arXiv:2603.01163v1 [cs.CV] 1 Mar 2026

BeautyGRPO：通过动态路径引导和细粒度偏好建模实现人脸修饰的审美对齐

Jiachen Yang[1] XianhuiLin[2†] Yi Dong[2] Zebiao Zheng[2]Xing Liu[2] Hong Gu[2] YanmeiFang[1,3*]

[1]中山大学深圳校区网络空间安全学院，中国[2]vivo BlueImageLab，维沃移动通信有限公司，中国[3]广东省信息安全技术重点实验室，中国项目主页

[†]项目负责人 *通讯作者

摘要

人脸修饰需要去除细微瑕疵，同时保留独特的面部身份特征，以提升整体美学吸引力。然而，现有方法面临根本性的权衡。基于标注数据的监督学习受限于像素级标签模仿，无法捕捉复杂、主观的人类审美偏好。相反，尽管在线强化学习（*RL*）擅长偏好对齐，其随机探索范式却与人脸修饰对高保真度的要求相冲突，并且由于随机漂移不断累积，往往会引入明显的噪声伪影。为解决这些局限，我们提出 *BeautyGRPO*，这是一种使人脸修饰与人类审美偏好对齐的强化学习框架。我们构建了 *FRPref-10K*，这是一个涵盖五个关键修饰维度的细粒度偏好数据集，并训练了一个能够评估细微感知差异的专用奖励模型。为兼顾探索与保真度，我们提出动态路径引导（*DPG*）。*DPG*通过动态计算基于锚点的常微分方程（*ODE*）路径，并在每个采样时间步重新规划引导轨迹，从而稳定随机采样轨迹，有效纠正随机漂移，同时保持受控探索。大量实验表明，*BeautyGRPO*优于专用人脸修饰方法和通用图像编辑模型，在纹理质量、瑕疵去除准确性方面表现更佳，整体效果也更符合人类审美偏好。

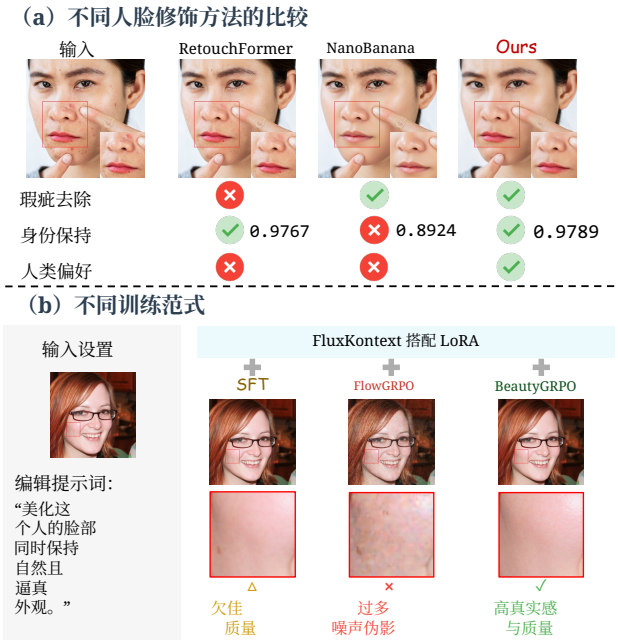


图 1. 人脸修饰方法与训练范式的比较。(a) 在瑕疵去除、身份保持和人类偏好方面，我们的方法优于专用模型（Retouch- Former）和通用模型（NanoBanana）。(b) 通过稳定随机探索，BeautyGRPO 克服了监督微调（SFT）和标准在线强化学习（Flow- GRPO）的局限性，取得了自然且高度逼真的效果。

1. 引言

随着社交媒体和成像设备的迅速普及，数字人像已成为日常不可或缺的一部分

生活，并推动了对高质量人脸修饰日益增长的需求。该任务本质上十分复杂，因为它需要去除痤疮和瑕疵等细微缺陷，同时保留痣、毛孔和自然皮肤纹理等身份特征，以提升整体美学吸引力。本质上，人脸修饰必须在对原始肖像的高视觉保真度与主观人类审美偏好的一致性之间取得平衡。

