arXiv:2601.00423v1 [cs.LG] 1 Jan 2026

E-GRPO：高熵步骤推动流模型的有效强化学习

Shengjun Zhang, Zhang Zhang,Chensheng Dai,Yueqi Duan†清华大学
{zhangsj23, z-z23}@mails.tsinghua.edu.cn, duanyueqi@tsinghua.edu.cn

摘要

近期，强化学习提升了流匹配模型在人类偏好对齐方面的表现。虽然随机采样能够探索去噪方向，但现有的跨多个去噪步骤进行优化的方法会受到稀疏且模糊的奖励信号的影响。我们发现，高熵步骤能够实现更高效、更有效的探索，而低熵步骤则会产生难以区分的滚动采样结果。为此，我们提出 *E-GRPO*，这是一种熵感知的组相对策略优化方法，旨在提高 *SDE* 采样步骤的熵。由于随机微分方程的积分过程会受到多步随机性所导致的模糊奖励信号影响，我们专门将连续的低熵步骤合并为一个高熵步骤，用于 *SDE* 采样，同时在其他步骤采用 *ODE* 采样。在此基础上，我们提出多步群组归一化优势，在共享同一合并后 *SDE* 去噪步骤的样本中计算群组相对优势。不同奖励设置下的实验结果证明了我们方法的有效性。代码可在以下地址获取：*https://github.com/shengjun-zhang/VisualGRPO*。

1. 引言

生成模型的最新进展极大地推动了视觉内容创作领域的发展，催生了从艺术设计和娱乐到医学成像与虚拟现实等广泛应用。最先进的扩散模型 [13, 29, 33] 和基于流的方法 [19, 21] 在生成高质量图像和视频 [5, 8, 26] 方面取得了显著的逼真度。

在大语言模型中，强化学习已展现出其在对齐人类偏好方面的有效性，包括近端策略优化（PPO） [30], 直接偏好优化（DPO） [28], 和组相对策略优化（GRPO） [32]。因此，基于人类反馈的强化学习（RLHF） [4, 9] 已被用于视觉生成的后训练阶段。由于 GRPO 通过移除价值网络简化了架构，并利用组内相对奖励直接计算优势，近期研究 [20, 38] 将其整合到采用随机微分方程（SDE）的流模型中。为提高采样效率，一些方法 [12, 17, 22, 40]引入了随机微分方程采样与常微分方程采样的混合方式，而另一些方法 [10, 18]则提出了基于树的结构，以减少采样步数。

尽管取得了这些进展，现有的基于 GRPO 的方法会在多个去噪时间步上进行策略优化，导致奖励信号稀疏且含糊，妨碍了有效对齐。我们观察到，只有高熵时间步对训练动态有实质性贡献。如图 1(b) 所示，在噪声水平较高的时间步，通过随机微分方程进行随机探索会产生更大的熵。我们展示了不同情况下生成的图像：高熵时间步会产生多样且奖励差异明显的图像，而低熵时间步产生的奖励差异较小，这与在最终图像中添加 10% 随机噪声所造成的效果相似。这一现象表明，在低熵区域，奖励模型难以辨别细微的轨迹偏差。此外，我们在四种设置下使用随机微分方程采样实现 GRPO：(i) 前 4 个时间步，(ii) 前 8 个时间步，(iii) 后 8 个时间步，以及 (iv) 全部 16 步。值得注意的是，在前半部分时间步上进行优化，效果甚至优于在所有时间步上优化，这表明后半部分时间步基本不提供有效信息。

为解决这一局限，我们提出 E-GRPO，这是一种熵感知 SDE 采样策略，旨在提升 GRPO 训练期间的探索效果。一种直观的做法是采用多步连续 SDE 采样来拓宽探索范围。然而，这会引入累积随机性，导致各步骤间的奖励归因含糊不清，使得某一步中有益的探索可能因后续轨迹偏离最优而受到惩罚，从而导致优化方向相反。为此，我们将多个低熵 SDE 步骤整合为一个有效的 SDE 步，同时将其余步骤

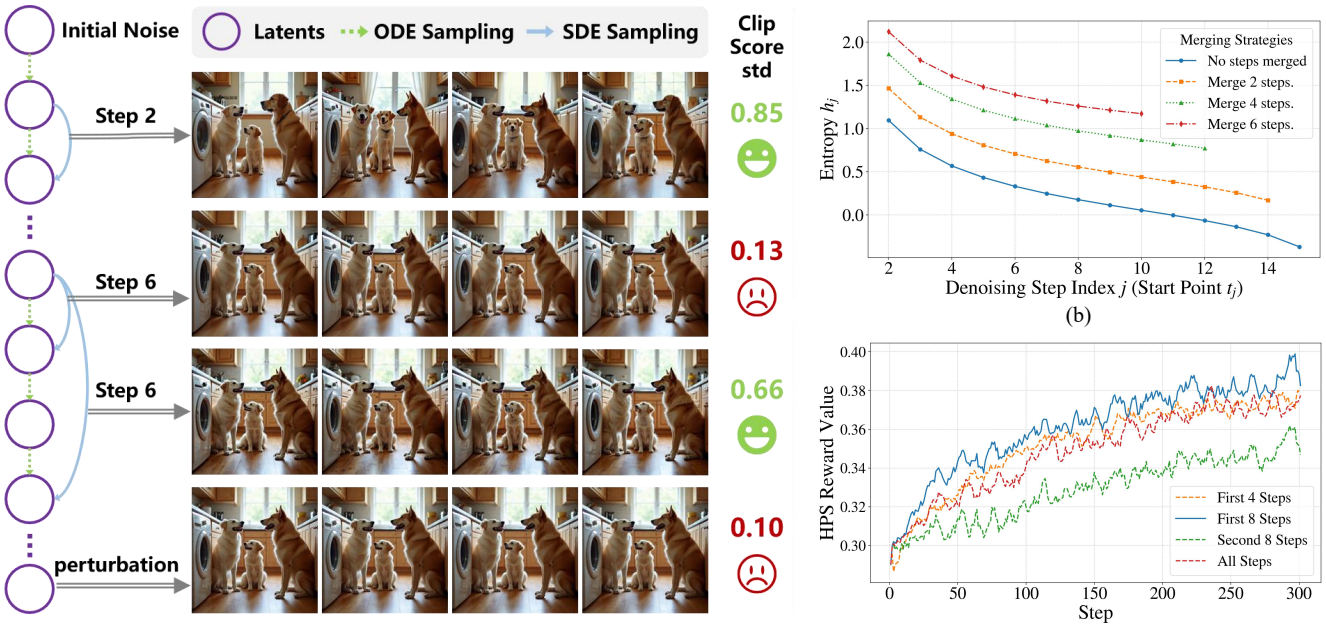


图1. **熵对采样结果的影响**。(a) 我们展示了采用不同 SDE采样策略生成的图像，包括在第 2 步和第 6 步进行单步SDE采样，以及在第 6 步进行合并步SDE采样。我们还报告了生成图像的 CLIP 分数方差。来自初始步骤和合并步骤的样本差异较大，而后验步骤生成的样本难以区分，其方差与对原始图像施加小扰动后的方差相近。(b) 我们报告了采用不同合并步数时，SDE采样在每个时间步的熵。合并步数越多，熵越高，强化学习训练中的探索范围也越大。(c) 我们展示了分别在所有时间步、前半部分时间步和后半部分时间步上训练的模型的训练奖励曲线。

保持为确定性的 ODE 采样，从而仅在有信息价值的地方保留高熵探索，并确保奖励归因可靠。在此基础上，我们引入多步群组归一化优势，在共享同一整合 SDE 步的样本中计算群组相对优势。该机制能够提供密集且可信的奖励信号，增强生成轨迹与人类偏好的对齐。

我们在单奖励设置和多奖励设置下进行了实验，并在域内和域外矩阵上进行了评估。实验结果证明了我们方法的有效性和效率。我们的主要贡献可总结如下：

1. 我们对 GRPO 训练过程中的去噪时间步进行了全面的基于熵的分析，发现只需在高熵步骤上进行优化，就能实现有效对齐。

2. 我们提出了 E-GRPO，这是一种用于流模型 GRPO 训练的熵感知 SDE 采样策略。该策略将多个低熵步骤合并为一个高熵随机微分方程步，从而拓展有效探索空间，同时消除奖励归因歧义。

3. 我们在单一奖励和多奖励设置下开展了广泛的实验，清晰地证明了 E-GRPO 始终优于以往的方法，验证了有针对性的熵引导随机优化的有效性和鲁棒性。

2. 相关工作

图像生成的强化学习对齐。 基于人类反馈的强化学习（RLHF） [3, 24]和基于可验证奖励的强化学习（RLVR） [16]已成为将大语言模型（LLM）与人类偏好对齐的强大范式 [1, 3, 31]。早期基于近端策略优化（PPO）的框架 [30]依靠价值模型来指导策略更新，而近期的方法（如组相对策略优化（GRPO）） [7, 23, 32] 则通过利用组内相对比较而非绝对奖励，实现了更高的稳定性和效率。语言对齐领域的这些进展激发了人们日益浓厚的兴趣，促使研究者将强化学习技术迁移到视觉生成模型，以使其与人类偏好对齐。在视觉生成领域，扩散 [13, 33]和流匹配模型 [19, 21, 25] 通过迭代去噪过程展现出强大的生成能力 [26, 29]。为增强与人类反馈的对齐，近期研究已将RLHF应用于这些模型。Diffusion-DPO [34], 和 D3PO [39] 扩展了直接偏好优

