arXiv:2602.01382v1 [cs.CV] 1 Feb 2026

提示词强化学习：提示词对基于流的图像生成中的强化学习至关重要

Fu-Yun Wang [1 2 *] **Han Zhang** [3] **Micha¨el Gharbi** [2] **Hongsheng Li** [1] **Taesung Park** [2]

摘要

流匹配模型（流匹配模型）彻底改变了文本到图像生成，而强化学习则是实现奖励目标对齐的关键后训练策略。在这项研究中，我们发现，当前针对流匹配模型的强化学习流程存在两个尚未得到充分重视却十分重要的局限：由于生成多样性不足而导致的样本效率低下，以及显著的提示词过拟合——模型会记住特定的训练表述，并在面对语义等效但风格各异的提示词时，出现明显的性能崩溃。我们提出了**PromptRL**（**P**rompt**M**atters in **RL** for 基于流的图像生成），这是一个将语言模型作为可训练的提示词优化智能体，直接整合到基于流的强化学习优化循环中的框架。这一设计带来了两项互补优势：能够快速开发复杂的提示词改写能力；更重要的是，它形成了一种协同训练机制，重塑了优化动态。PromptRL 在多个基准上取得了最先进的性能，GenEval 得分为 0.97，光学字符识别准确率为 0.98，PickScore 得分为 24.05。此外，我们还验证了该强化学习方法在大规模图像编辑模型上的有效性：仅通过 0.06 百万次 rollout，就将 FLUX.1-Kontext 的 EditReward 从 1.19 提升至 1.43，超过得分为 1.37 的 Gemini 2.5 Flash Image（又称 Nano Banana），并达到与 ReasonNet（1.44）相当的性能；后者依赖细粒度数据标注和复杂的多阶段训练。大量实验从实证角度表明，与朴素的纯流强化学习相比，PromptRL 始终能够达到更高的性能上限，同时所需的 rollout 数量减少超过 2×。我们的代码可在 https://github.com/G-U-N/UniRL获取。

1. 引言

流匹配模型（FMs）的出现（Liu 等人，2022；Lipman 等人，2022；Labs，2024年；Song 等人，2020）改变了文本到图像生成，使其能够根据自然语言描述合成逼真图像。为使这些模型与人类偏好及特定奖励目标对齐，强化学习（RL）（Sutton，2018；Fan 等人，2023；Black 等人，2023）已成为标准的后训练机制，在监督式预训练的范围之外进一步优化模型行为。尽管取得了这些进展，将强化学习应用于流匹配模型仍然存在严重的样本低效问题。

我们的研究发现，当前基于流的强化学习流程中存在两种尚未受到充分重视、却至关重要的失效模式。（*i*）首先，我们观察到一种反直觉的探索悖论：随着文本到图像模型在精确遵循提示词方面不断改进，它们在相同提示词下的生成多样性却同时下降。这种提升的提示词遵循度限制了有效强化学习探索所需的行为变化，导致优化停滞在生成空间的狭窄模式中。（*ii*）其次，我们发现了严重的提示词过拟合，即模型学会利用训练提示中的表层语言模式，而非形成真正的视觉理解。这种过拟合表现为，当模型在测试时遇到语义等价但句法不同的提示词时，性能会急剧崩溃。这使得提示词增强（PE）（Hao 等人，2023；Rosenman 等人，2024年；Ma˜nas 等人，2024年；Mo 等人，2024年）这一提升生成质量的关键技术，对经强化学习微调的基础模型变得无效甚至适得其反。我们在第3节中为这两种现象提供了详细的实证证据。

这些局限揭示了现有方法在设计上的一个根本疏漏：将提示词视为固定输入，而非优化过程中的可变组成部分。随机同义词替换或基于规则的改写等朴素增强技术并不够用，往往无法大规模生成语义连贯的变体。本文探讨这样一种假设：大型语言模型（LM），在通过联合强化学习训练为自适应协同学习者时

[*]在 Reve 实习期间完成的工作。[1]香港中文大学，香港 [2] Reve，美国 [3]Meta Superintelligence Labs，美国。通信作者：Han Zhang <hanzhang.ai@gmail.com>，Hongsheng Li<hsli@ee.cuhk.edu.hk>。

预印本。*2026*年*2*月*3*日。

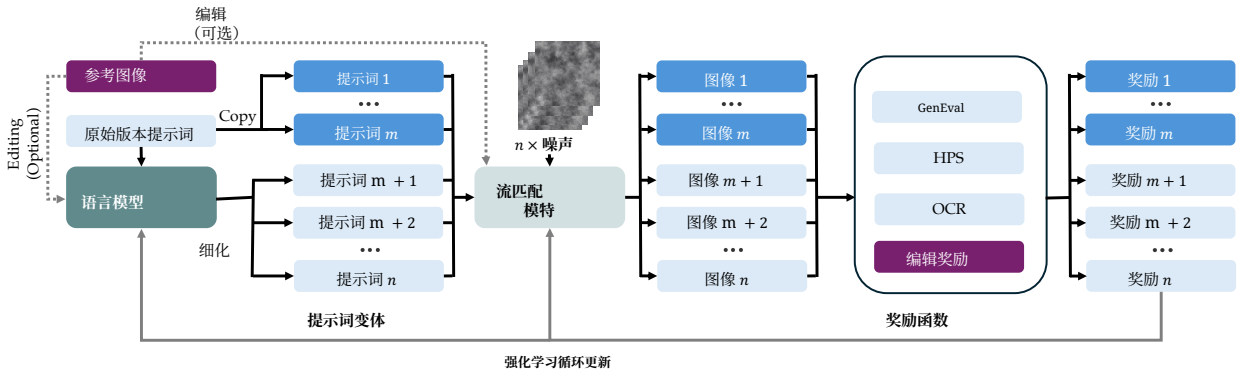


图*1.* **PromptRL框架概览。** PromptRL在统一的强化学习回路中联合训练语言模型与流匹配图像生成器。给定原始提示词（以及可选的参考图像），语言模型会生成具有语义依据的提示词变体，从而将探索空间扩展到固定提示训练之外。这些提示词与相互独立的噪声样本配对，并输入流匹配模型以生成多样化的图像。多种奖励函数的组合会对每张图像进行评估，并引导语言模型（以改进提示词改写）与流匹配模型（以改进视觉生成）的演化。

能够生成有语义依据的提示词变体，从而提升探索效率，并在实际部署中充当协同训练的提示增强（PE）模块。

我们提出**PromptRL**（**提示词**在**强化**学习**中至关重要**），这是一个将语言模型作为自适应协同学习者整合到基于流的强化学习训练循环中的框架，如图1所示。我们并非将语言模型用作静态预处理器，而是训练它们生成既能保持语义意图、又能最大化下游图像生成奖励的提示词变体。由此形成互利的训练动态：语言模型生成的多样化提示词扩展了基础模型的探索空间，从而加速策略改进；而来自流模型输出的奖励信号则引导语言模型发现语言上富于变化、语境上仍然恰当的改写。

我们的实验结果表明，PromptRL在多个基准上取得了最先进的性能，在GenEval（Ghosh等人，2023）上获得0.97分，在OCR准确率（Cui等人，2025）上获得0.98分，在PickScore（Kirstain等人，2023）上获得24.05分。此外，我们在大规模图像编辑模型上验证了我们的强化学习方法的有效性，提升了EditReward（Wu 等人，2025b）所对应的FLUX.1-Kontext（Labs等人，2025）的得分，使其从1.19提升至1.43，仅使用6万次rollout。大量实验表明，PromptRL始终能达到更高的性能上限，同时相比现有方法，所需rollout最多可减少 2×，并对多样化的提示词表述保持稳健的泛化能力。这些发现将语言-视觉协同优化确立为生成模型中高效且稳健的偏好学习的基本原则。