为应对这一挑战，研究人员探索了多种专用架构，范围从基于卷积神经网络的架构 [20, 37]到先进的基于Transformer的模型[44–46, 54, 55]。近年来，通用图像编辑模型[11, 24, 30, 35, 59] 也展现出在人脸修饰方面的巨大潜力。然而，现有的大多数方法仍依赖监督学习，要么从头开始训练，要么进行监督微调（SFT）。这类监督目标的局限不仅在于精心整理的数据稀缺，还在于其像素级学习目标。它们训练模型去模仿带标签的参考图像，而非与人类审美偏好对齐。这一目标强调局部保真度，却忽视感知对齐，导致生成结果虽然忠实于标签，却仍与人类的审美期望不符。因此，输出往往呈现出视觉上僵硬或不自然的修饰效果，如图1所示。此外，由于将优化限制在像素级重建上，监督学习会促使模型过拟合特定的修图风格，并阻碍模型探索超越训练数据审美质量的方案。这一观察引出了一个根本问题：**我们能否使人脸修饰与人类偏好对齐?**

强化学习（RL）近期已成为一种颇具前景的框架，可用于使图像生成与人类偏好保持一致 [4, 7, 10, 13, 32–34, 40, 41, 56, 58, 60]。在文生图（T2I）生成中，FlowGRPO 等方法[22] 通过将在线强化学习融入流匹配模型，展现出强大的可控性 [9, 18, 21, 25]。这种在线强化学习机制支持主动探索，使模型能够发现监督训练期间未曾见过的方案，并更好地契合人类判断。尽管取得了这些成功，将文生图强化学习框架直接应用于人脸修饰仍面临两项根本挑战。第一项挑战是保真度与探索之间的冲突。为鼓励探索，FlowGRPO 会在采样过程中注入随机噪声，将确定性常微分方程轨迹转变为随机微分方程（SDE）。虽然这种随机性提升了文生图生成的多样性，但对人脸修饰却有害，因为人脸修饰对视觉稳定性和精度要求极高。如图1所示，累积噪声会使轨迹偏离高保真流形，并产生明显的噪声伪影。第二项挑战在于奖励监督的粒度。现有的文生图生成或通用图像编辑 [26, 43, 48, 49, 51, 52]主要关注整体美学或指令遵循，缺乏人脸修饰所需的细粒度感知能力；人脸修饰依赖于对皮肤平滑、瑕疵去除、纹理保留和身份一致性的准确评估。

本文提出 BeautyGRPO，一种将人脸修饰与人类审美偏好对齐的强化学习框架。该领域的一个关键挑战是缺乏能够评估细粒度美化质量的奖励模型。为此，我们构建了 FRPref-10K，一个大规模**人**脸**修**饰**偏**好数据集，涵盖五个关键维度：皮肤平滑、瑕疵去除、纹理质量、清晰度和身份保持。每张图像均由多个视觉语言模型（视觉语言模型）通过基于推理的评分进行评估，随后经过人工优化，以确保与审美判断一致。基于该数据集，我们训练了一个奖励模型，用于捕捉美化图像对之间细微的感知差异，并反映人类偏好。

为解决保真度与探索之间的关键冲突，我们提出动态路径引导（DPG），即 BeautyGRPO 的核心算法。DPG 通过将随机轨迹柔和地引导至理想的确定性路径，同时保留受控随机性，从而稳定该轨迹。与监督微调不同——监督微调通过像素级监督将模型输出与标签图像显式对齐——DPG 仅在采样过程中将标签图像作为稳定性锚点纳入。这些锚点是从我们的 FRPref-10K 数据集中选出的高偏好示例，用于将轨迹正则化到高保真流形附近，而非充当成对的真实参考图像。这些锚点提供温和的方向性稳定，使采样轨迹保持在高保真流形附近，同时不干扰由奖励驱动的优化。该设计允许策略在更高奖励表明结果更具吸引力时超越锚点，从而在保持轨迹稳定性的同时，保留在线强化学习的探索特性。通过这种既受锚点约束又由奖励驱动的机制，DPG 在探索与保真度之间建立了平衡状态，有效防止轨迹漂移，并能够发现审美质量超越参考数据的修饰结果。

我们的贡献总结如下：1. 我们构建了大规模、细粒度的人脸修饰偏好数据集 FRPref-10K，并训练了一个专用奖励模型，能够高精度地捕捉多维人类审美偏好。2. 我们提出了 BeautyGRPO，这是一种配备动态路径引导的强化学习框架，能够在修饰过程中有效平衡探索与保真度。3. 大量实验证明，所提出的奖-励模型与人类审美判断高度一致，而 BeautyGRPO 在用户偏好方面优于现有方-