化（DPO） [28] 在扩散模型中的应用。然而，这些方法在训练过程中不会生成新样本，因此会受到分布偏移的影响。DanceGRPO [38] 和Flow-GRPO [20] 将基于确定性常微分方程（ODE）的采样重新表述为随机微分方程（SDE）轨迹，从而能够在视觉领域进行GRPO式策略更新。在此基础上，Granular-GRPO [40]细化了时间步粒度，以便在去噪步骤中进行更精确、密集的信用分配；TempFlow-GRPO [12]则引入时间感知加权，以缓解跨时间步均匀优化的局限。MixGRPO [17]通过混合ODE–SDE采样机制进一步提升训练效率，而BranchGRPO [18]则借助分支轨迹和结构化剪枝提高探索效率。尽管取得了这些进展，现有的流模型GRPO框架通常仍在所有时间步上进行均匀优化，忽视了去噪过程中探索潜力的差异，并且面临奖励信号稀疏或噪声较大的问题。我们的工作以逐步熵作为探索能力的衡量指标来应对这些挑战，通过针对高熵步骤进行优化，提升稳定性和效率。

熵引导的探索与对齐。 强化学习（RL）的早期研究已经认识到，熵是促进有效探索的重要机制。具体而言，策略熵正则化等策略已被广泛用于稳定学习并鼓励多样化行为 [2]。例如，软演员-评论家（SAC） [11] 在最大化期望奖励的同时，还会显式最大化策略熵，从而实现更加稳健且样本高效的探索。近年来，基于熵的洞见被应用于大语言模型（LLM）中的推理强化学习。研究表明，少量高熵词元会不成比例地推动策略改进，凸显了词元级不确定性在引导探索方面的重要性 [35]。相关研究进一步将熵形式化为理解探索动态的视角，证明高熵区域对应于对学习最具信息量的关键决策点 [6]。受这些发现启发，我们探究流匹配模型中是否存在类似的熵驱动模式，并提出一种具备熵感知能力的 GRPO 框架，优先处理信息量更高的去噪步骤，从而更高效、稳定地实现与人类偏好的对齐。

3. 方法

3.1. 预备知识

为了在强化学习中实现探索，基于流的群体相对策略优化（GRPO）将确定性的常微分方程采样转变为——确定性的常微分方程采样：

$$\mathrm{d}\mathbf{x}_t = \mathbf{v}_\theta(\mathbf{x}_t, t)\,\mathrm{d}t \tag{1}$$

转化为等价的随机微分方程：

$$\begin{aligned}\mathbf{x}_{t+\Delta t} =& \mathbf{x}_t + \left[\mathbf{v}_\theta(\mathbf{x}_t, t) + \frac{\sigma_t^2}{2t}\big(\mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t)\big)\right]\Delta t \\ &+ \sigma_t\sqrt{\Delta t}\,\boldsymbol{\epsilon},\end{aligned} \tag{2}$$

其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$ 和 $\sigma_t = a\sqrt{\frac{t}{1-t}}$通过随机微分方程采样，基于流的 GRPO 将在线强化学习融入流匹配模型，方法是将反向采样视为一个马尔可夫决策过程（MDP），其状态为$\mathbf{s}_t = (\mathbf{x}_t, t)$，动作为 $\mathbf{a}_t = \mathbf{x}_{t-1} \sim \pi_\theta(\cdot|\mathbf{s}_t)$，并针对提示词 c 设置终端奖励 $R(\mathbf{x}_0, c)$。策略优化

$$J_{\text{Flow-GRPO}}(\theta) = \mathbb{E}_{c\sim\mathcal{C},\ \{\mathbf{x}^{(i)}\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|c)}\left[f(r, A, \theta, \epsilon)\right].$$

裁剪代理目标 $f(r, A, \theta, \epsilon)$ 定义如下：

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\left[\min\big(r_t^{(i)}A^{(i)}, \text{clip}(r_t^{(i)}, 1-\epsilon, 1+\epsilon)A^{(i)}\big)\right],$$

其中 $r_t^{(i)}(\theta) = \frac{p_\theta(\mathbf{x}_{t-1}^{(i)}|\mathbf{x}_t^{(i)}, c)}{p_{\theta_{\text{old}}}(\mathbf{x}_{t-1}^{(i)}|\mathbf{x}_t^{(i)}, c)}$，而 $p_\theta(\mathbf{x}_{t-1}^{(i)}\,|\,\mathbf{x}_t^{(i)}, c)$ 是输出 $\mathbf{x}^{(i)}$ 在时间步 $t-1$ 的策略函数。组归一化优势 $A^{(i)}$ 的形式为：

$$A^{(i)} = \frac{R(\mathbf{x}_0^{(i)}, c) - \text{mean}\{R(\mathbf{x}_0^{(j)}, c)\}_{j=1}^{G}}{\text{std}\{R(\mathbf{x}_0^{(j)}, c)\}_{j=1}^{G}}. \tag{3}$$

遵循先前方法的实践 [17, 38], ，目标函数中省略了 KL 正则化项。

3.2. 熵分析

在基于流的群体相对策略优化中，来自随机微分方程的反向采样过程被构建为马尔可夫决策过程（MDP）。为了推导反向随机微分方程步的熵，我们从给定的前向随机微分方程出发，并应用贝叶斯定理。前向随机微分方程如下：

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \mu_\theta(\mathbf{x}_t, t)\Delta t + \sigma_t\sqrt{\Delta t}\boldsymbol{\epsilon}, \tag{4}$$

其中，$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$ 注入随机性，漂移项为：

$$\mu_\theta(\mathbf{x}_t, t) = \mathbf{v}_\theta(\mathbf{x}_t, t) + \frac{\sigma_t^2}{2t}\big(\mathbf{x}_t + (1-t)\mathbf{v}_\theta(\mathbf{x}_t, t)\big).$$

前向随机微分方程的转移概率是一个高斯分布：

$$p_{\text{f}}(\mathbf{x}_{t+\Delta t}\,|\,\mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t+\Delta t}\,|\,\mathbf{x}_t + \mu_\theta(\mathbf{x}_t, t)\Delta t,\ \sigma_t^2\Delta t\, I\right).$$

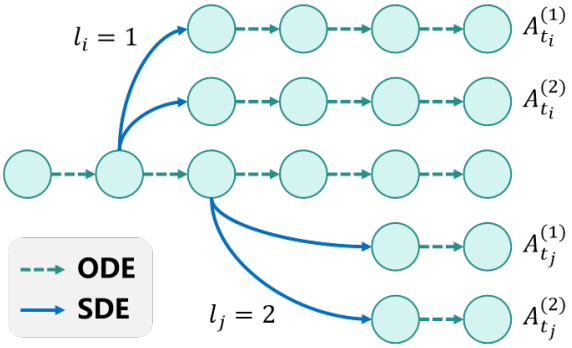


图 2. **E-GRPO 采样策略**。首先，我们生成一组与不同时间步对应的锚点噪声潜变量。对于每个活跃的随机微分方程时间步 t_i，根据熵分析选择合并步骤 $\mathcal{T}_i$。我们根据合并步骤的特定随机微分方程采样生成一组结果，并在每组中计算优势。

通过贝叶斯定理求解反向随机微分方程。 反向转移概率 $p_{\mathrm{r}}(\mathbf{x}_t \mid \mathbf{x}_{t+\Delta t})$对应于GRPO中的策略$\pi_\theta$，可通过贝叶斯定理推导得出：

$$p_{\mathrm{r}}(\mathbf{x}_t \mid \mathbf{x}_{t+\Delta t}) = \frac{p_{\mathrm{f}}(\mathbf{x}_{t+\Delta t} \mid \mathbf{x}_t)\, p(\mathbf{x}_t)}{p(\mathbf{x}_{t+\Delta t})}.$$

对于高斯过程，反向转移也服从高斯分布：

$$p_{\mathrm{r}}(\mathbf{x}_t \mid \mathbf{x}_{t+\Delta t}) = \mathcal{N}\left(\mathbf{x}_t \mid \tilde{\mu}_\theta(\mathbf{x}_{t+\Delta t}, t),\ \tilde{\sigma}_t^2 \Delta t\, I\right),$$

其中，$\tilde{\mu}_\theta$ 是反向漂移，$\tilde{\sigma}_t$ 是反向扩散系数。对于线性高斯随机微分方程，当过程具有时间可逆性时，两个方向上的扩散系数相同，其中 $\tilde{\sigma}_t = \sigma_t$。对于反向漂移 $\tilde{\mu}_\theta$，前向转移概率的对数为：

$$\begin{aligned}\log p_{\mathrm{f}} = &-\frac{1}{2}\log\det(2\pi\sigma_t^2\Delta t I) \\ &-\frac{1}{2\sigma_t^2\Delta t}\left\|\mathbf{x}_{t+\Delta t} - \mathbf{x}_t - \mu_\theta(\mathbf{x}_t, t)\Delta t\right\|^2\end{aligned}$$

对$\mathbf{x}_t$求导后，我们得到反向漂移的表达式为：

$$\tilde{\mu}_\theta(\mathbf{x}_{t+\Delta t}, t) = \mathbf{x}_{t+\Delta t} - \mu_\theta(\mathbf{x}_t, t)\Delta t + \sigma_t^2 \Delta t\, \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t).$$

反向随机微分方程步的熵。多元高斯分布 $\mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ 的熵为：

$$h(\mathbf{y}) = \frac{d}{2}\log\left((2\pi e)^d \det(\boldsymbol{\Sigma})\right) \tag{5}$$