2. 相关工作

图像生成的强化学习。面向基于流的图像生成的强化学习已历经多种范式的演进。早期的可微奖励方法（例如DRaFT（Clark等人，2023）、AlignProp（Prabhudesai等人，2023）、ReFL（Xu等人，2023））从ImageReward等预训练奖励模型反向传播梯度，方法简单，但容易出现奖励黑客问题，如过饱和。基于强化学习的方法（例如DDPO（Black等，2023）、DPOK（Fan等，2023））将去噪视为马尔可夫决策过程，并应用PPO或其变体进行对齐，其规模化版本已展现出潜力（Zhang等人，2024年）。直接偏好优化借助Diffusion-DPO（Wallace等人，2024年）取得进展，可在偏好对上进行免仿真训练，随后由D3PO（Yang等人，2024年）和SPO（Liang等人，2024年）进一步扩展，从而在无需绝对分数的情况下结合强化学习与直接目标。DiffusionNPO（Wang等人，2025a）利用负向/反向偏好来避免不良模式，并提升细节、光照和结构。Flow-GRPO（Liu等人，2025a）与DanceGRPO（Xue等人，2025）将基于DDPM的强化学习方法（Black等，2023）扩展至流匹配模型，通过将流常微分方程转换为随机微分方程来实现。相关工作还包括DiffusionNFT（Zheng 等，2025），它通过流匹配在前向加噪过程上执行高效的在线强化学习——对比正负样本，且无需反向梯度、轨迹存储或求解器约束。MixGRPO（Li等人，2025）进一步通过混合ODE–SDE采样和滑动窗口调度，提升了流设定下GRPO的效率。

面向图像生成的提示词增强 提示词增强（PE）已成为提升文本到图像（T2I）生成质量与对齐的关键。早期方法依赖人工优化，而近期方法则利用语言模型实现自动化提示词优化。Promptist（Hao 等人，



图 *2*。 **基于流的 T2I 模型中的质量-多样性困境及其通过提示词优化进行的缓解**。 随着模型从 Stable Diffusion v1-5 (a) 发展到 FLUX.1-dev (b)，其文本-图像对齐（TI-Sim）和美学质量（P.S.）得到提升，但输出多样性（II-Sim）却大幅降低，从而为强化学习优化造成探索瓶颈。基于语言模型的提示词优化（c）在保持质量的同时部分恢复了多样性，表明语言变体能够拓展探索空间。每一行中的所有图像均使用相同的随机种子，以隔离提示词条件控制的影响。

2023）将监督微调与强化学习相结合，在保留用户意图的同时优化提示词以提升美学吸引力。NeuroPrompts（Rosenman 等人，2024年）提出受约束文本解码，用于具有用户可控风格的自动提示词增强。OPT2I（Ma˜nas 等人，2024年）利用语言模型迭代精炼提示词，以最大化一致性得分。RePrompt（Wu 等人，2025c）融入思维链推理与奖励引导训练，以实现结构化提示词重构。PAE（Mo 等人，2024年）通过带有多目标奖励的在线强化学习，动态优化词权重与注入时间步。然而，这些方法仅限于提示词级强化学习，收益有限——例如，RePrompt 仅将 FLUX.1-dev 的 GenEval 得分从 0.66 提升至 0.76——且缺乏对提示词级与流级强化学习协同作用的探索。除生成之外，近期工作还将语言模型用于基于指令的图像编辑。MGIE（Fu 等人，2023）将简短的用户指令转化为富有表现力的编辑指引。SmartEdit（Huang 等人，2024年）引入双向交互模块，以应对复杂推理场景。Emu Edit（Sheynin 等人，2024年）利用语言模型输出详细的编辑指令与目标图像描述。

3. 理解流强化学习的低效问题

3.1. 生成困境：质量对比。多样 ity

我们观察到，在现代文生图基础模型中，生成质量与输出多样性之间存在根本性张力。随着模型精确遵循文本提示词的能力不断提升，它们同时也牺牲了有效强化学习探索所必需的生成多样性。为量化这一现象，我们使用 CLIP ViT-g-14（Radford 等人，2021）测量文本-图像相似度（TI-Sim）以评估提示词对齐，使用 PickScore（P.S.）评估美学质量，并借助 CLIP ViT-g-14 计算图像-图像相似度（II-Sim），以衡量相同提示词下的输出多样性。图2展示了这种质量-多样性权衡：Stable Diffusion v1-5（列 a）的生成结果具有中等的 TI-Sim 和 P.S.，但保持了相当大的多样性（II-Sim 为 0.58–0.72）。相比之下，FLUX.1-dev（列 b ）取得了明显更高的美学分数，但生成结果的多样性大幅降低（II-Sim 为 0.92–0.93）。这一模式在语义多样的提示词中始终一致，从奇幻生物到超现实构图场景皆是如此。值得注意的是，基于语言模型的提示词细化（列 c）部分缓解了这种坍缩，将 II-Sim 降至 0.84–0.85，同时保留了质量指标。

这种质量-多样性困境直接削弱了基于流的模型中的强化学习优化

表1。不同模型和评估指标的性能比较。所有指标均在分辨率为 1024 、欧拉步数为 20 的条件下进行测试。PE 表示使用 Qwen-2.5-VL 进行提示词增强。绿色 和 红色 数字表示应用 PE 后的性能变化。P.S. 和 U.R. 分别表示 PickScore 和 UnifiedReward。

模型	使用 PE 的 GenEval		OCR	P.S.	HPS	U.R.
SD3	✗	0.58	0.48	22.30	26.95	2.982
SD3	✓	0.63 (+0.05)	0.53 (+0.05)	22.34 (+0.04)	27.67 (+0.72)	3.140 (+0.158)
DiffusionNFT	✗	0.88	0.89	23.63	31.78	3.392
DiffusionNFT	✓	0.77 (-0.11)	0.86 (-0.03)	23.21 (-0.42)	30.66 (-1.12)	3.268 (-0.124)
FlowGRPO	✗	0.92	0.89	23.33	—	—
FlowGRPO	✓	0.81 (-0.11)	0.86 (-0.03)	23.13 (-0.20)	—	—

。随着生成策略变得日益确定，展开轨迹会坍缩到输出空间的狭窄模式中，导致奖励信号退化：当所有样本都聚集在相似的高质量输出附近时，优势估计器会失去策略改进所必需的比较信息。这一探索瓶颈又因严重的提示词过拟合而进一步加剧，在这种情况下，模型利用的是表面的词汇模式，而非语义理解；我们将在下一小节对此进行探讨。

3.2. 提示词语言操纵

除了质量-多样性困境，我们还发现了第二种关键失效模式：提示词语言操纵，即经强化学习训练的模型利用表层词汇模式，而非形成稳健的语义理解。我们通过在原始提示词和由 Qwen-2.5-VL 生成的语义保持改写版本（即，提示增强）(Bai 等人，2025)

如表1所示，预训练 SD3（Esser 等人，2024年）展现出语言鲁棒性，在所有指标上，即使进行释义改写，性能也始终保持稳定或有所提升。然而，仅流匹配强化学习模型表现出严重的提示词过拟合。DiffusionNFT 在原始提示词上取得了出色的性能，但在释义改写下遭遇灾难性退化。同样，在 GenEval 上训练的 FlowGRPO 在提示词经过释义改写后，得分从 0.92 降至 0.81。这表明学习到的策略记忆的是表层语言特征，而不是理解潜在的视觉概念。这种过拟合发生在提示词条件层面，无法通过标准正则化解决。更为关键的是，对预训练流匹配模型有益的提示增强技术，在仅流匹配强化学习之后会失效，甚至产生有害影响，因为微调后的模型会对特定的提示词分布产生过拟合。这一观察促使我们采用联合 LM-FM 优化方法：我们不将提示增强作为固定的预处理步骤，而是以共生方式共同演化提示增强器与生成器。

4. PromptRL

4.1. 将语言模型作为动态提示词优化器

为应对第3节中发现的探索崩溃和提示词过拟合这两大挑战，我们提出在强化学习训练循环中引入语言模型，作为自适应提示词优化器。不同于依赖基于规则的变换或同义词替换的静态增强技术，我们的方法利用预训练语言模型的语义理解能力和组合灵活性，生成具有上下文依据的提示词变体。

