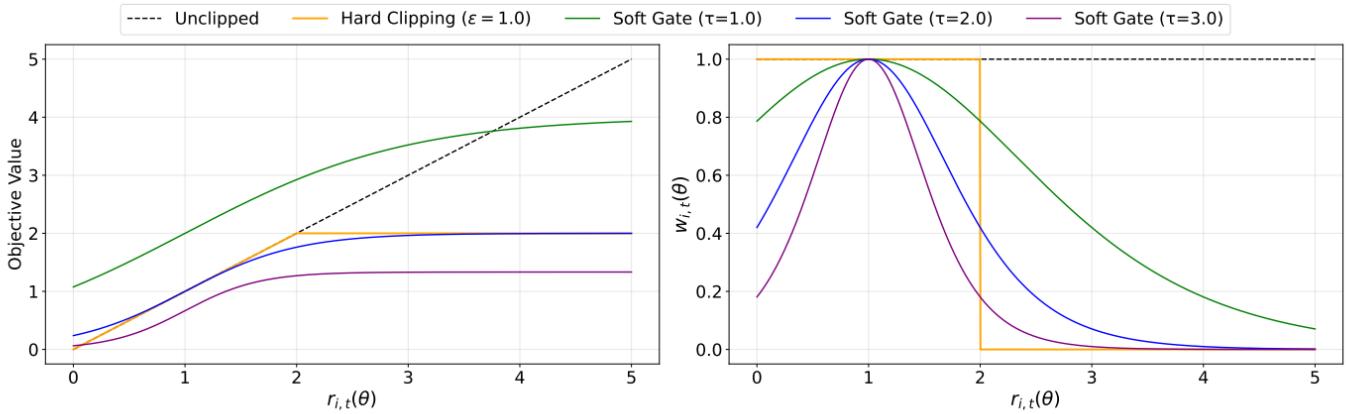


Figure 1: Comparison of policy-update objectives under positive advantage. The left panel shows the surrogate objective value; the right panel shows the corresponding gradient weight $w_{i,t}(\theta)$ as a function of the policy ratio $r_{i,t}(\theta)$.

3.2 非对称温度参数

为了解决负样本的不稳定性，SAPO引入了两个温度参数：

- **配置**: $\tau_{neg} > \tau_{pos}$
- **效果**: 较大的 τ_{neg} 使得负样本的梯度权重衰减得更快，从而更严格地限制 Off-policy 的负向更新

3.3 理论视角：与GSPO/GRPO的联系

- **与 GSPO 的联系**:
  - 在满足 (A1) Small steps/on policy 和 (A2) Low intra-sequence dispersion 的假设下，SAPO 的 Token 级平均门控近似于一个序列级的软门控。

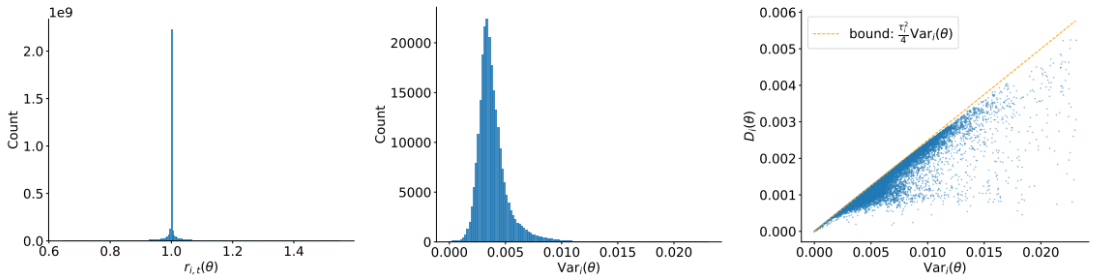


Figure 2: Empirical validation of assumptions (A1)–(A2) on the MoE model (Qwen3-30B-A3B). Left: histogram of token importance ratios $r_{i,t}(\theta)$. Middle: histogram of per-sequence log-ratio variance $\mathrm{Var}_i(\theta)$. Right: scatter of $\mathrm{Var}_i(\theta)$ versus $D_i(\theta)$.