其中，d 是随机变量$\mathbf{y}$的维度。对于反向随机微分方程步，协方差矩阵为：

$$\boldsymbol{\Sigma}_{\mathrm{r}} = \sigma_t^2 \Delta t\, I. \tag{6}$$

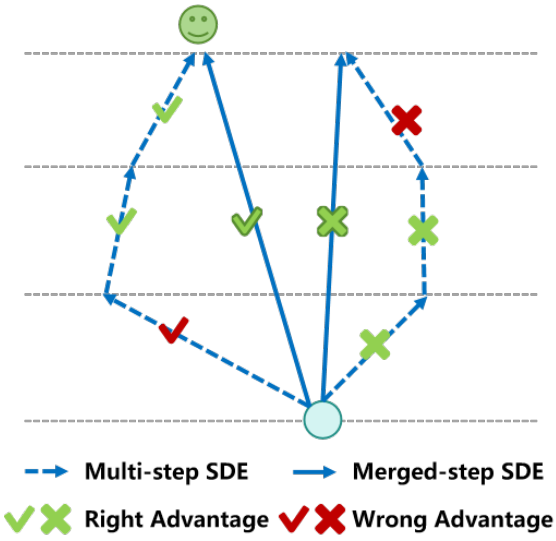


图3. **模糊奖励信号**。 对于连续的多步随机微分方程采样，优势对应于多个时间步，这可能导致特定时间步的优化方向错误。我们的合并步随机微分方程采样不仅扩大了探索范围，还通过将最终优势与一个合并后的随机微分方程步对齐，消除了模糊奖励。

这个对角矩阵的行列式为 $(\sigma_t^2\Delta t)^d$。将其代入熵公式：

$$\begin{aligned} h(t) &= \frac{1}{2}\log\left((2\pi e)^d(\sigma_t^2\Delta t)^d\right) &(7)\\ &= \frac{1}{2}\left[d\log(2\pi e) + d\log(\sigma_t^2\Delta t)\right] &(8)\\ &= \frac{d}{2}\log\left(2\pi e\sigma_t^2\Delta t\right) &(9)\end{aligned}$$

代入 $\sigma_t = a\sqrt{\frac{t}{1-t}}$，得：

$$h(t) = \frac{d}{2}\log\left(2\pi e \cdot a^2 \cdot \frac{t}{1-t}\cdot\Delta t\right) \tag{10}$$

3.3. 熵感知 GRPO

为了解决在各时间步进行均匀优化时奖励归因稀疏且模糊的问题，我们提出了熵感知 GRPO（E-GRPO）框架，该框架整合了熵驱动的步合并策略和多步组归一化优势估计。其核心设计是将低熵随机微分方程步骤整合为信息丰富的采样事件，以优先进行有意义的探索。

3.3.1. 熵驱动的步合并策略

给定一系列去噪时间步 $\{t_T, ..., t_1, t_0\}$，时间步与熵之间的关系可表示为：

$$e^{h(t_k)} \propto \frac{t_k}{1-t_k}(t_k - t_{k-1}). \tag{11}$$

实际上，流模型会调整时间偏移，以平衡质量与效率：

$$t_k = \frac{s\hat{t}_k}{1+(s-1)\hat{t}_k},$$

其中 $\hat{t}_k = \frac{k}{T}$。将其代入（11），得到：

$$e^{h(t_k)} \propto \frac{s^2 Tk}{(T-k)[T+(s-1)k][T+(s-1)(k-1)]}.$$

我们定义自适应熵阈值 τ，将时间步分为高熵时间步 $\{t_T, \cdots, t_{M+1}\}$，其 $e^{h(t_k)} \geq \tau$和低熵时间步 $\{t_M, \cdots, t_0\}$，其 $e^{h(t_k)} < \tau$。对于低熵时间步 t_m，我们可以在连续时间步上引入多步随机微分方程采样 $\{t_m, ..., t_{m-l}\}$。如图 3所示，这会引入累积随机性，导致各步骤间的奖励归因模糊。因此，一步中的有益探索可能会因下游轨迹偏差不理想而受到惩罚，从而导致优化方向相反。因此，我们将连续时间步合并为一个等效的随机微分方程步，以消除模糊奖励信号。合并 l 连续的低熵 SDE 步骤需要在减少步骤数量的同时保留总扩散效应。对于合并后的时间步，时间间隔为 $\Delta t = t_m - t_{m-l}$。将 Δt 代入（4），得到：

$$\mathbf{x}_{t_{m-l}} = \mathbf{x}_{t_m} + \mu_\theta(\mathbf{x}_{t_m}, t_m)(t_m - t_{m-l}) + \sigma_t \sqrt{t_m - t_{m-l}}\,\boldsymbol{\epsilon}.$$

根据（6），反向随机微分方程步的形式为：

$$\boldsymbol{\Sigma} = \sigma_t^2 (t_m - t_{m-l})\, I.$$

因此，合并后时间步的熵为：

$$\begin{aligned} e^{h(t_k)} &\propto \frac{t_m}{1-t_m}(t_m - t_{m-l}) \\ &\propto \frac{s^2 Tml}{(T-m)[T+(s-1)m][T+(s-1)(m-l)]}, \end{aligned}$$

其中 $e^{h(t_k)}$ 是关于 l的递增函数。

我们不再对所有低熵时间步使用统一的 l，而是提出一种自适应策略，为每个低熵时间步选择最优的 l，其中 l 的确定方式是使合并步的熵刚好超过阈值 τ。这一设计可避免单个合并步的熵过高，否则在探索尝试次数有限的情况下，将难以找到恰当的优化方向。对 l 进行自适应选择，可确保每个合并步都保持适中的熵水平——既足以保留有意义的探索信号，又能防止熵过高，以免无法有效指导优化。通过使合并后的熵与预定义阈值相匹配，我们在步合并带来的效率提升与奖励引导探索的可靠性之间取得了平衡。

算法1 熵感知 GRPO（E-GRPO）

输入： 初始策略 θ_{old}、提示集 $\mathcal{C}$、总时间步数 T，活跃的随机微分方程采样时间步 $\{t_T, \cdots, t_N\}$，合并步数 $\{l_T, \cdots, l_N\}$，裁剪系数 ϵ，轨迹数量 $\{G^{(T)}, \cdots, G^{(N)}\}$

输出： 优化后的策略 θ

对于迭代 = 1 至 K（总迭代次数）**执行**
  对于 $c \sim \mathcal{C}$（采样提示）**执行**
    对于 $N \leq n \leq T$ **执行**
      $\mathcal{T}_n \leftarrow \{t_n, t_{n-1}, \cdots, t_{n-l_n}\}$
      生成 $G^{(n)}$ 轨迹，使用 $\mathcal{T}_n$
      计算奖励 $\{R(\mathbf{x}_{0,t_n}^{(j)}, c)\}_{j=1}^{G^{(n)}}$
      $A_{t_n}^{(i)} \leftarrow \frac{R(\mathbf{x}_{0,t_n}^{(i)}, c) - \text{均值}\{R(\mathbf{x}_{0,t_n}^{(j)}, c)\}_{j=1}^{G^{(n)}}}{\text{标准差}\{R(\mathbf{x}_{0,t_n}^{(j)}, c)\}_{j=1}^{G^{(n)}}}$
      $r_{t_n}^{(i)}(\theta) \leftarrow \frac{p_\theta(\mathbf{x}_{t_{n-l_n}}^{(i)} | \mathbf{x}_{t_n}^{(i)}, c)}{p_{\theta_{\text{old}}}(\mathbf{x}_{t_{n-l_n}}^{(i)} | \mathbf{x}_{t_n}^{(i)}, c)}$
    结束循环
    构造裁剪代理目标：$f(r, A, \theta, \epsilon)$
    通过最小化 θ 更新 $J_{\text{E-GRPO}}(\theta)$
    设置 $\theta_{\text{old}} \leftarrow \theta$
  结束循环
结束循环返回优化后的策略 θ

3.3.2. 策略优化目标函数

为解决奖励归因歧义，我们通过定义合并分组样本，将 GRPO 的组归一化扩展到合并步骤。我们指定一组活跃的 SDE 时间步 $\{t_T, \cdots, t_N\}$，其中每个时间步 t_n(满足 $N \leq n \leq T$) 都关联一个合并步数 l_n，该步数由第3.3.1节中的熵驱动策略确定。对于给定的提示词 c，我们针对每个活跃时间步 t_n生成 G^n 条轨迹，所有 G^n 条轨迹都共享相同的合并后时间步 $\mathcal{T}_n \triangleq \{t_n, t_{n-1}, \cdots, t_{n-l_n}\}$。

在每个合并组 $\mathcal{T}_n$ 内，我们首先使用 G^n 轨迹计算优势估计，确保奖励信号一致地归因于合并后的时间步。第 i 条轨迹在状态 $\mathbf{x}^{t_n}$处、基于合并组 $\mathcal{T}_n$ 估算的优势为：

$$A_{t_n}^{(i)} = \frac{R(\mathbf{x}_{0,t_n}^{(i)}, c) - \text{mean}\{R(\mathbf{x}_{0,t_n}^{(j)}, c)\}_{j=1}^{G_n}}{\text{std}\{R(\mathbf{x}_{0,t_n}^{(j)}, c)\}_{j=1}^{G_n}}, \tag{12}$$

其中，$\mathbf{x}_{0,t_n}^{(j)}$ 表示采用活跃随机微分方程时间步 t_n 生成的第 j 个结果。针对合并分组样本调整后的最终裁剪代理目标函数$f(r, A, \theta, \epsilon)$，其表述如下：

$$\frac{1}{\hat{T}} \sum_{n=N}^{T} \frac{1}{G^{(n)}} \sum_{i=1}^{G^{(n)}} \min\left(r_{t_n}^{(i)} A_{t_n}^{(i)}, \text{clip}(r_{t_n}^{(i)}, 1-\epsilon, 1+\epsilon) A_{t_n}^{(i)}\right),$$