形式上，给定一个来自训练分布的原始提示词 p_0 ，我们使用语言模型 $\pi_{\text{LM}}(\cdot|p_0)$ 生成一组经过精炼的提示词 $\{p_1, p_2, \ldots, p_k\}$，在保留核心语义意图的同时引入语言多样性。随后，将每个精炼后的提示词 p_i 与一个独立的噪声样本 $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I)$ 配对，并输入流匹配模型 $\pi_{\text{FM}}(\cdot|p_i, \epsilon_i)$，以生成多样化的图像样本 $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$。这种架构构建了一种分层探索机制：语言模型探索语义等价描述的语言流形，而流匹配模型则探索以每种提示词变体为条件的视觉流形。

关键在于，我们在训练过程中引入了提示词保留机制：对于总计 n 个样本的每个批次，我们保留$m < n$ 个使用原始提示词 p_0 （未经语言模型精炼）的样本，其余 $n-m$ 个样本则通过语言模型进行增强。这一设计有两个互补的目的。首先，未经修改的提示词提供了强基线，用于优势估计——在策略优化期间，奖励低于 p_0 的增强提示词实际上会被剔除，从而避免在提示词空间中的低奖励区域进行无效探索。其次，持续接触原始提示词可确保流匹配模型在训练分布上保持稳健的性能，避免模型在推理时过度依赖语言模型的精炼结果。

4.2. 对分离的语言模型进行联合强化学习训练以及 FMs

在将语言模型确立为动态提示词优化器之后，我们现在描述联合强化学习训练流程，该流程在统一的策略梯度框架内同时优化 π_{LM} 和 π_{FM} 。关键在于，尽管这两个模型共享奖励信号，它们在架构上仍然分离——梯度不会在语言模型与流匹配模型之间传播，从而保持模块化与计算效率。

在每次训练迭代中，我们采样一个批次的 B 原始提示词 $\{p_0^{(1)}, \ldots, p_0^{(B)}\}$，这些提示词来自训练分布。对于每个提示词 $p_0^{(j)}$ ，我们生成 n 个样本，所用过程见第4.1节：m 个样本直接从 $p_0^{(j)}$ 以不同的噪声种子生成，而其余

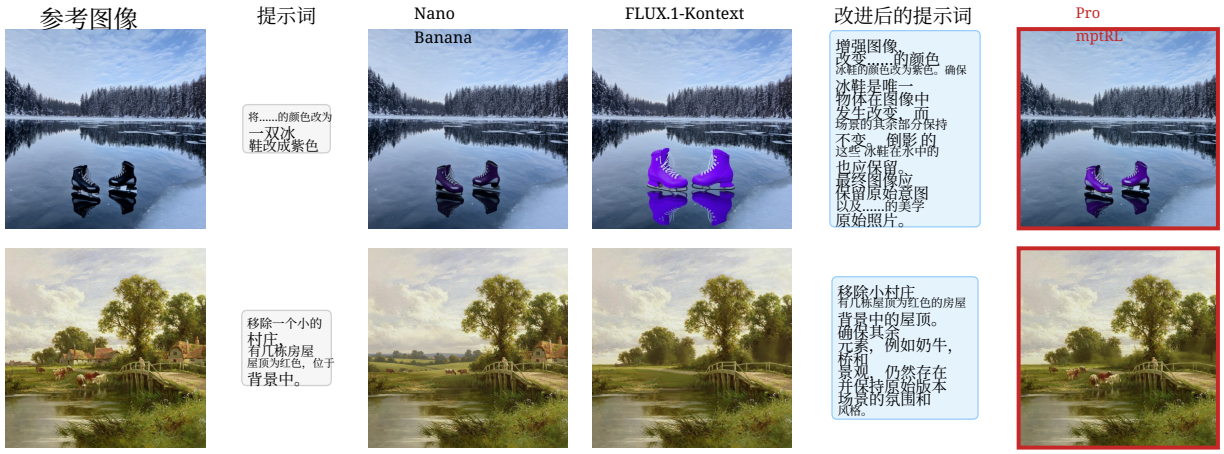


图 3. 指令式图像编辑任务的定性比较。我们的方法使语言模型能够利用原始图像的视觉信号，将模糊的编辑指令转化为更明确且针对图像的提示词，从而提升编辑性能。

$n-m$ 个样本则使用经语言模型细化的提示词 $\{p_i^{(j)}\}_{i=1}^{n-m}$，这些提示词采样自 $\pi_{\text{LM}}(\cdot|p_0^{(j)})$。每个提示词（原始版本或细化后的提示词）都与一个独立的潜在噪声向量 $\boldsymbol{\epsilon} \sim \mathcal{N}(0,\ I)$ 配对，以通过流匹配模型生成图像。

生成的图像使用复合奖励函数 $R(\cdot)$ 进行评估，该函数汇总了格式奖励和图像生成奖励：

$$R(\mathbf{x}, p) = \lambda_{\text{Format}} R_{\text{Format}}(p) + \lambda_{\text{Gen}} R_{\text{Gen}}(\mathbf{x}, p), \quad (1)$$

其中，$R_{\text{Format}}(p)$ 是一种二元奖励，要求语言模型输出以 XML 标签<answer> 和 </answer>括起来的细化提示词变体。这种格式约束确保了结构化输出解析，并对格式错误的生成进行惩罚；若输出符合所需格式，则 $R_{\text{Format}}(p) = 1$，否则为 0。术语 $R_{\text{Gen}}(\mathbf{x}\ p)$ 表示图像生成奖励模型（例如，用于评估组合准确性的 GenEval）。超参数 λ_{Format}和 λ_{Gen} 用于平衡格式合规性与生成质量；为简洁起见，实践中我们设定 $\lambda_{\text{Format}} = 1.0$ 和 $\lambda_{\text{Gen}} = 1.0$。遵循 GRPO，我们通过组内归一化计算优势。对于每个提示词 $p_0^{(j)}$ 及其 n个关联样本 $\{\mathbf{x}_1^{(j)}, \ldots, \mathbf{x}_n^{(j)}\}$，我们计算：

$$A(\mathbf{x}_i^{(j)}, p_i^{(j)}) = \frac{R(\mathbf{x}_i^{(j)}, p_i^{(j)}) - \mu^j}{\sigma^j + \epsilon}, \quad (2)$$

其中，μ^j 和 σ^j 分别是第 j组内奖励的均值和标准差，ϵ 是用于保证数值稳定性的小常数。这种按组归一化使优势信号不受不同提示词间奖励尺度的影响，从而能够在多样化的语义类别上实现稳定优化。它还天然实现了一种自我竞争机制：由相同提示词生成的样本会相互竞争，促进组内多样性。

至关重要的是，尽管语言模型和流匹配模型在 PromptRL 中共享奖励信号，但二者在架构上仍然相互独立——梯度不会在语言模型和流匹配模型之间传播，从而保持模块化并提高计算效率。对于语言模型，我们只保留与经 $n-m$ 语言模型细化的样本相对应的优势，不包括由原始提示词生成的样本。语言模型的策略梯度为：

$$\nabla_{\theta_{\text{LM}}} \mathcal{J}_{\text{LM}} = \mathbb{E}_{p_0}\left[\sum_{i=m+1}^{n} A(\mathbf{x}_i, p_i) \cdot \nabla_{\theta_{\text{LM}}} \log \pi_{\text{LM}}(p_i|p_0)\right]. \quad (3)$$

这种选择性优势分配确保语言模型学会生成优于基线的提示词变体。表现不佳的变体会获得负优势并被降低权重，从而防止语言模型引入适得其反的语言变化。

对于流匹配模型，我们利用所有 n 个样本（包括原始提示词和细化后的提示词）的优势来更新图像生成策略：

$$\nabla_{\theta_{\text{FM}}} \mathcal{J}_{\text{FM}} = \mathbb{E}_{p_0}\left[\sum_{i=1}^{n} A(\mathbf{x}_i, p_i) \cdot \nabla_{\theta_{\text{FM}}} \log \pi_{\text{FM}}(\mathbf{x}_i|p_i, \boldsymbol{\epsilon}_i)\right]. \quad (4)$$