法，尤其在皮肤平滑、瑕疵去除和纹理保留方面表现突出。

2. 相关工作

2.1. 人脸修饰

人脸修饰旨在去除面部瑕疵、提升面部美学，同时保留面部身份特征。早期研究从传统滤波[1, 3, 39] 和修复方法 [23, 44, 45] 发展到基于卷积神经网络的模型 [20, 50, 53]，近来又发展到基于 Transformer 的架构 [46, 54, 55]。这些专用修图模型大多仍依赖监督学习。其局限在于，模型的目标是让输出在像素层面与标签图像对齐，而非符合人类审美偏好。此外，大规模通用图像编辑模型 [11, 24, 30, 47, 59]能够通过指令驱动编辑执行复杂的修饰操作，但往往会引入不自然的视觉效果，例如改变面部身份、使皮肤纹理显得油腻或像塑料，以及产生虚假的视觉伪影，导致结果不一致且不够真实。

2.2. 扩散模型强化学习

强化学习已成为使扩散模型和流匹配模型与人类偏好对齐的有效方法。DDPO[4] 和DPOK [10] 等策略梯度方法通过优化偏好分数来实现对齐，但扩散模型中的梯度路径较长，导致更新仅限于少数靠后的时间步。DiffusionDPO [40] 将DPO [34] 扩展到离线设置下的文生图，利用成对的正样本和负样本进行训练。这种方法能够实现稳定优化，但仍受静态数据质量和多样性的限制，且难以超越训练分布。FlowGRPO [22] 等在线强化学习方法通过将确定性常微分方程转换为随机微分方程，使GRPO [12]适用于流匹配模型，从而实现主动探索。将其用于人脸修饰时，基于随机微分方程的探索会引入随机漂移，违反高保真约束并产生残留伪影。尽管这种随机性有助于提升文生图的多样性，却与人像增强所要求的精确性和稳定性相冲突。

2.3. 人类偏好奖励建模

奖励模型是强化学习的核心，作为优化信号，使模型行为与人类偏好保持一致。近期，文本到图像生成和图像编辑领域的进展通过在大规模人类偏好数据集上训练奖励模型，实现了这种对齐[17, 49]。代表性模型包括用于文本到图像生成的 ImageReward [51],HPSv2 [49], 和 UnifiedReward [43]，以及用于图像编辑的 EditScore [26], 、EditReward [48], 和 UnifiedReward-Edit [43] 。尽管这些奖励模型能够使生成结果符合人类审美判断，但它们主要关注全局图像质量和文本-图像一致性。它们无法捕捉细微的感知维度，例如皮肤平滑度与纹理自然度之间的平衡，以及瑕疵去除的洁净程度。为弥补这一差距，我们构建了专门的人脸修饰偏好数据集，并训练了一个奖励模型，以提供符合人类审美偏好的细粒度奖励指导。

3. 方法论

我们提出了 BeautyGRPO，这是一种强化学习框架，可使人脸修饰模型与人类审美偏好对齐。该框架包含三个组成部分：（1）大规模、多维偏好数据集（第3.1节）；（2）用于细粒度审美评估的专用奖励模型（第3.2节）；以及（3）动态路径引导（DPG）算法，通过协调随机探索与确定性保真度来稳定在线强化学习（第3.3节）。

3.1. 细粒度偏好数据整理

审美对齐需要能够反映人脸修饰中细微判断的数据。为克服现有粗粒度数据集的局限性，我们构建了 FRPref-10K，这是一个包含 10,000 对高分辨率偏好对、并带有详细多维标注的细粒度数据集。

数据收集与候选项生成。我们从 FFHQR [36]和一个专有高分辨率数据集中筛选源肖像，确保涵盖多样的人口统计特征和成像条件。对于每张输入图像 I_{in}，我们使用多个先进模型（NanoBanana $\{I^i_{out}\}$ Flux.1 Kon- text [11], RetouchFormer [19], ），并采用不同的随机种子，生成多样化的修饰候选图像 [46]。随后，通过模型间比较（输出与输出）以及输出与标签的评估来构建偏好对 (I_{in}， I_A， I_B)。关于数据集统计信息的更多细节，请参阅补充材料 E。