表1. **评估结果。** 不同方法之间的比较。每列中最优和次优的结果分别以**粗体** 和下划线标示。

方法	训练奖励模型：HPS				训练奖励模型：HPS&CLIP			
	HPS	CLIP	PickScore	ImageScore	HPS	CLIP	PickScore	ImageScore
FLUX.1-dev [15]	0.311	0.388	0.231	1.089	0.311	0.388	0.231	1.089
DanceGRPO [38]	0.353	**0.375**	0.228	1.233	0.331	0.389	0.227	1.128
MixGRPO [17]	0.378	0.358	0.225	1.266	0.363	0.399	0.230	1.436
GranularGRPO [40]	<u>0.385</u>	0.355	0.229	<u>1.313</u>	<u>0.377</u>	<u>0.400</u>	<u>0.236</u>	<u>1.490</u>
BranchGRPO [18]	0.358	<u>0.365</u>	<u>0.231</u>	1.311	0.342	0.384	0.230	1.243
TempFlowGRPO [12]	0.382	0.357	<u>0.231</u>	1.264	0.310	0.388	0.230	1.106
Ours	**0.391**	0.355	**0.232**	**1.324**	**0.382**	**0.401**	**0.237**	**1.494**

其中，$\hat{}T = T - N$ 和比率 $r_{t_n}^{(i)}$ 定义如下：

$$r_{t_n}^{(i)}(\theta) = \frac{p_\theta(\mathbf{x}_{t_{n-l_n}}^{(i)} \mid \mathbf{x}_{t_n}^{(i)}, c)}{p_{\theta_{\text{old}}}(\mathbf{x}_{t_{n-l_n}}^{(i)} \mid \mathbf{x}_{t_n}^{(i)}, c)}. \tag{13}$$

基于 GRPO 的裁剪代理目标，E-GRPO 将优化限制在合并后的高熵步骤上。目标函数修改为：

$$J_{\text{E-GRPO}}(\theta) = \mathbb{E}_{c\sim\mathcal{C},\{\mathbf{x}_{t_n}^{(i)}\}_{i=1}^{G^{(n)}}\sim\pi_{\theta_{\text{old}}}(\cdot|c),N\le n\le T} f(r, A, \theta, \epsilon).$$

我们的策略如算法 1所示。我们首先使用 ($h(t_k)$)计算所有时间步的10，并确定τ。然后，我们将连续的低熵步骤 $\mathcal{T}_i$ 按时间步 t_i进行聚类。对于其他步骤，我们使用常微分方程生成轨迹；对于合并步骤，则使用合并后的随机微分方程生成轨迹。最后，我们通过 ($A_{t_n}^{(i)}$ 12) 和 (13) 估计优势 p_θ 及比率，并最小化 $J_{\text{E-GRPO}}(\theta)$，以一致地将奖励信号归因于合并后的时间步，并更新。

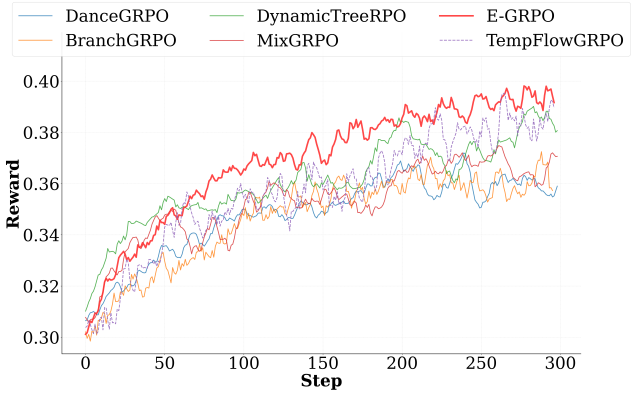


图4. **训练奖励曲线对比。** 与基线方法相比，E-GRPO 的奖励曲线表明，其在训练期间提升更快且更稳定。这说明，由高熵步骤引导的探索能够提升学习效率，同时减轻奖励信号中的噪声。

4. 实验

4.1. 实验设置

数据集与模型。 我们在 HPD 数据集 [36]，上开展实验，该数据集是用于人类偏好评估的大规模数据集，包含约 103,000 条训练文本提示和 400 条测试提示。实验中，我们采用 FLUX.1-dev [15] 作为骨干流匹配模型，与 DanceGRPO [38]和 MixGRPO [17]等先前工作保持一致。

评估设置。 为了评估与人类偏好的对齐程度，我们采用了几种具有代表性的奖励模型，它们分别捕捉生成图像的不同方面。 *HPS- v2。1* [36] 和*PickScore* [14] 均基于大规模人类偏好数据训练，因此能够反映人类对整体图像质量和文本-图像一致性的判断。 *CLIP* 分数 [27]主要衡量生成图像与输入提示之间的语义对齐。 *ImageRe-ward* [37] 则关注图像的感知质量和审美吸引力

，为基于偏好的指标提供了互补的视角。

采样策略。 遵循 DanceGRPO 的做法，训练过程中生成的图像从相同的初始化噪声中采样，以形成轨迹组。我们将采样步数总数设为 $T = 16$ ，并将方程 $\sigma_t = a\sqrt{\frac{t}{1-t}}$ 中的参数 a设为 0。7。训练期间，步骤合并的熵阈值 τ 设为 2。2。

训练细节。 我们首先仅使用 HPS-v2.1 奖励模型训练模型，以估计我们方法的性能上界。为增强鲁棒性并缓解潜在的奖励劫持问题，我们进一步使用 HPS-v2.1 和 CLIP 作为联合奖励信号训练模型。训练仅在前半部分采样步数上进行。

优化细节。 所有实验均在8 ×NVIDIA A800 GPU上进行，批量大小为 1。我们采用 AdamW优化器，学习率为 2×10^{-6}

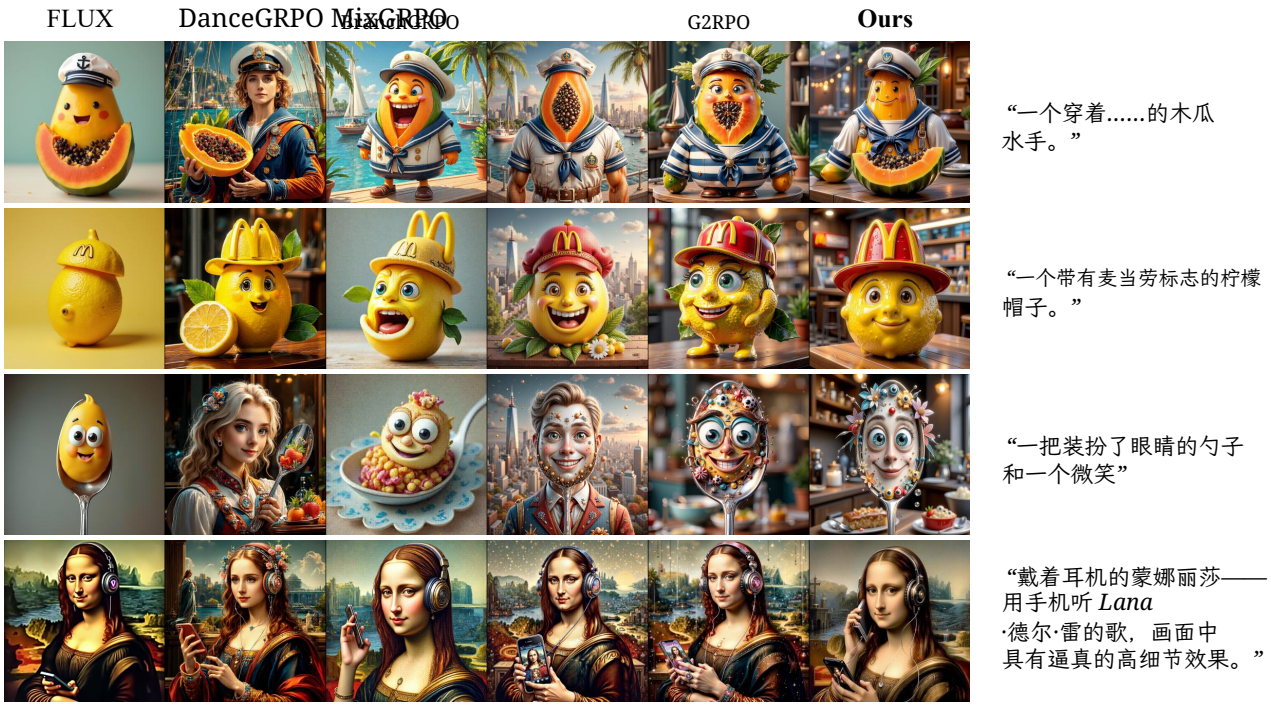


图 5. **可视化对比**。 E-GRPO 与其他基线方法的比较。E-GRPO 更好地融合了语义与细粒度细节。

权重衰减为 1×10^{-4}。使用 bfloat16 格式启用混合精度训练。训练迭代总数为 300。

4.2. 主要实验

如表1所示，我们将该方法与多种近期方法进行比较，包括基线模型 FLUX.1-dev [15],DanceGRPO [38], MixGRPO [17], BranchGRPO [18],TempFlowGRPO [12]和 GranularGRPO [40]。在仅使用 HPS-v2.1 奖励进行训练时，我们的方法取得了新的最先进性能，在 HPS 指标上领先 DanceGRPO 10.8%。这表明，我们的熵引导探索能够有效识别高价值去噪步骤，从而实现更精准、稳定的策略优化。