纳入原始提示词的优势，可确保流匹配模型在基础分布上保持强劲表现，同时受益于语言模型优化所带来的扩展探索空间。

值得注意的是，这种联合优化无需对任一模型进行架构或算法上的修改——我们只需利用共享奖励作为更新信号，通过各自的参数分别反向传播策略梯度。我们的框架不依赖于特定的强化学习算法；虽然我们采用 GRPO（Shao等人，2024年）作为基线实现，因其简单易用，但 PromptRL 的组件

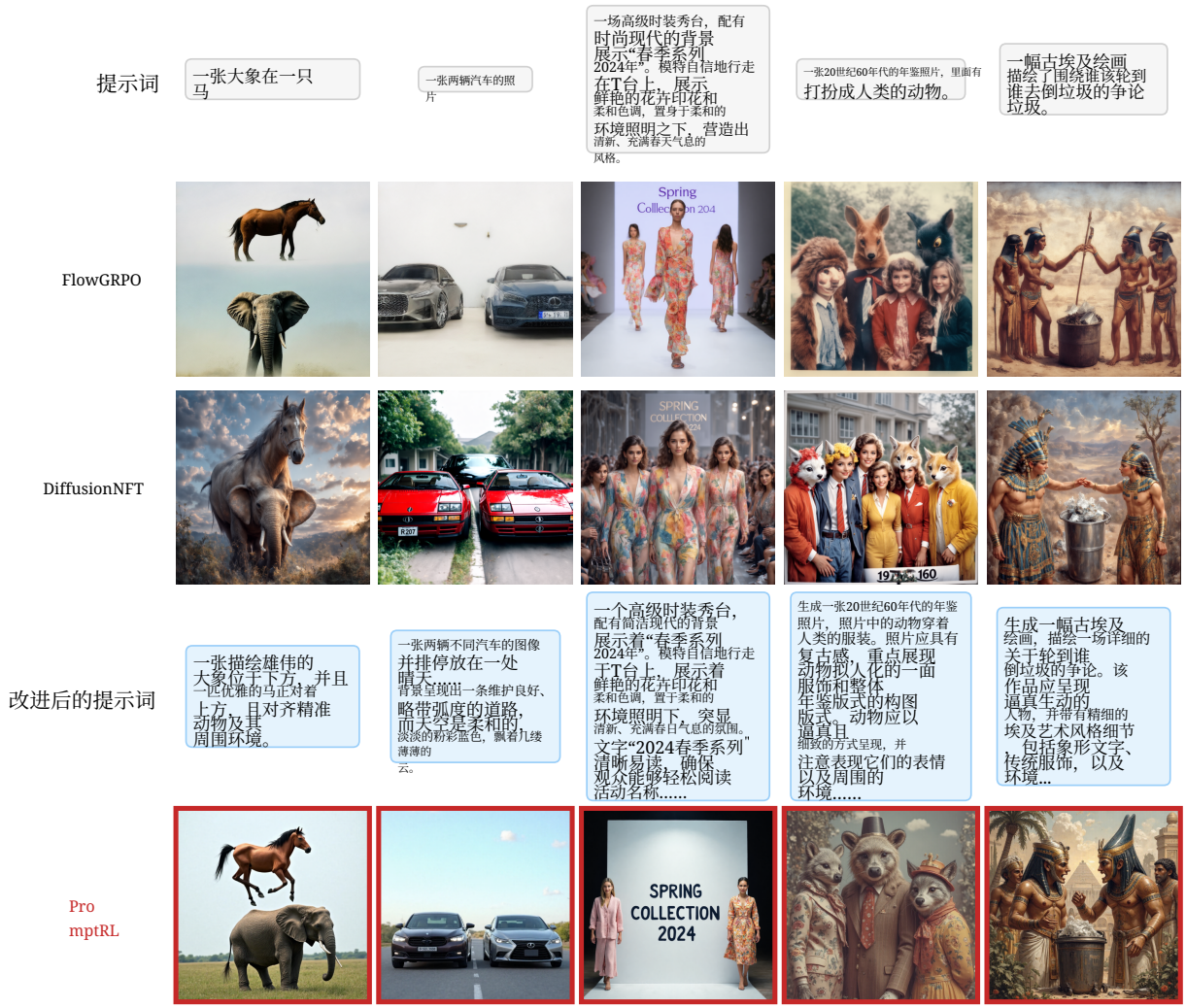


图*4.* 文本到图像生成的定性比较。前两个提示词来自 GenEval。第三个提示词来自 OCR-1k。最后两个提示词来自 Drawbench。

应能够与其他适用于语言模型的在线强化学习方法（例如，ReMax（Li 等人，2023））以及流匹配模型（例如，DiffusionNFT（Zheng 等人，2025））集成。

4.3. 通过奖励标记进行多奖励训练 g

除单一奖励场景外，我们还在 GenEval、PickScore 和 OCR 上验证了 PromptRL 在多重奖励优化下的表现。一个已知挑战是，不同奖励模型的尺度和方差差异极大，因此需要繁琐地手动调节奖励系数。替代方案采用多阶段训练流程（例如，DiffusionNFT 先基于 PickScore 进行训练，然后再过渡到其他奖励），但这引入了调度复杂性。我们提出了一种简单而有效的解决方案：单一奖励-逐样本训练。我们不再计算加权奖励总和，而是为每个训练提示分配一个类别标签，该标签指明由哪个奖励函数评估其生成图像。在进行按组优势计算时，我们在每个奖励类别内部执行归一化，从而确保尽管尺度不同，优势信号仍具有可比性。这完全消除了奖励系数调优——每个奖励模型都在其原生尺度下运行，互不干扰。实证结果表明，这种简单的标记机制无需显式的奖励工程或多阶段课程学习，便能在所有目标上同时取得出色表现。

5. 实验

5.1. 实验设置

基础模型。 为了全面验证 PromptRL 在基于流的图像生成中的效果，我们在文本到图像合成和指令式图像编辑任务上进行评估。我们采用 FLUX.1-

dev 作为文本到图像生成的基础流匹配模型，并采用 FLUX.1-Kontext 进行图像编辑。对于语言模型组件，我们使用 Qwen2.5-VL-3B-Instruct 作为提示词优化器。

数据集与基准。 对于文本到图像生成，遵循 FlowGRPO 的做法，我们从三个互补维度进行评估：组合准确性（GenEval）、文本渲染能力（光学字符识别）和人类偏好对齐（PickScore）。对于 GenEval 和光学字符识别目标，我们使用 FlowGRPO 中对应的训练集。对于 PickScore 优化，我们在 Pick-a-Pic 数据集上训练。对于指令式图像编辑，我们从 OmniEdit（Wei 等人，2024年）训练集中随机抽取 10,000 个样本，仅保留编辑指令及相应的参考图像。对于文本到图像生成，我们使用官方 GenEval 验证集进行组合评估，使用 Draw-Bench（Saharia 等人，2022）进行美学质量评估（PickScore、HPS（Wu 等人，2023）、UnifiedReward（Wang等人，2025b）），并使用三个光学字符识别基准：FlowGRPO OCR-1k 验证集、TMDB，以及 MARIO-Eval（Chen 等人，2023）中的 OpenLib。对于图像编辑，我们直接在 OmniEdit 验证集上进行评估，该验证集涵盖六类编辑：对象替换、对象添加、对象移除、属性修改、环境变更和风格迁移。

训练详情。 为提高计算效率，我们在训练 FLUX.1-dev 时以 512×512 分辨率进行 rollout，推理步数设为 20。然而，我们发现，降低分辨率会显著削弱 FLUX.1-Kontext 的编辑能力；因此，在编辑实验中，我们仍采用 1024×1024分辨率。为提高高分辨率训练这一高计算需求场景下的训练效率，我们将推理步数降至 8，并遵循先前工作（Li 等人，2025）的做法，仅在每次 rollout 的前 4 个去噪步数中进行随机微分方程求解。该研究表明，图像结构在早期时间步中基本趋于稳定。所有模型均使用 GRPO 算法进行训练，每个提示词的组大小为 $n = 8$个样本，提示词保留数量为 $m = 2$。