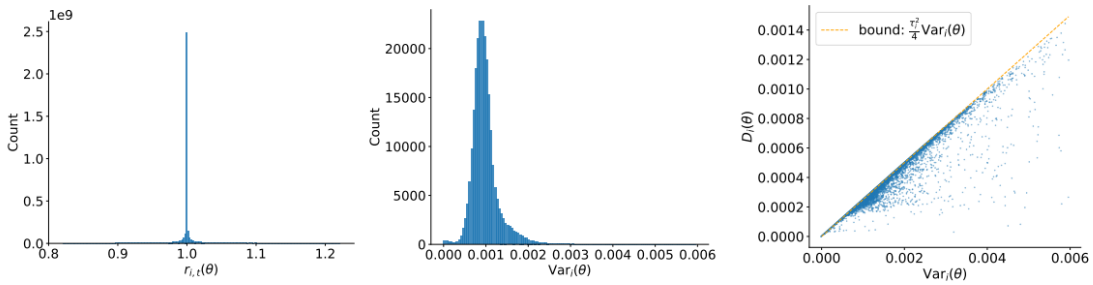


Figure 3: Empirical validation of assumptions (A1)–(A2) on the dense model (Qwen3-4B). Left: histogram of token importance ratios $r_{i,t}(\theta)$. Middle: histogram of per-sequence log-ratio variance $\mathrm{Var}_i(\theta)$. Right: scatter of $\mathrm{Var}_i(\theta)$ versus $D_i(\theta)$.

- 这证明了 SAPO 在稳定时具有序列连贯性，但在遇到特殊Token时能退回Token级处理。

SAPO's token-level soft gate Using $\sigma(x)(1-\sigma(x)) = \frac{1}{(e^{x/2}+e^{-x/2})^2} = \frac{1}{4}\mathrm{sech}^2(x/2)$, we have

$$f_{i,t}^{\mathrm{SAPO}'}(r_{i,t}(\theta)) = 4\,\sigma(\tau_i\,(r_{i,t}(\theta)-1))\Big(1-\sigma(\tau_i\,(r_{i,t}(\theta)-1))\Big) = \mathrm{sech}^2\Big(\frac{\tau_i}{2}\,(r_{i,t}(\theta)-1)\Big), \tag{16}$$

Assumptions We invoke two common assumptions:

(A1) Small-step/on-policy: $r_{i,t}(\theta) \approx 1$. Thus, $\log r_{i,t}(\theta) \approx r_{i,t}(\theta) - 1$.

(A2) Low intra-sequence dispersion: letting $z_{i,t}(\theta) := \log r_{i,t}(\theta)$ and $\mu_i(\theta) := \frac{1}{|y_i|}\sum_t z_{i,t}(\theta) = \log s_i(\theta)$, the variance $\mathrm{Var}_i(\theta) := \frac{1}{|y_i|}\sum_t (z_{i,t}(\theta)-\mu_i(\theta))^2$ is small for most sequences.

Under (A1), we have

$$f_{i,t}^{\mathrm{SAPO}'}(r_{i,t}(\theta)) = \mathrm{sech}^2\Big(\frac{\tau_i}{2}\,(r_{i,t}(\theta)-1)\Big) \;\approx\; \mathrm{sech}^2\Big(\frac{\tau_i}{2}\,\log r_{i,t}(\theta)\Big) =: g_{\tau_i}(z_{i,t}(\theta)). \tag{17}$$

Average token gates $\Rightarrow$ sequence gate By a second-order Taylor expansion of the smooth function $g_\tau(z) = \mathrm{sech}^2(\frac{\tau}{2}z)$ around $\mu_i(\theta) = \log s_i(\theta)$,

$$g_{\tau_i}(z_{i,t}(\theta)) = g_{\tau_i}(\mu_i(\theta)) + g'_{\tau_i}(\mu_i(\theta))\,(z_{i,t}(\theta)-\mu_i(\theta)) + \frac{1}{2}g''_{\tau_i}(\xi_{i,t}(\theta))\,(z_{i,t}(\theta)-\mu_i(\theta))^2, \tag{18}$$

for some $\xi_{i,t}(\theta)$ between $z_{i,t}(\theta)$ and $\mu_i(\theta)$. Averaging over tokens cancels the linear term:

$$\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g_{\tau_i}(z_{i,t}(\theta)) = g_{\tau_i}(\mu_i(\theta)) + \frac{1}{2}\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g''_{\tau_i}(\xi_{i,t}(\theta))\,(z_{i,t}(\theta)-\mu_i(\theta))^2\right). \tag{19}$$

For $g_\tau(z) = \mathrm{sech}^2(\alpha z)$ with $\alpha = \frac{\tau}{2}$, a direct calculation gives

$$g''_\tau(z) = \alpha^2\Big(4\,\mathrm{sech}^2(\alpha z) - 6\,\mathrm{sech}^4(\alpha z)\Big), \quad \sup_z |g''_\tau(z)| = 2\alpha^2 = \frac{\tau^2}{2}. \tag{20}$$