然而，正如 DanceGRPO [38], 中所讨论的，仅使用 HPS-v2.1 进行训练可能会导致奖励劫持，产生过度饱和、与真实人类偏好不符的视觉效果。为了解决这一问题，我们遵循先前工作 [17, 38, 40]，在训练期间采用联合奖励方案，同时使用 HPS-v2.1 和 CLIP 分数作为奖励。在这种更稳健的多重奖励设置下，我们的方法不仅保持了在域内 HPS 指标上的最先进性能，还在域外指标上取得了显著提升。具体而言，与 Dance-GRPO 相比，我们的方法使 ImageReward 提升了 32.4%，并使 PickScore 提升了 4.4%，这表明熵引导的优化能够促进在不同奖励模型上的更广泛泛化，并有效缓解奖励劫持。

图 5 展示了 FLUX.1-dev、DanceGRPO、BranchGRPO、MixGRPO、G2RPO 与我们提出的 E-GRPO 之间的定性比较。如第一行所示（提示词：“一个打扮成水手的木瓜”），E-GRPO 生成的构图自然地将木瓜的结构与类人的服饰相结合，呈现出美学质量更高、真实感更强的图像。相比之下，基线方法要么误解提示词（例如，DanceGRPO 生成了一个手拿木瓜的人），要么产生视觉上不连贯的结果（例如，MixGRPO 和 G2RPO）。在第三行（提示词：“一把长着眼睛、带着微笑的勺子”），E-GRPO 在保留勺子金属质感的同时，生成富有表现力且视觉上协调一致的拟人化面孔；而其他方法则会生成不真实的面部融合效果，或丧失材质保真度。这些结果表明，E-GRPO 在语义锚定和视觉连贯性方面表现更优，因此生成的图像能更忠实地体现文本意图，并更符合人类的审美偏好。

图 4 展示了训练期间的奖励轨迹。与先前工作相比，我们的方法在训练早期奖励增长更快，收敛也更加平稳，最终奖励更高。这表明，

表 2. **步合并策略比较**。 训练过程中不同步骤合并策略的定量结果。所提出的熵感知自适应合并策略在 HPS、CLIP、PickScore 和 ImageScore 上始终取得更高分数，表明其具有更好的语义对齐和生成质量。每列中的最佳结果以**粗体**标示

.

合并策略	HPS	CLIP	PickScore	ImageScore
2 步	0.382	0.290	0.232	1.223
4 步	0.374	0.302	0.230	1.216
六步	0.337	**0.372**	0.226	1.298
自适应	**0.391**	0.355	**0.232**	**1.324**

熵引导的步骤选择通过将更新集中在信息量最大的去噪步骤上，使优化更加稳定，同时提升效率和可靠性。

4.3. 消融研究

为了评估所提出方法的有效性，我们开展了一系列消融实验，以了解 E-GRPO 的设计。

步骤合并策略。我们评估了多种步骤合并策略，以验证训练过程中基于熵的自适应合并方案的有效性。如表2所示，我们的方法在几乎所有评估指标上都稳定优于朴素的两步、四步和六步合并基线，展现出效率和鲁棒性。与不考虑探索程度、固定地合并多个步骤的策略相比，我们的熵感知自适应合并会动态调整合并行为，使各步骤的探索程度保持相当，从而实现更准确、高效的优化。

步骤熵分析。 为验证基于熵的分析是否合理，以及所提出的熵感知 GRPO 方法是否有效，我们通过在不同的去噪步骤子集上训练模型进行实验。具体而言，我们分别使用（1）前 4 步、（2）前 8 步、（3）最后 8 步和（4）全部步骤训练模型。如图 1(c) 所示，在前 8 个高熵步骤上训练取得了最佳性能，其次是使用前 4 步。相比之下，在全部步骤上训练的结果与仅使用前 4 步的情况相近，但计算成本显著更高。当模型在最后 8 个（低熵）步骤上训练时，性能大幅下降。这些结果表明，专注于训练早期的高熵步骤便足以取得出色性能，而纳入过多后期低熵步骤则会引入不必要的噪声并降低效率。因此，我们采用前 8 个去噪步骤作为默认训练配置。表3 进一步给出了不同训练步骤子集的定量结果。

表 3. **不同训练去噪步骤的比较**。在高熵（早期）步骤上训练的模型取得了更高的对齐分数，同时计算成本更低，这证实了高熵步骤对有效优化的贡献最大。每列中的最高分以**粗体**标出。请注意，CLIP 分数出现了意外偏差，这是由于仅使用 HPS 奖励进行训练所致，从而造成了一定程度的奖励劫持，正如前文所述。

合并策略	HPS	CLIP	PickScore	ImageScore
前 4 个步骤	0.370	0.348	0.231	1.252
前 8 个步骤	**0.391**	0.355	**0.232**	**1.324**
后 8 个步骤	0.357	**0.381**	0.231	1.250
全部步骤	0.366	0.359	0.231	1.169

5. 结论

本文针对现有基于群组相对策略优化（GRPO）的流模型方法中稀疏且模糊奖励信号这一关键挑战展开研究，该问题源于对所有去噪时间步进行统一优化。通过熵分析，我们揭示了一个关键洞见：高熵时间步能够有效促进探索和人类偏好对齐，而低熵时间步产生的轨迹缺乏区分度，从而阻碍奖励判别。为解决这一局限，我们提出 E-GRPO，这是一种熵感知框架，包含两项核心创新：自适应熵驱动步骤合并策略，以及多步组归一化优势估计。步骤合并策略将连续的低熵 SDE 步骤合并为单个高熵 SDE 步骤，同时对其他步骤保留 ODE 采样，从而消除累积随机性导致的奖励归因歧义。多步组归一化优势通过在共享同一合并步骤的样本内计算相对优势，确保奖励信号密集且可靠。在以 FLUX.1-dev 为骨干网络的 HPD 数据集上开展的大量实验验证了 E-GRPO 的有效性。

局限性与未来工作。 推进视觉生成模型发展的一个关键瓶颈，在于奖励信号的设计与对齐。奖励是引导强化学习范式生成高质量、人类偏好的内容的基石，但现有的奖励形式往往无法充分契合细腻的人类偏好，例如审美吸引力、语义一致性和语境适切性。这种不对齐不仅会导致生成结果不理想，还会使模型容易遭受奖励投机：模型可能利用奖励函数中的漏洞来最大化得分，却并未真正满足人类的期望。因此，开发更稳健、更有效的奖励模型仍是视觉强化学习驱动的生成领域未来研究的重要方向。