5.2. 定量比较

文本到图像生成。 如表2所示，PromptRL 在 GenEval 上取得最先进性能，总分达到 **0.97**，优于得分为 0.92 的 FlowGRPO 和得分为 0.88 的 DiffusionNFT。PromptRL w/ PE 在位置和计数两项上均获得近乎完美的 0.99 分，展现出卓越的组合准确性。值得注意的是，即使推理时不进行提示词增强，PromptRL w/o PE 仍达到 0.94 分，证实联合训练能够赋予模型稳健的视觉理解能力，不依赖语言模型优化。表3 进一步验证了 PromptRL 在美学和光学字符识别指标上的表现。我们的方法取得了 24.04 的 PickScore、32.03 的 HPS

表 *2*。不同模型在 GenEval基准上的性能比较。分数越高表示性能越好。最佳结果以**粗体**显示。带 * 的模型指标取自 Qwen-Image 论文。

模型	单物体	双物体	Cnt.	Clr.	Pos.	属性	平均值↑
Show-o* (Xie 等人，2024年b)	0.95	0.52	0.49	0.82	0.11	0.28	0.53
Emu3-Gen* (Wang等人，2024年)	0.98	0.71	0.34	0.81	0.17	0.21	0.54
SD3 Medium* (Esser 等人，2024年)	0.98	0.74	0.63	0.67	0.34	0.36	0.62
FLUX.1-dev* (Labs，2024年)	0.98	0.81	0.74	0.79	0.22	0.45	0.66
SD3.5 Large*	0.98	0.89	0.73	0.83	0.34	0.47	0.71
JanusFlow*(Ma 等人，2025)	0.97	0.59	0.45	0.83	0.53	0.42	0.63
Janus-Pro-7B* (Chen 等人，2025)	0.99	0.89	0.59	0.90	0.79	0.66	0.80
HiDream (Cai 等人，2025b)	1.00	0.98	0.79	0.91	0.60	0.72	0.83
Seedream 3.0* (Gao 等人，2025)	0.99	0.96	0.91	0.93	0.47	0.80	0.84
Qwen-Image* (Wu 等人，2025a)	0.99	0.92	0.89	0.88	0.76	0.77	0.87
基于强化学习							
RePrompt (Wu 等人，2025c)	0.98	0.87	0.77				0.76
FlowGRPO (Liu 等，2025a)	1.00	0.99	0.91			0.80	0.92
DiffusionNFT (Zheng 等，2025)	1.00	0.98	0.74	0.92	0.85	0.80	0.88
PromptRL（不使用提示增强）	1.00	0.96	0.95	0.95	0.93	0.85	0.94
PromptRL（使用提示增强）	**1.00**	**0.99**	**0.99**				**0.97**

表 *3.* 不同模型在美学和光学字符识别指标上的性能比较。

	美学			*OCR*		
模型	P.S.	HPS	U.R.	OCR-1k	TMDB	OpenLib
SD1.5(Rombach 等，2022)	20.92	23.71	2.00	0.05	0.13	0.08
SDXL (Podell 等人，2023)	22.14	26.67	2.78	0.13	0.20	0.09
SD3Medium(Esser 等人，2024年)	22.38	28.56	3.09	—	0.44	0.33
FLUX.1-schnell(Labs, 2024年)	22.64	29.39	3.25	0.54	0.66	0.50
FLUX.2-klein (Labs, 2025)	22.79	29.03	3.29	0.55	0.22	0.46
Z-Image (Cai 等人，2025a)	20.14	28.22	3.51	0.70	0.71	0.83
Qwen-Image（Wu 等人，2025a）	23.05	30.40	3.53	0.65	0.79	0.94
Qwen-Image-2512	23.16	30.79	3.40	0.72	0.81	0.87
基于强化学习						
FlowGRPO	23.33			0.89	0.83	0.73
DiffusionNFT	23.63			0.89	0.91	0.86
PromptRL（不含提示增强）	24.01			0.97	0.92	0.95
PromptRL（含提示增强）	24.05	32.03	3.44	0.98	0.91	0.95

以及在 OCR-1k 上 0.98 的 OCR准确率，始终优于先前的基于强化学习的方法。这些在多种指标上的提升表明，语言模型与流匹配模型的联合优化能够泛化到单目标奖励优化之外。

指令式图像编辑。 对于指令式图像编辑，如表 5所示，采用提示增强的 PromptRL 将 FLUX.1-Kontext 基线的表现从 1.19 提升至**1.43**，接近依赖细粒度密集数据标注及复杂多阶段训练流程的 ReasonEdit-Think（1.44）。在移除（+0.69）和环境（+0.28）上的提升最为显著。值得注意的是，不进行联合训练而直接应用提示词增强，会使 FLUX.1-Kontext 的表现从 1.19 降至 1.01，这表明 PromptRL 的联合优化对于有效的提示词优化至关重要。

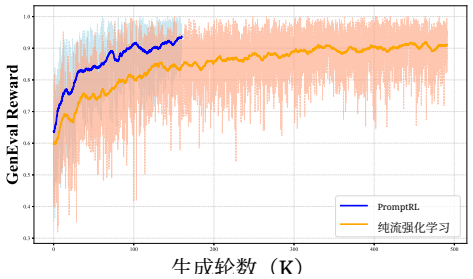
多重奖励训练。 我们在表 4 中评估了基于标签的多重奖励训练策略。尽管没有调整奖励系数，也没有采用多阶段课程，多重奖励模型在所有目标上都取得了有竞争力的表现（GenEval：0.93，光学字符识别：0.96，PickScore：23.94），与单奖励专用模型相比，性能仅略有下降。

表4. 单奖励训练与多奖励训练的比较。单奖励训练针对每个目标分别进行，而多奖励训练则采用基于标签的联合优化。

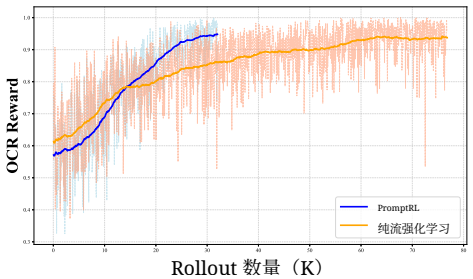
训练	GenEval↑	OCR↑	PickScore↑
单奖励	0.97	0.98	24.05
多重奖励	0.93	0.96	23.94

表5。 使用 EditReward 衡量的图像编辑任务性能比较。我们在六类编辑任务上评估模型：交换、风格、添加（添加）、属性修改（属性）、环境（环境）和移除。FLUX.1-Kontext w/ 提示增强的编辑指令由预训练的 Qwen2.5-VL 细化。

模型	Swap	风格	Add.	属性	Env.	移除	平均值↑
InstructPix2Pix（Brooks等人，2023）	-0.24	0.91	-0.45	0.45	0.48	-0.80	0.02
MagicBrush（Zhang等人，2023）	-0.38	0.36	-0.78	-0.80	0.91	-0.85	-0.27
LEDITS++（Brack等人，2024年）	-0.81	-0.32	-0.30	-0.60	-0.37	-0.97	-0.60
Qwen-Image-Edit	1.11	1.14	0.95	0.90	1.39	0.61	1.03
FLUX.2-klein	1.42	1.73	1.29	1.42	1.80	0.32	1.34
Nano Banana	1.58	1.20	1.28	1.18	1.61	1.13	1.37
Step1X-Edit (Liu 等，2025b)	1.39	1.58	1.19	1.34	1.57	0.22	1.24
ReasonEdit (Yin 等，2025)	1.51	1.43	1.19	1.47	1.58	1.14	1.40
ReasonEdit-Think	**1.52**	**1.47**	1.19	**1.44**	1.69	**1.27**	**1.44**
FLUX.1-Kontext	1.35	1.36	1.16	1.15	1.44	0.55	1.19
FLUX.1-Kontext 搭配提示增强	1.35	0.97	1.04	0.48	1.22	0.65	1.01
PromptRL 不使用提示增强	1.45	1.46	1.28	1.35	1.56	0.98	1.36
采用提示增强的 PromptRL	**1.47**	**1.43**	**1.29**	**1.39**	**1.72**	**1.24**	**1.43**