Hence, the average token gate is well-approximated by the sequence gate with a uniform bound:

$$D_i(\theta) = \left|\frac{1}{|y_i|}\sum_t g_{\tau_i}(z_{i,t}(\theta)) - g_{\tau_i}(\mu_i(\theta))\right| \le \frac{1}{2}\sup_z |g''_{\tau_i}(z)|\;\mathrm{Var}_i(\theta) = \frac{\tau_i^2}{4}\,\mathrm{Var}_i(\theta). \tag{21}$$

Starting from (15) and applying $r_{i,t}(\theta) \approx 1$ (A1), we have

$$\nabla_\theta \mathcal{J}_{\mathrm{SAPO}} \approx \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} g_{\tau_i}(z_{i,t}(\theta))\;\nabla_\theta \log \pi_\theta(y_{i,t}\mid q, y_{i,<t})\;\widehat{A}_i\right]. \tag{22}$$

Using (21), we have

$$\begin{aligned}\nabla_\theta \mathcal{J}_{\mathrm{SAPO}} &\approx \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} g_{\tau_i}(\log s_i(\theta))\;\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\nabla_\theta \log \pi_\theta(y_{i,t}\mid q, y_{i,<t})\right)\;\widehat{A}_i\right] \\ &= \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} g_{\tau_i}(\log s_i(\theta))\;\nabla_\theta \log s_i(\theta)\;\widehat{A}_i\right].\end{aligned} \tag{23}$$

Thus, under (A1)–(A2), SAPO reduces to a sequence-level update structurally similar to GSPO with a smooth gate $g_{\tau_i}(\log s_i(\theta)) = \mathrm{sech}^2(\frac{\tau_i}{2}\log s_i(\theta))$.

- **与 GRPO 的联系**:
  - GRPO 是二值化的 Trust Region（要么1要么0），SAPO 是连续的

4. Experiments

4.1 训练稳定性与性能

- **设置**: Qwen3-30B-A3B-Base 冷启动，数学推理任务 (AIME25, HMMT25 等)，对比 SAPO，GSPO，GRPO-R2（装了 Routing Replay 的GRPO）