致谢。 本研究部分受到北京市自然科学基金项目（ L252011）和国家自然科学基金委员会项目（ 62576185）的资助。

参考文献

[1] Marwa Abdulhai、Isadora White、Charlie Snell、Charles Sun、Joey Hong、Yuexiang Zhai、Kelvin Xu 和 Sergey Levine。《LMRL gym：语言模型多轮强化学习基准》，2023年。2[2] Zafarali Ahmed、Nicolas Le Roux、Mohammad Norouzi 和 Dale Schuurmans。《理解熵对策略优化的影响》。载于第*36*届国际机器学习会议论文集，第151–160页。PMLR，2019年。3[3] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhao-hai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Junyang Lin。《Qwen2.5-VL 技术报告》，2025。2[4] Kevin Black、Michael Janner、Yilun Du、Ilya Kostrikov 和 Sergey Levine。《使用强化学习训练扩散模型》。*arXiv* 预印本 *arXiv:2305.13301*，2023年。1[5] Andreas Blattmann、Tim Dockhorn、Sumith Kulal、Daniel Mendelevitch、Maciej Kilian、Dominik Lorenz、Yam Levi、Zion English、Vikram Voleti、Adam Letts 等。《Stable Video Diffusion：将潜空间视频扩散模型扩展至大型数据集》。*arXiv* 预印本 *arXiv:2311.15127*，2023年。1[6] Daixuan Cheng、Shaohan Huang、Xuekai Zhu、Bo Dai、Wayne Xin Zhao、Zhenliang Zhang 和 Furu Wei。《通过探索进行推理：从熵的视角看大语言模型的强化学习》，2025年。3[7] DeepSeek-AI、Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shi-rong Ma、Peiyi Wang、Xiao Bi、Xiaokang Zhang、Xingkai Yu、Yu Wu、Z. F. Wu、Zhibin Gou、Zhihong Shao、Zhuoshu Li、Ziyi Gao、Aixin Liu、Bing Xue、Bingxuan Wang、Bochao Wu、Bei Feng、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan、Damai Dai、Deli Chen、Dongjie Ji、Erhang Li、Fangyun Lin、Fucong Dai、Fuli Luo、Guangbo Hao、Guanting Chen、Guowei Li、H. Zhang、Han Bao、Hanwei Xu、Haocheng Wang、Honghui Ding、Huajian Xin、Huazuo Gao、Hui Qu、Hui Li、Jianzhong Guo、Jiashi Li、Jiawei Wang、Jingchang Chen、Jingyang Yuan、Junjie Qiu、Junlong Li、J. L. Cai、Jiaqi Ni、Jian Liang、Jin Chen、Kai Dong、Kai Hu、Kaige Gao、Kang Guan、Kexin Huang、Kuai Yu、Lean Wang、Lecong Zhang、Liang Zhao、Litong Wang、Liyue Zhang、Lei Xu、Leyi Xia、Mingchuan Zhang、Minghua Zhang、Minghui Tang、Meng Li、Miaojun Wang、Ming-ming Li、Ning Tian、Panpan Huang、Peng Zhang、Qiancheng Wang、Qinyu Chen、Qiushi Du、Ruiqi Ge、Ruisong Zhang、Ruizhe Pan、Runji Wang、R. J. Chen、R. L. Jin、Ruyi Chen、Shanghao Lu、Shangyan Zhou、Shanhuang Chen、Shengfeng Ye、Shiyu Wang、Shuiping Yu、Shunfeng Zhou、Shuting Pan、S. S. Li、Shuang Zhou、Shaoqing Wu、Shengfeng Ye、Tao Yun、Tian Pei、Tianyu Sun、T. Wang、Wangding Zeng、Wan- jia Zhao、Wen Liu、Wenfeng Liang、Wenjun Gao、Wenqin Yu、Wentao Zhang、W. L. Xiao、Wei An、Xiaodong Liu、Xiaohan Wang、Xiaokang Chen、Xiaotao Nie、Xin Cheng、Xin Liu、Xin Xie、Xingchao Liu、Xinyu Yang、Xinyuan Li、Xuecheng Su、Xuheng Lin、X. Q. Li、Xiangyue Jin、Xiaojin Shen、Xi- aosha Chen、Xiaowen Sun、Xiaoxiang Wang、Xinnan Song、Xinyi Zhou、Xianzu Wang、Xinxia Shan、Y. K. Li、Y. Q. Wang、Y. X. Wei、Yang Zhang、Yanhong Xu、Yao Li、Yao Zhao、Yaofeng Sun、Yaohui Wang、Yi Yu、Yichao Zhang、Yi- fan Shi、Yiliang Xiong、Ying He、Yishi Piao、Yisong Wang、Yixuan Tan、Yiyang Ma、Yiyuan Liu、Yongqiang Guo、Yuan Ou、Yuduan Wang、Yue Gong、Yuheng Zou、Yujia He、Yun- fa n Xiong、Yuxiang Luo、Yuxiang You、Yuxuan Liu、Yuyang Zhou、Y. X. Zhu、Yanhong Xu、Yanping Huang、Yaohui Li、Yi Zheng、Yuchen Zhu、Yunxian Ma、Ying Tang、Yukun Zha、Yuting Yan、Z. Z. Ren、Zehui Ren、Zhangli Sha、Zhe Fu、Zhean Xu、Zhenda Xie、Zhengyan Zhang、Zhewen Hao、Zhicheng Ma、Zhigang Yan、Zhiyu Wu、Zihui Gu、Zijia Zhu、Zijun Liu、Zilin Li、Ziwei Xie、Ziyang Song、Zizheng Pan、Zhen Huang、Zhipeng Xu、Zhongyu Zhang、and Zhen Zhang。Deepseek-r1：通过强化学习激励大语言模型的推理能力，2025。2[8] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel、et al。扩展用于高分辨率图像合成的整流流 Transformer。在第四十一届国际机器学习会议，2024。1[9] Ying Fan、Olivia Watkins、Yuqing Du、Hao Liu、Moonkyung Ryu、Craig Boutilier、Pieter Abbeel、Moham- mad Ghavamzadeh、Kangwook Lee、and Kimin Lee。DPOK：用于微调文本到图像扩散模型的强化学习。神经信息处理系统进展，36：79858–79885，2023。1[10]Xiaolong Fu、Lichen Ma、Zipeng Guo、Gaojing Zhou、Chongxiao Wang、ShiPing Dong、Shizhe Zhou、Ximan Liu、Jingling Fu、Tan Lit Sin、et al。Dynamic-treeRPO：通过结构化采样打破独立轨迹瓶颈。arXiv预印本 arXiv:2509.23352，2025。1[11] Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel、and Sergey Levine。软演员-评论家：由随机演员实现的离策略最大熵深度强化学习，2018。3[12] Xiaoxuan He、Siming Fu、Yuke Zhao、Wanli Li、Jian Yang、Dacheng Yin、Fengyun Rao、and Bo Zhang。Te mpFlow-GRPO：在流模型中，GRPO的时机何时重要。arXiv预印本 arXiv:2508.04324，2025。1, 3, 6, 7[13] Jonathan Ho、Ajay Jain、and Pieter Abbeel。去噪扩散概率模型。神经信息处理系统进展，33：6840–6851，2020。1, 2[14] Yuval Kirstain、Adam Polyak、Uriel Singer、Shahbuland Ma- tiana、Joe Penna、and Omer Levy。Pick-a-Pic：一个用于文本到图像生成的用户偏好开放数据集。在神经信息处理系统进展，页码36652–36663。Curran Associates、Inc.，2023。6[15] Black Forest Labs。FLUX。https://github.com/ black-forest-labs/flux，2024。6, 7

[16]Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester J ames Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi。Tulu 3：推动开放语言模型后训练的前沿。在第二届语言建模会议，2025。2
[17] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong。MixGRPO：通过混合常微分方程-随机微分方程释放基于流的GRPO效率。arXiv预印本 arXiv:2507.21802，2025。1, 3, 6, 7[18] Yuming Li, Yikai Wang, Yuying Zhu, Zhongyu Zhao, Ming Lu, Qi She, and Shanghang Zhang。BranchGRPO：通过扩散模型中的结构化分支实现稳定且高效的GRPO。arXiv预印本 arXiv:2509.06040，2025。1, 3, 6, 7[19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximil- ian Nickel, and Matt Le。生成建模的流匹配。arXiv预印本 arXiv:2210.02747，2022。1, 2[20] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang。Flow-GRPO：通过在线强化学习训练流匹配模型。arXiv预印本 arXiv:2505.05470，2025。1, 3[21] Xingchao Liu, Chengyue Gong, and Qiang Liu。流模型又直又快：利用整流流学习生成和迁移数据。arXiv预印本 arXiv:2209.03003，2022。1, 2[22] Yifu Luo, Penghui Du, Bo Li, Sinan Du, Tiantian Zhang, Yongzhe Chang, Kai Wu, Kun Gai, and Xueqian Wang。逐步采样，按块优化：用于文本到图像生成的块级GRPO。arXiv预印本 arXiv:2510 21583，2025 1 . .[23] OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Hel- yar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexan- der Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, An- drea Vallone, Andrew Duberstein, Andrew Kondrich, An- drey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Ben- jamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Mi- naiev, Botao Hao, Bowen Baker, Brandon Houghton, Bran- don McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Che- ung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman、Hyung Won Chung、Ian Kivlichan、Ian O'Connell、Ian Osband、Ignasi Clavera Gilaberte、Ilge Akkaya、Ilya Kostrikov、Ilya Sutskever、Irina Kofman、Jakub Pachocki、James Lennon、Jason Wei、Jean Harb、Jerry Twore、Jiacheng Feng、Jiahui Yu、Jiayi Weng、Jie Tang、Jieqi Yu、Joaquin Qui˜nonero Candela、Joe Palermo、Joel Parish、Johannes Heidecke、John Hallman、John Rizzo、Jonathan Gordon、Jonathan Uesato、Jonathan Ward、Joost Huizinga、Julie Wang、Kai Chen、Kai Xiao、Karan Singhal、Karina Nguyen、Karl Cobbe、Katy Shi、Kayla Wood、Kendra Rimbach、Keren Gu-Lemberg、Kevin Liu、Kevin Lu、Kevin Stone、Kevin Yu、Lama Ahmad、Lauren Yang、Leo Liu、Leon Maksin、Leyton Ho、Liam Fedus、Lilian Weng、Linden Li、Lind- say McCallum、Lindsey Held、Lorenz Kuhn、Lukas Kon- draciuk、Lukasz Kaiser、Luke Metz、Madelaine Boyd、Maja Trebacz、Manas Joglekar、Mark Chen、Marko Tintor、Mason Meyer、Matt Jones、Matt Kaufer、Max Schwarzer、Meghan Shah、Mehmet Yatbaz、Melody Y. Guan、Mengyuan Xu、Mengyuan Yan、Mia Glaese、Mianna Chen、Michael Lampe、Michael Malek、Michele Wang、Michelle Fradin、Mike Mc- Clay、Mikhail Pavlov、Miles Wang、Mingxuan Wang、Mira Murati、Mo Bavarian、Mostafa Rohaninejad、Nat McAleese、Neil Chowdhury、Neil Chowdhury、Nick Ryder、Nikolas Tezak、Noam Brown、Ofir Nachum、Oleg Boiko、Oleg Murk、Olivia Watkins、Patrick Chao、Paul Ashbourne、Pavel Iz- mailov、Peter Zhokhov、Rachel Dias、Rahul Arora、Randall Lin、Rapha Gontijo Lopes、Raz Gaon、Reah Miyara、Reimar Leike、Renny Hwang、Rhythm Garg、Robin Brown、Roshan James、Rui Shu、Ryan Cheu、Ryan Greene、Saachi Jain、Sam Altman、Sam Toizer、Sam Toyer、Samuel Miserendino、Sand- hini Agarwal、Santiago Hernandez、Sasha Baker、Scott McK- inney、Scottie Yan、Shengjia Zhao、Shengli Hu、Shibani San- turkar、Shraman Ray Chaudhuri、Shuyuan Zhang、Siyuan Fu、Spencer Papay、Steph Lin、Suchir Balaji、Suvansh Sanjeev、Szymon Sidor、Tal Broda、Aidan Clark、Tao Wang、Tay- lor Gordon、Ted Sanders、Tejal Patwardhan、Thibault Sotti- aux、Thomas Degry、Thomas Dimson、Tianhao Zheng、Timur Garipov、Tom Stasi、Trapit Bansal、Trevor Creech、Troy Pe- terson、Tyna Eloundou、Valerie Qi、Vineet Kosaraju、Vinnie Monaco、Vitchyr Pong、Vlad Fomenko、Weiyi Zheng、Wenda Zhou、Wes McCabe、Wojciech Zaremba、Yann Dubois、Ying- hai Lu、Yining Chen、Young Cha、Yu Bai、Yuchen He、Yuchen Zhang、Yunyun Wang、Zheng Shao，以及 Zhuohan Li。《OpenAI o1 系统卡》，2024 年。2
[24] Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Car- roll Wainwright、Pamela Mishkin、Chong Zhang、Sandhini Agarwal、Katarina Slama、Alex Ray 等。通过人类反馈训练遵循指令的语言模型。神经信息处理系统进展，35：27730–27744，2022年。2
[25] William Peebles 和 Saining Xie。《可扩展的扩散模型——采用 Transformer 的模型》。载于 *IEEE/CVF* 国际计算机视觉会议（*ICCV*），页码 4195–4205，2023年。2
[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann、Tim Dockhorn、Jonas M¨uller、Joe Penna 和 Robin Rombach。SDXL：改进潜在扩散模——用于高分辨率图像合成的模型。*arXiv* 预印本