(a) GenEval 奖励的训练曲线对比。



(b) 光学字符识别奖励的训练曲线对比。

图5。 不同奖励的训练曲线对比。

5.3 定性比较

图 3 和4 展示了文本到图像生成和图像编辑任务的定性比较。对于组合式提示，PromptRL 能够正确呈现物体，并准确还原颜色和空间布局；而基线方法则出现颜色泄漏或遗漏。在 OCR任务上，PromptRL 能生成清晰易读、拼写正确的文本，同时保持美学质量。对于指令式编辑，PromptRL 联合训练的语言模型利用来自参考图像的视觉信号，生成具备图像感知能力的提示词优化结果，将模糊的指令转化为精准的图像特定提示，在修改指定区域的同时保留前景主体。

5.4 消融研究

联合强化学习可提升训练效率。我们比较了生成轮数增加时的训练动态，两种方法均使用 FlowGRPO 进行流匹配优化，并采用每个样本单次更新以确保稳定性。如图所示，

5a 和图 5b，在 GenEval 和光学字符识别目标上，PromptRL 始终以更少的 rollout 获得更高的奖励。在 GenEval 上，PromptRL 仅用约少 50% 的 rollout，就达到了与 FlowGRPO 收敛点相当的表现。在光学字符识别任务上也观察到了类似趋势，PromptRL 初期提升更快，最终表现更优。这些结果验证了语言模型生成的提示词变体所提供的扩展探索空间能够提升样本效率。

提示词保留机制。我们研究了训练期间保留原始提示词对模型的影响

表6。 提示词保留机制的消融实验。我们报告了推理时未使用提示增强、且保留的原始提示词数量不同的情况下的 GenEval 得分 m（组大小 $n = 8$）。

保留	单物体	双物体	Cnt.	Clr.	Pos.	属性	平均值↑
$m = 0$	0.99	0.94	0.84	0.89	0.51	0.83	0.83
$m = 1$	0.98	0.88	0.83	0.85	0.37	0.64	0.77
$m = 2$	**1.00**	**0.96**	**0.95**	**0.95**	**0.93**	**0.85**	**0.94**
$m = 4$	1.00	0.96	0.94	0.88	0.93	0.81	0.92

在推理时不使用 PE 的性能。对于组大小$n = 8$，我们改变保留的原始提示词数量$m \in \{0, 1, 2, 4\}$，并评估不使用 PE 时的 GenEval 性能（表6）。当 $m = 0$时，模型在训练期间从未见过原始提示词，因此在未修改的测试提示词上性能下降至 0.83。 $m = 1$ 则导致结果更差（0.76）。我们通过实证发现，语言模型很快就会发现更适合流匹配模型的提示词变体，导致单个原始提示词持续获得负优势，因而无法进行有效的梯度更新。将 $m = 2$设为该值后，性能显著提升至 0.94，因为多个原始提示词在组内相互竞争，从而产生有效的优势信号，并在原始提示词分布上实现稳健学习。

6. 结论

我们提出了 PromptRL，这是一种在统一的强化学习循环中联合训练语言模型和流匹配模型的框架，用于文本到图像生成。我们的方法利用语言模型生成的提示词变体来扩展探索空间，同时协同

进化出一个提示词增强模块，以提升推理时的生成质量。大量实验表明，PromptRL 在多个基准上均达到了最先进的性能。

影响声明

本文旨在推动语言模型和扩散模型联合强化学习训练领域的发展。我们的工作可能会产生许多潜在的社会影响，但我们认为无需在此特别强调其中任何一项。

参考文献

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S.,Dang, K., Wang, P., Wang, S., Tang, J., 等。Qwen2. 5-vl 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025。

Black, K., Janner, M., Du, Y., Kostrikov, I.,andLevine, S. 使用强化学习训练扩散模型。*arXiv* 预印本 *arXiv:2305.13301*，2023。

Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., and Passos, A. Ledits++：使用文生图模型进行无限制的图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 8861–8870 页，2024年。

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix：学习遵循图像编辑指令。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18392–18402 页，2023。

Cai, H., Cao, S., Du, R., Gao, P., Hoi, S., Hou, Z., Huang, S., Jiang, D., Jin, X., Li, L., et al. Z-image：一种采用单流扩散Transformer的高效图像生成基础模型。*arXiv*预印本 *arXiv:2511.22699*，2025a。

Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., 等。Hidream-i1：一种高效的图像生成基础模型，采用稀疏扩散Transformer。*arXiv* 预印本 *arXiv:2505.22705*，2025b。

Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., 和 Wei, F. Textdiffuser：将扩散模型用作文本绘制器。神经信息处理系统大会，36：9353–9387，2023。

Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., 和 Ruan, C. Janus-pro：通过数据和模型缩放实现统一的多模态理解与生成。*arXiv* 预印本 *arXiv:2501.17811*，2025。

Clark, K., Vicol, P., Swersky, K., 和 Fleet, D. J. 直接在可微奖励上微调扩散模型。*arXiv* 预印本 *arXiv:2309.17400*，2023。Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H. 等。Paddleocr 3.0 技术报告。*arXiv* 预印本 *arXiv:2507.05595*，2025。

Esser, P., Kulal, S., Blattmann, A., Entezari, R., M¨ uller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., 等。扩展用于高分辨率图像合成的整流流变换器。发表于第四十一届国际机器学习大会，2024年。

Fan, Y.、Watkins, O.、Du, Y.、Liu, H.、Ryu, M.、Boutilier, C.、Abbeel, P.、Ghavamzadeh, M.、Lee, K. 和 Lee, K. 用于微调文本到图像扩散模型的强化学习。载于第三十七届神经信息处理系统大会（*NeurIPS*）*2023*。神经信息处理系统基金会，2023。

Fu, T.-J.、Hu, W.、Du, X.、Wang, W. Y.、Yang, Y. 和 Gan, Z. 通过多模态大语言模型引导基于指令的图像编辑。*arXiv* 预印本 *arXiv:2309.17102*，2023。

Gao, Y.、Gong, L.、Guo, Q.、Hou, X.、Lai, Z.、Li, F.、Li, L.、Lian, X.、Liao, C.、Liu, L. 等。Seedream 3.0 技术报告。*arXiv* 预印本 *arXiv:2504.11346*，2025。

Ghosh, D.、Hajishirzi, H.和Schmidt, L. Geneval：一种用于评估文本到图像对齐的以对象为中心的框架。神经信息处理系统进展，36:52132–52152，2023。

Hao, Y.、Chi, Z.、Dong, L.和Wei, F. 优化用于文本到图像生成的提示词。神经信息处理系统进展，36:66923–66939，2023。

Huang, Y.、Xie, L.、Wang, X.、Yuan, Z.、Cun, X.、Ge, Y.、Zhou, J.、Dong, C.、Huang, R.、Zhang, R.等。SmartEdit：利用多模态大语言模型探索复杂的基于指令的图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第8362–8371页，2024年。

Kirstain, Y.、Polyak, A.、Singer, U.、Matiana, S.、Penna, J.和Levy, O. Pick-a-Pic：一个面向文本到图像生成的用户偏好开放数据集。神经信息处理系统进展，36:36652–36663，2023。

黑森林实验室 Flux。https://github.com/black-forest-labs/flux，2024年。

Labs, B. F. FLUX.2：前沿视觉智能。https://bfl.ai/blog/flux-2，2025。

Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Con- sul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., 等。Flux. 1 kontext：用于潜空间上下文图像生成与编辑的流匹配。*arXiv* 预印本 *arXiv:2506.15742*，2025。

Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., 和 Zhong, Z. Mixgrpo：通过混合 ODE-SDE 提升基于流的 GRPO 效率。*arXiv* 预印本 *arXiv:2507.21802*，2025。

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

Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Li, J., 和 Zheng, L. 步进感知偏好优化：使偏好与每一步的去噪性能保持一致。*arXiv* 预印本 *arXiv:2406.04314*，2024年。

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. 和 Le, M. 生成建模的流匹配。*arXiv* 预印本 *arXiv:2210.02747*，2022。

Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D. 和 Ouyang, W. Flow-GRPO：通过在线强化学习训练流匹配模型。*arXiv* 预印本 *arXiv:2505.05470*，2025a。

Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C. 等。Step1X-Edit：通用图像编辑的实用框架。*arXiv* 预印本 *arXiv:2504.17761*，2025b。