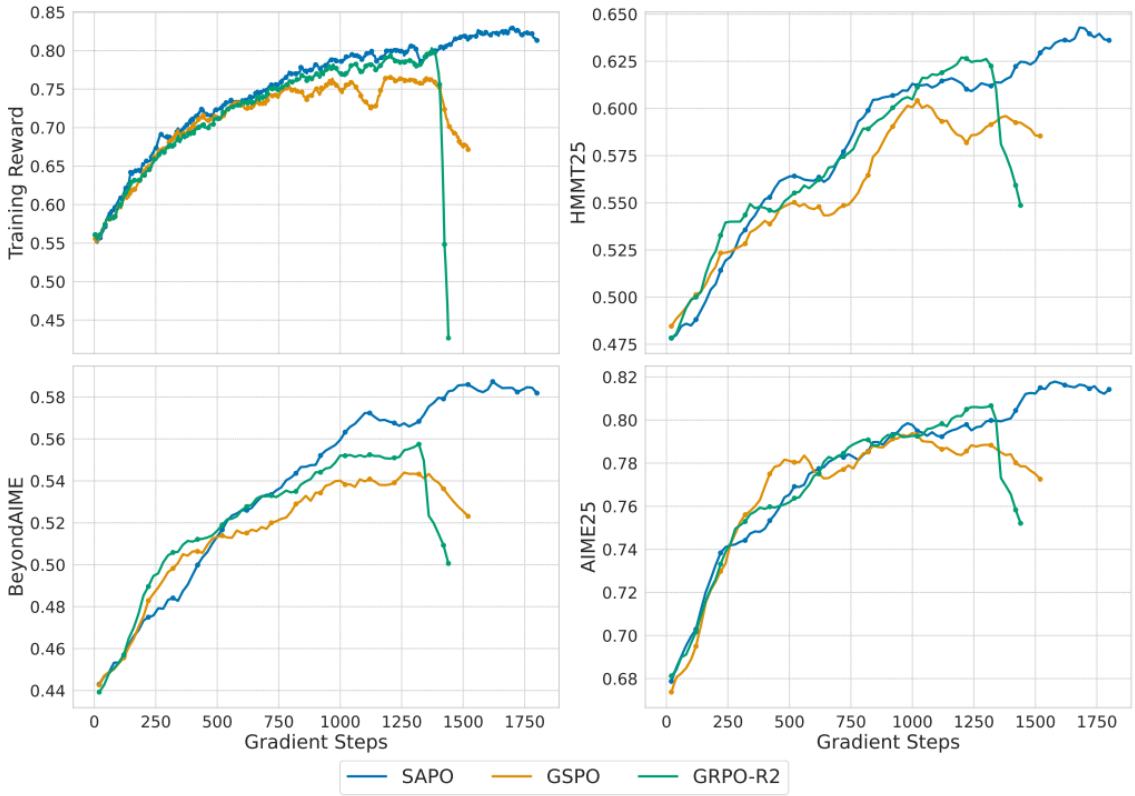
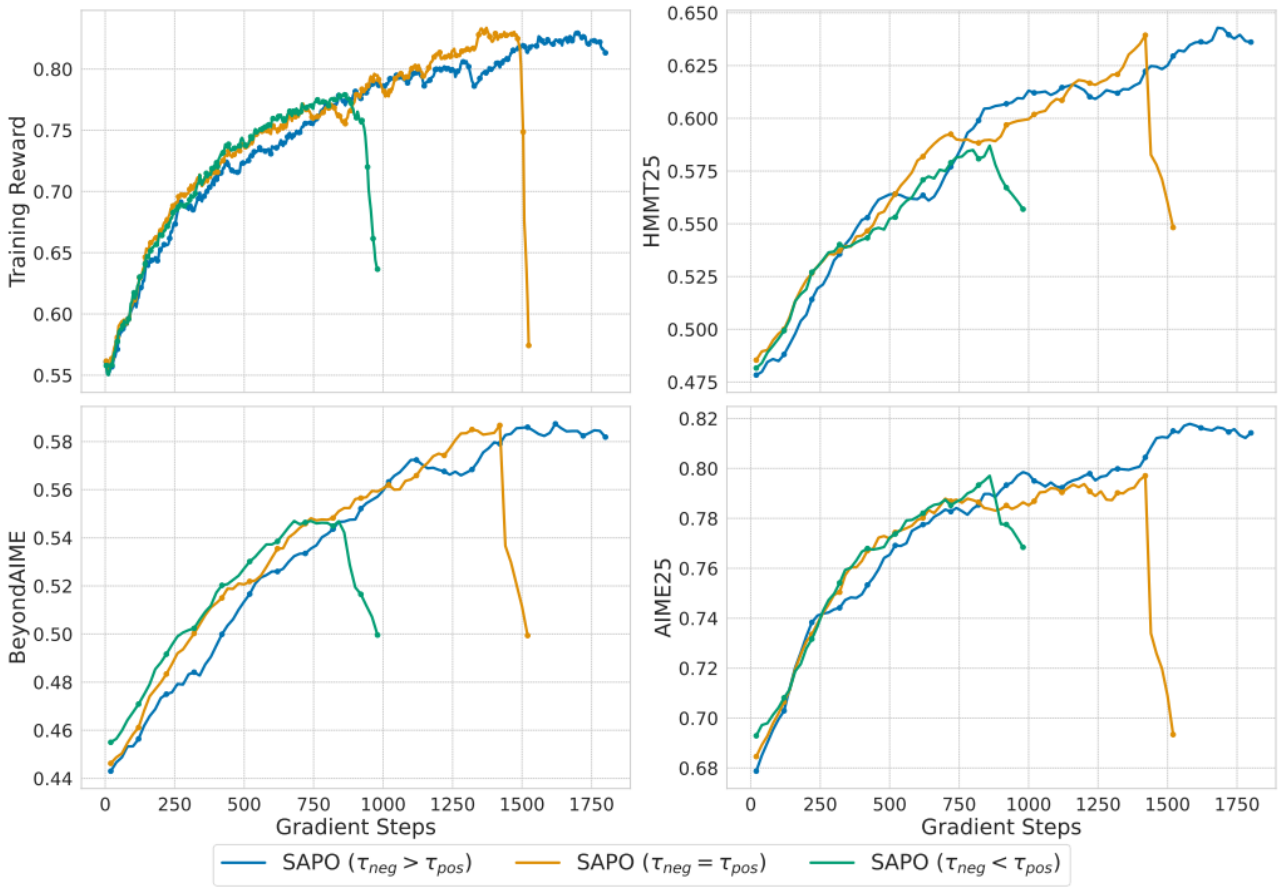


Figure 4: Training reward and validation performance of a cold-start model fine-tuned from Qwen3-30B-A3B-Base under different RL algorithms. SAPO exhibits consistently stable learning and achieves higher final performance compared with GSPO and GRPO-R2, both of which experience early training collapse.

4.2 温度参数消融实验

- **验证不对称性**: 对比了 $\tau_{neg} > \tau_{pos}$, $\tau_{neg} = \tau_{pos}$, $\tau_{neg} < \tau_{pos}$ 三种情况



4.3 扩展性

- **Qwen3-VL**: 在多模态模型和MoE架构上，SAPO 同样优于 GSPO 和 GRPO，证明了其在不同模态和模型规模下的通用性