arXiv:2307.01952，2023年。1、2[27]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。从自然语言监督中学习可迁移视觉模型。载于《第38届国际机器学习会议论文集》，第8748–8763页。PMLR，2021年。6[28] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christo- pher D Manning、Stefano Ermon 和 Chelsea Finn。直接偏好优化：你的语言模型其实是一个奖励模型。《神经信息处理系统进展》，36：53728–53741，2023年。1、3[29]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10684–10695页，2022年。1、2[30] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Rad- ford 和 Oleg Klimov。近端策略优化算法。arXiv预印本 arXiv:1707.06347，2017年。1、2[31]Lior Shani、Aviv Rosenberg、Asaf Cassel、Oran Lang、Daniele Calandriello、Avital Zipori、Hila Noga、Orgad Keller、Bilal Piot、Idan Szpektor、Avinatan Hassidim、Yossi Matias 和 R´emi Munos。基于人类偏好的多轮强化学习，2024年。2[32] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu 等。Deepseekmath：突破开放语言模型数学推理的极限。arXiv预印本 arXiv:2402.03300，2024年。1、2[33] Jiaming Song、Chenlin Meng 和 Stefano Ermon。去噪扩散隐式模型。arXiv预印本 arXiv:2010.02502，2020年。1、2[34] Bram Wallace、Meihua Dang、Rafael Rafailov、Linqi Zhou、Aaron Lou、Senthil Purushwalkam、Stefano Ermon、Caiming Xiong、Shafiq Joty 和 Nikhil Naik。使用直接偏好优化进行扩散模型对齐。载于《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），第8228–8238页，2024年。2[35] Shenzhi Wang、Le Yu、Chang Gao、Chujie Zheng、Shix- uan Liu、Rui Lu、Kai Dang、Xionghui Chen、Jianxin Yang、Zhenru Zhang、Yuqiong Liu、An Yang、Andrew Zhao、Yang Yue、Shiji Song、Bowen Yu、Gao Huang 和 Junyang Lin。超越80/20法则：高熵少数令牌推动大语言模型推理中的有效强化学习，2025年。3[36] Xiaoshi Wu、Yiming Hao、Keqiang Sun、Yixiong Chen、Feng Zhu、Rui Zhao 和 Hongsheng Li。人类偏好分数 v2：评估文本到图像合成人类偏好的可靠基准。arXiv预印本 arXiv:2306 09341，2023年 6 。 。[37] Jiazheng Xu、Xiao Liu、Yuchen Wu、Yuxuan Tong、Qinkai Li、Ming Ding、Jie Tang 和 Yuxiao Dong。Imagere- ward：学习和评估文本到图像生成人类偏好。载于《神经信息处理系统进展》，第15903–15935页。Curran Associates, Inc.，2023年。6

[38] 薛泽越、吴杰、高宇、孔方圆、朱灵婷、陈梦昭、刘志恒、刘伟、郭秋山、黄伟林等。DanceGRPO：释放 GRPO 在视觉生成中的潜力。*arXiv* 预印本 *arXiv:2505.07818*，2025。1，3，6，7[39] 杨凯、陶建、吕佳飞、葛春江、陈佳欣、沈伟涵、朱小龙和李秀。无需奖励模型，利用人类反馈微调扩散模型。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第8941–8951页，2024年。2[40] 周宇杰、凌鹏扬、卜佳子、王一斌、臧宇航、王佳琪、牛力和翟光涛。G2RPO：面向流模型精确奖励的粒度化 GRPO。*arXiv* 预印本 *arXiv:2510.01982*，2025年。1，3，6，7

E-GRPO：高熵步骤推动流模型的有效强化学习

补充材料

6. 熵阈值的消融研究 τ

在我们的主论文中，我们引入了一种自适应熵阈 τ 值，用于将时间步划分为高熵组和低熵组。在采样过程中，连续的低熵步骤会被合并，直至其熵达到阈值。该阈值是熵驱动的步骤合并策略中的一个关键超参数。为验证所提出方法的有效性，并评估其对 τ的敏感性，我们针对不同的阈值开展了一系列实验。具体而言，我们在 HPS 奖励配置下训练了 E-GRPO，其中 τ设为 0（即所有步骤一视同仁，不进行合并）、1.8、2.0、2.2（默认设置）和 2.6。结果汇总于表4。

如表4所示，阈值取值不同，模型表现也有明显差异。随着 τ 增大，可达到的 HPS 分数也随之提高，这表明熵在训练过程中作为引导信号是有效的。然而，当 τ 过大时，较长的步骤序列可能会被合并，有时甚至会将仍包含有用熵信息或梯度信息的步骤合并在一起。这会导致更新过于粗略，进而使性能略有下降。值得注意的是，我们默认选择的 $\tau = 2.2$在利用熵进行引导和避免过度合并之间取得了有效平衡，因此在实验中获得了最佳的整体性能。

7. 补充可视化

7.1. 更多质量结果

为了进一步证明我们提出的 E-GRPO 的优越性，我们提供了更多与基线方法在图 6 和图 7中的定性比较。如这些图所示，E-GRPO 始终能生成更忠实于文本提示的结果。例如，对于提示词“以 1930 年代为灵感、风格低调的太空时代美学中的一幅获奖柠檬肖像”，E-GRPO 成功生成了一幅将太空时代美学与预期构图结构相结合的肖像。同样，对于提示词“道路两侧有许多建筑，中间是一条非常蜿蜒的道路”，与基线方法相比，我们的方法更准确地捕捉了“蜿蜒”这一特征，并实现了更高的美学质量。这些结果进一步验证了，专注于高熵步骤能够使 E-GRPO 进行更有效的探索，并更好地契合复杂的人类偏好。

表4. **熵阈值消融研究** τ。结果基于 HPS 奖励设置报告。阈值为 $\tau = 0$ 时对应不进行步合并的基线。默认选择（$\tau = 2.2$）取得了整体最佳性能。每列中的最佳结果以 **粗体** 标示。

阈值（τ）	HPS	CLIP	PickScore	ImageScore
0（不合并）	0.384	0.349	0.230	1.297
1.8	0.383	0.352	0.232	1.293
2.0	0.384	0.344	0.231	1.269
2.2（我们的方法）	**0.391**	**0.355**	**0.233**	**1.324**
2.6	0.388	0.355	0.233	1.320

7.2. 失败案例

尽管 E-GRPO 具有鲁棒性，但在处理具有挑战性的提示词时，我们观察到若干反复出现的失败模式。**奖励劫持**。如正文所述，仅使用 HPS 奖励往往会生成过度饱和图像，因此需要 CLIP 奖励作为制衡。尽管如此，某些情况下仍会发生奖励劫持。例如，在图 8所示的提示词中，例如“一个在空间站舱体中睡觉的水母。”以及“图像描绘了外星花卉和植物，周围环绕着外骨骼构造，背景是神话山脉，采用戏剧性对比照明，并以三维渲染呈现超现实主义精细刻画。”模型偶尔会引入本不应出现的人脸或类人形状。这些瑕疵反映出模型倾向于利用奖励模型中的偏差；这是许多基于强化学习的训练框架普遍存在的局限。提高奖励模型的可靠性，对于推动强化学习在视觉生成领域的发展至关重要。