Liu, X.、Gong, C. 和 Liu, Q. 通过整流流实现流的直化与加速：学习生成和迁移数据。*arXiv* 预印本 *arXiv:2209.03003*，2022年。

Ma, Y.、Liu, X.、Chen, X.、Liu, W.、Wu, C.、Wu, Z.、Pan, Z.、Xie, Z.、Zhang, H.、Yu, X. 等。JanusFlow：协调自回归与整流流，实现统一多模态理解与生成。载于 计算机视觉与模式识别会议论文集，第 7739–7751 页，2025。

Ma˜nas, O.、Astolfi, P.、Hall, M.、Ross, C.、Urbanek, J.、Williams, A.、Agrawal, A.、Romero-Soriano, A. 和 Drozdzal, M. 通过自动提示优化提升文本到图像一致性。*arXiv* 预印本 *arXiv:2403.17804*，2024年。

Mo, W., Zhang, T., Bai, Y., Su, B., Wen, J.-R., and Yang, Q. Dynamicprompt 优化文生图生成。载于 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 26627–26636 页，2024年。

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., M¨uller, J., Penna, J., and Rombach, R. SDXL：改进用于高分辨率图像合成的潜扩散模型。*arXiv* 预印本 *arXiv:2307.01952*，2023。

Prabhudesai, M., Goyal, A., Pathak, D., and Fragkiadaki, K. 通过奖励反向传播对文生图扩散模型进行对齐。*arXiv* 预印本 *arXiv:2310.03739*，2023。

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. 从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议，第 8748–8763 页。PMLR，2021。

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. 和 Ommer, B. 使用潜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。

Rosenman, S., Lal, V. 和 Howard, P. NeuroPrompts：用于优化文本到图像生成提示词的自适应框架。载于欧洲计算语言学协会第 *18* 届会议：系统演示，第 159–167 页，2024年。Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T. 等。具备深度语言理解能力的照片级真实感文本到图像扩散模型。神经信息处理系统大会进展，35：36479–36494，2022年。Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y. 等。DeepSeekMath：突破开放语言模型数学推理能力的极限。*arXiv* 预印本 *arXiv:2402.03300*，2024年。

Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., 和 Taigman, Y. Emu Edit：通过识别和生成任务实现精准图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 8871–8879 页，2024年。

Song, J., Meng, C., 和 Ermon, S. 去噪扩散隐式模型。*arXiv*预印本 *arXiv:2010.02502*，2020年。

Sutton, R. S. 强化学习：导论。*ABradford Book*，2018年。

Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., 和 Naik, N. 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，第 8228–8238 页，2024 年。

Wang, F.-Y., Shui, Y., Piao, J., Sun, K., 和 Li, H. Diffusion-npo：用于生成偏好对齐效果更佳的扩散模型的负偏好优化。*arXiv* 预印本 *arXiv:2505.11245*，2025a。

Wang, X.、Zhang, X.、Luo, Z.、Sun, Q.、Cui, Y.、Wang, J.、Zhang, F.、Wang, Y.、Li, Z.、Yu, Q. 等。Emu3：下一词元预测就是你所需要的一切。*arXiv* 预印本 *arXiv:2409.18869*，2024年。

Wang, Y.、Zang, Y.、Li, H.、Jin, C. 和 Wang, J.。用于多模态理解与生成的统一奖励模型。*arXiv* 预印本 *arXiv:2503.05236*，2025b。

Wei, C.、Xiong, Z.、Ren, W.、Du, X.、Zhang, G. 和 Chen, W.。OmniEdit：通过专家监督构建图像编辑通用模型。载于第十三届国际学习表征会议，2024年。

Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., 等。Qwen-image 技术报告。*arXiv* 预印本 *arXiv:2508.02324*，2025a。

Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., 和 Chen, W. Ed-itreward：一种用于指令引导图像编辑的人类偏好对齐奖励模型。*arXiv* 预印本 *arXiv:2509.26346*，2025b。Wu, M., Wang, L., Zhao, P., Yang, F., Zhang, J., Liu, J., Zhan, Y., Han, W., Sun, H., Ji, J., 等。Re-prompt：通过强化学习实现文本到图像生成的推理增强重新提示。*arXiv* 预印本 *arXiv:2505.17540*，2025c。Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., 和 Li, H. 人类偏好分数 v2：用于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。Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., 等。SANA：采用线性扩散变换器的高效高分辨率图像合成。*arXiv* 预印本 *arXiv:2410.10629*，2024a。Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Q., Gu, Y., Chen, Z., Yang, Z., 和 Shou, M. Z. Show-o：用单一变换器统一多模态理解与生成。*arXiv* 预印本 *arXiv:2408.12528*，2024b。Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., 和 Dong, Y. ImageReward：学习和评估文本到图像生成的人类偏好。神经信息处理系统大会进展，36：15903–15935，2023。Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., 等。DanceGRPO：释放 GRPO 在视觉生成中的潜力。*arXiv* 预印本 *arXiv:2505.07818*，2025。

Yang, K.、Tao, J.、Lyu, J.、Ge, C.、Chen, J.、Shen, W.、Zhu, X.和Li, X. 使用人类反馈微调扩散模型而无需任何奖励模型。载于*IEEE/CVF*计算机视觉与模式识别会议，第8941–8951页，2024年。Yin, F.、Liu, S.、Han, Y.、Wang, Z.、Xing, P.、Wang, R.、Cheng, W.、Wang, Y.、Li, A.、Yin, Z.等。ReasonEdit：迈向推理增强图像编辑模型。*arXiv*预印本 *arXiv:2511.22625*，2025。Zhang, K.、Mo, L.、Chen, W.、Sun, H.和Su, Y. MagicBrush：一个用于指令引导图像编辑的人工标注数据集。神经信息处理系统进展，36:31428–31449，2023。Zhang, Y.、Tzeng, E.、Du, Y.和Kislyuk, D. 面向扩散模型的大规模强化学习。*arXiv*预印本 *arXiv:2401.12244*，2024年。Zheng, K.、Chen, H.、Ye, H.、Wang, H.、Zhang, Q.、Jiang, K.、Su, H.、Ermon, S.、Zhu, J. 和Liu, M.-Y. DiffusionNFT：结合前向过程的在线扩散强化。*arXiv*预印本 *arXiv:2509.16117*，2025。

算法 1 PromptRL：语言模型和流匹配模型的联合强化学习训练

要求： 训练提示 $\mathcal{D}$、语言模型 π_{LM}、流匹配模型 π_{FM}、奖励函数 R、组大小 n、保留数量 m
确保： 优化后的 π_{LM} 和 π_{FM}
1: **对于**每次迭代 **执行**
2: 采样一批提示词 $\{p_0^{(j)}\}_{j=}^{B}$ $_1$ 来自 $\mathcal{D}$
3: **针对**每个提示词 $p_0^{(j)}$ **do**
4: // 生成提示词变体
5: $\mathcal{P}^{(j)} \leftarrow \{p_0^{(j)}\}^m$ // 保留 m原始提示词
6: **对于** $i = m+1$ 来 n **执行**
7: $p_i^{(j)} \sim \pi_{\text{LM}}(\cdot \mid p_0^{(j)})$ // 语言模型精炼
8: $\mathcal{P}^{(j)} \leftarrow \mathcal{P}^{(j)} \cup \{p_i^{(j)}\}$
9: **循环结束**
10: // 生成图像并计算奖励
11: **对**每个 $p_i \in \mathcal{P}^{(j)}$ **执行**
12: 采样 $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I)$
13: $\mathbf{x}_i \leftarrow \pi_{\text{FM}}(\cdot \mid p_i, \boldsymbol{\epsilon}_i)$
14: $r_i \leftarrow R(\mathbf{x}_i, p_i)$
15: **结束循环**
16: // 组内优势归一化
17: $\mu^{(j)}, \sigma^{(j)} \leftarrow$ 均值($\{r_i\}$)，标准差($\{r_i\}$)
18: **对于** $i = 1$到 n **执行**
19: $A_i^{(j)} \leftarrow (r_i - \mu^{(j)})/(\sigma^{(j)} + \epsilon)$
20: **结束循环**
21: **结束循环**
22: // 更新语言模型（仅针对细化后的提示词）
23: $\theta_{\text{LM}} \leftarrow \theta_{\text{LM}} + \alpha_{\text{LM}} \sum_{j,i>m} A_i^{(j)} \nabla$记录 $\pi_{\text{LM}}(p_i^{(j)} \mid p_0^{(j)})$
24: // 更新流匹配模型（对所有样本）
25: $\theta_{\text{FM}} \leftarrow \theta_{\text{FM}} + \alpha_{\text{FM}} \sum_{j,i} A_i^{(j)} \nabla$对数 $\pi_{\text{FM}}(\mathbf{x}_i^{(j)} \mid p_i^{(j)}, \boldsymbol{\epsilon}_i)$
26: **结束循环**