总的来说，这些观察结果凸显了基于强化学习的视觉生成系统面临的若干关键挑战。未来的研究可以针对这些已发现的局限性探索解决方案。

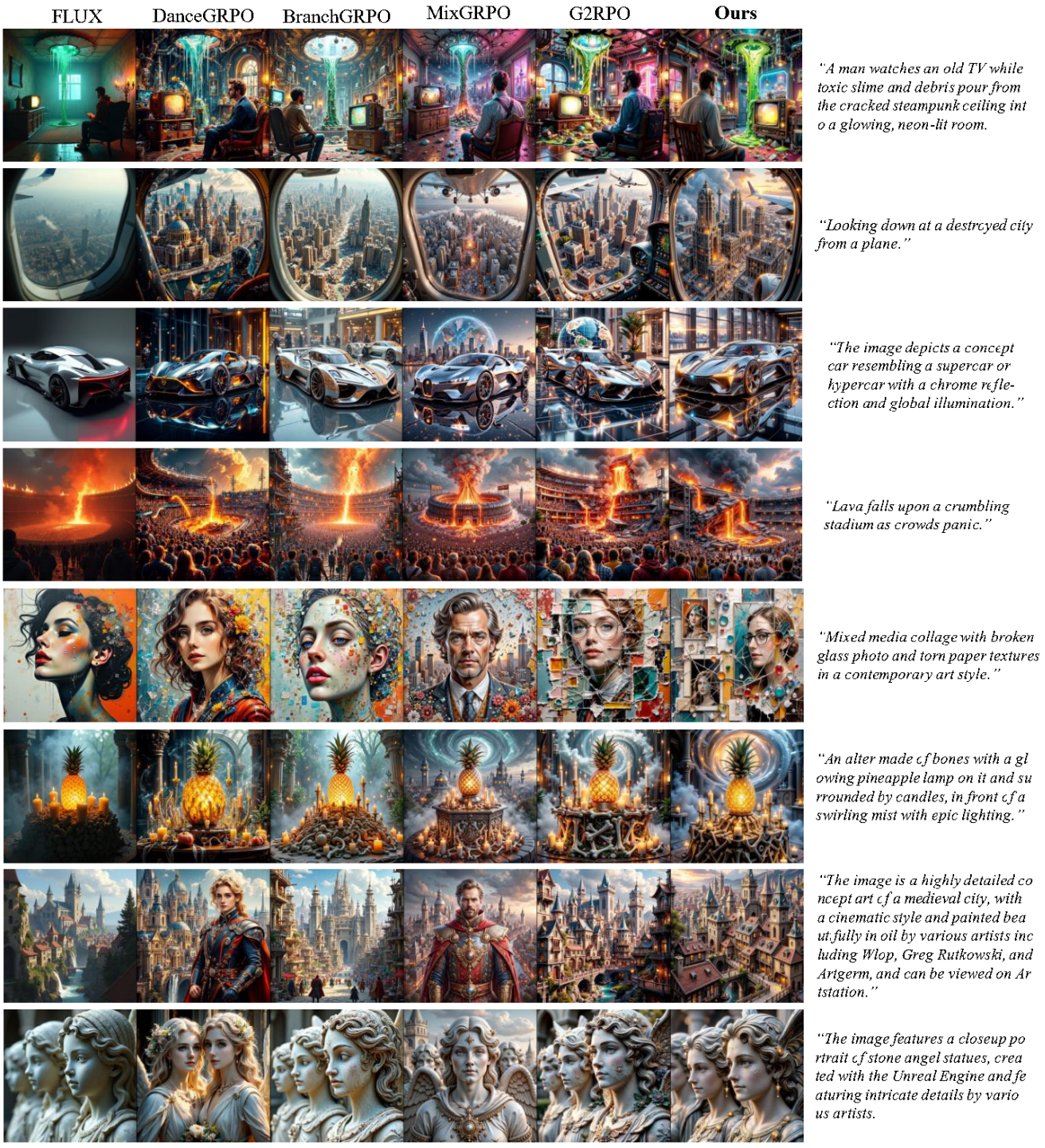


图6. **E-GRPO 与其他基线方法的更多可视化对比。**

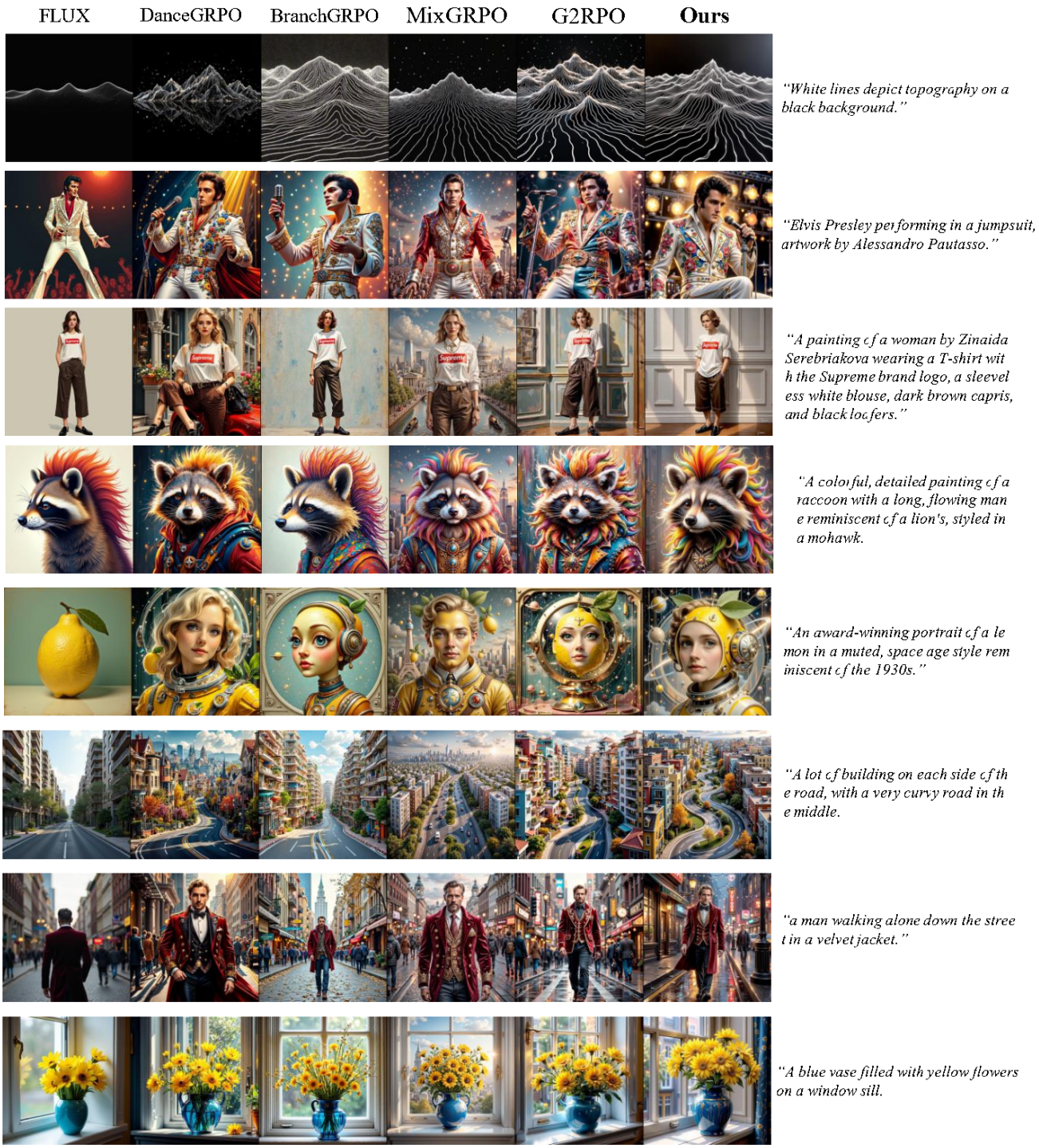


图7. **E-GRPO 与其他基线方法之间的更多可视化对比。**

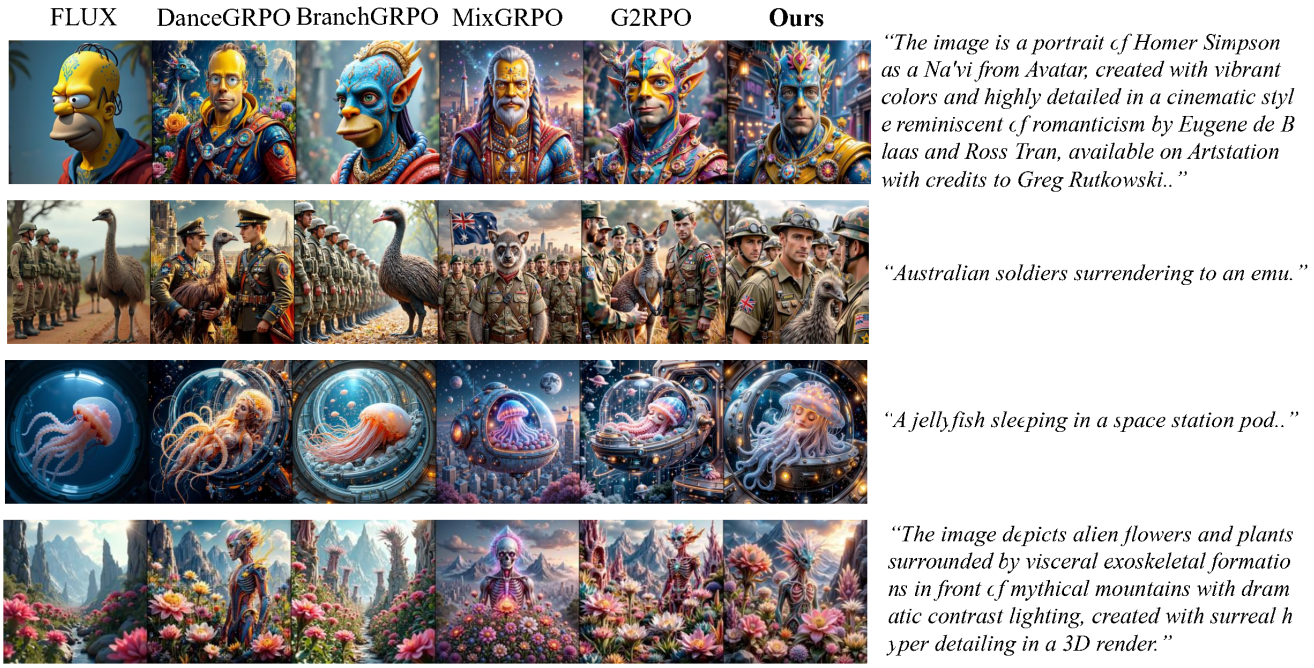


图 8. **E-GRPO 的失败案例**