A. 训练详情

伪代码。 算法 1 给出了完整的PromptRL训练流程。该算法交替进行以下操作：通过语言模型生成提示词变体、通过流匹配模型生成图像，以及使用组归一化优势更新两个模型。关键设计选择包括提示词保留机制（第4–9行），该机制为每组保留 m 条原始提示词，以及选择性梯度更新：仅在细化后的提示词上训练语言模型，而流匹配模型则从所有样本中学习。

训练配置。表7 总结了所有实验中使用的超参数和计算资源。为确保公平比较，我们尽可能采用一致的设置，并根据初步实验针对不同任务调整学习率和 KL 系数。对于图像编辑，我们采用更高的分辨率（1024×1024）和更少的 SDE 步数，以平衡质量与效率，详见第4节。

B. 讨论

提示增强器对未见过的流匹配模型的泛化能力。 我们研究了通过 PromptRL 训练的提示增强模块，能否泛化到训练期间未见过的流模型。具体而言，我们使用 GenEval，在 1024 分辨率和 20 个推理步数下评估提示增强器在 SANA（Xie 等人，2024a）和 SD3 上的表现（表 8）。尽管我们的提示增强器不及针对每个流匹配模型分别通过仅提示词强化学习训练的专用提示增强器，但与模型原始表现相比，仍有显著提升。对于 SANA，我们的增强器将 GenEval 得分从 0.62 提升至 0.70；对于 SD3，则从 0.58 提升至 0.77。这些结果表明，PromptRL 学会了可泛化的语言润色，能够惠及多种流模型；这说明语言模型捕捉到的语义模式广泛适用于文本到图像生成，而非过拟合于 FLUX.1-dev 的特定特征。

表 7. PromptRL 在不同奖励目标下的训练配置和超参数。所有实验均使用 GRPO 作为基础强化学习算法。“—”表示不适用或与默认值相同。

配置	文本到图像			多重奖励	图像编辑
	GenEval	OCR	PickScore		EditReward
模型设置					
基础流匹配模型	FLUX.1-dev	FLUX.1-dev	FLUX.1-dev	FLUX.1-dev	FLUX.1-Kontext
语言模型	Qwen2.5-VL-3B-Instruct				
FM 参数	12B				12B
LM 参数	3B				
训练详情					
训练数据集	FlowGRPO-GenEval	FlowGRPO-OCR	Pick-a-Pic	混合	OmniEdit（10k）
训练样本	50,000	19,653	25,432	95,085	10,000
训练分辨率	512×512	512×512	512×512	512×512	1024×1024
训练精度	bfloat16	bfloat16	bfloat16	bfloat16	bfloat16
推理步数（Rollout）	20	20	20	20	8
SDE 步数	20	20	20	10	4
组大小 (n)	8	8	8	8	8
提示词保留 (m)	2	2	2	2	2
批次大小	1	1	1	1	1
K 个训练周期	1	1	1	1	1
学习率（FM）	3×10^{-7}	3×10^{-7}	3×10^{-7}	10^{-7}	2×10^{-7}
学习率（LM）	10^{-6}	10^{-6}	10^{-6}	10^{-6}	4×10^{-7}
优化器	AdamW	AdamW	AdamW	AdamW	AdamW
KL 系数（FM）	4×10^{-3}	4×10^{-3}	2×10^{-3}	2×10^{-3}	10^{-2}
KL 系数（LM）	10^{-2}	10^{-2}	10^{-2}	10^{-2}	10^{-2}
奖励配置					
奖励模型	GenEval	OCR准确率	PickScore	基于标签	EditReward
格式奖励 (λ_{Format})	1.0	1.0	1.0	1.0	1.0
生成奖励 (λ_{Gen})	1.0	1.0	1.0	1.0	1.0
奖励归一化	按组	按组	按组	按标签	按组
训练成本					
GPU数量	8	8	8	8	8
GPU类型	H100	H100	H100	H100	H100
训练 Rollout	0.2M	0.05M	0.13M	0.5M	0.06M

与纯流强化学习的比较。一个自然的问题是：仅仅扩大纯流强化学习的训练规模，能否获得与 PromptRL 相当的提升？为探究这一点，我们训练了一个 rollout 数量翻倍的纯流基线，并在表 9 中将其与 PromptRL 进行比较。尽管具有 $2\times$ 计算优势，纯流强化学习在所有指标上的表现仍然不及 PromptRL（GenEval：0.93 对比 0.97，光学字符识别：0.93 对比 0.98，PickScore：23.85 对比 24.05）。这一结果表明，纯流强化学习的局限——即第 3 节指出的探索崩溃和提示词过拟合——无法仅靠增加 rollout 数量来克服。联合 LM-FM 优化通过持续.....从根本上重塑了优化景观

表 8. PromptRL 的提示增强器（PE）在 GenEval 上对未见过的流模型的泛化能力。所有评估均在 1024×1024 分辨率下进行，推理步数为 20。我们比较了不使用 PE 的官方预训练权重、使用我们的 PE 的权重，以及使用针对特定模型进行强化学习调优的 PE 的权重。两个 PE 模块均使用 Qwen-2.5-VL-Instruct-3B。我们学习得到的 PE 展现出强大的泛化能力，能够有效提升训练期间未见过的预训练模型。

模型	原始版本	我们的 PE	仅提示词强化学习
SANA	0.62	0.70	0.76
SD3	0.58	0.77	0.83

表9. PromptRL 与纯流强化学习的比较。纯流强化学习使用 2× 数量的 rollout 进行训练，以控制计算预算。

方法	Rollout 数量	GenEval↑	OCR↑	PickScore↑
纯流强化学习	2×	0.93	0.93	23.85
PromptRL	1×	0.97	0.98	24.05

注入语言多样性　　样性，使流匹配模型能够摆脱狭窄的奖励模式，并实现更高的性能上限　　s.

局限性。 尽管 PromptRL 通过联合 LM-FM 优化取得了强劲性能，但流匹配模型和语言模型在训练过程中会形成一定程度的协同适应。具体而言，在推理时将我们共同训练的提示增强器替换为其他语言模型（例如，Qwen-3）后，我们观察到其在 GenEval 上的性能从 0.97 降至 0.88。这表明，流匹配模型在一定程度上专门适应了其训练阶段的语言模型搭档所生成的语言模式。幸运的是，这种协同适应是有意设计的结果，而非根本性缺陷——PromptRL 的明确目标是对两个组件进行联合优化，以便将其部署为统一系统。在共同训练的语言模型和流匹配模型协同使用的实际场景中，这种紧密耦合带来了我们所报告的最先进性能。未来的工作可以探索多语言模型训练或正则化策略等技术，以便在需要更广泛兼容性时进一步提升跨语言模型泛化能力。

为什么 JointRL 很重要? 随着文本到图像模型变得更加强大，它们对提示词的表述方式也变得更加敏感——细微的语言变化就可能导致生成质量的显著差异。这使得 PE模块与流匹配模型本身同等重要。然而，现有方法通常分别训练这两个组件，要么通过独立的监督微调，要么通过强化学习。我们认为，这种孤立的方法忽略了一个关键点：细化提示词的最佳方式取决于当前的流匹配模型，反之亦然。通过将提示增强和流匹配模型视为一个统一系统，PromptRL 使二者能够共同改进——语言模型发现能带来更高奖励的提示词变体，而扩展的提示词多样性则提供更丰富的探索信号，从而加速流匹配模型的训练效率。