混合视觉语言模型-人工标注流程。 我们采用三阶段标注协议，将大型视觉语言模型的推理能力与人工验证相结合:

1. **多维度评估：** 我们定义了五个关键美学维度：皮肤平滑、瑕疵去除、纹理质量、清晰度和身份保持。由多个视觉语言模型（GPT-4o [31], Qwen2.5-VL-72B [2], Gemini 2.5 Pro [8]）组成的集成模型会从这些维度评估每一对样本，并提供量化评分和结构化推理。我们汇总视觉语言模型的输出，形成初步共识。随后，经过培训的人工标注员依据相同标准审核这些评估结果。

2. **专家裁决：** 当人类判断与 VLM 共识不一致时，将升级交由资深专家进行最终裁决。

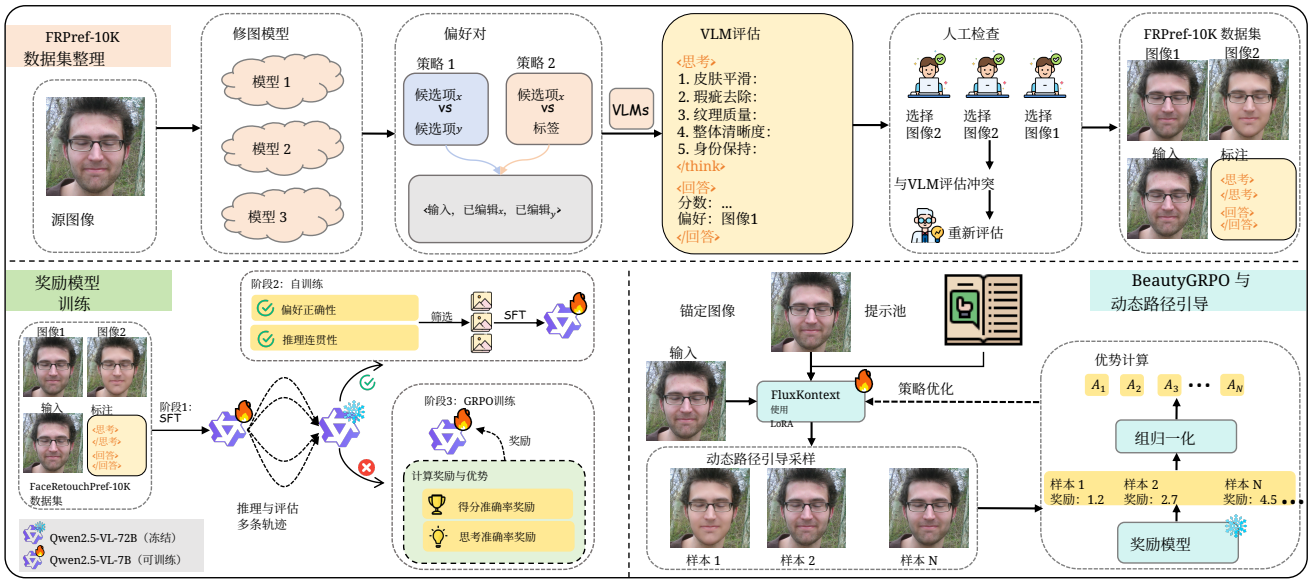


图 2. 我们的 FRPref-10K 构建、奖励模型训练以及采用动态路径引导（DPG）的 BeautyGRPO 概览。**顶部：** FRPref-10K 数据集整理流程。使用多种编辑模型生成不同的修饰候选图像，通过输出对输出/标签比较构建偏好对，并由视觉语言模型从五个质量维度进行标注，之后再经人工验证。**左下：** 三阶段奖励模型训练，包括监督微调、自训练与一致性筛选，以及 GRPO。**右下：** 基于 FluxKontext-LoRA 主干模型，采用 DPG 进行 BeautyGRPO 训练。

3.2. 多维度修图奖励模型

利用 FRPref-10K，我们训练了一个专用的奖励模型$R_\phi(I_{in}, I_{out})$，能够提供密集且可解释的反馈。遵循 UnifiedReward-Thinking 范式 [42]，我们采用三阶段训练策略来提升推理质量，并从 Qwen2.5-VL-7B 初始化 [2]。

阶段1：结构化推理初始化 我们在一个包含 2K 样本的子集上进行监督微调，使模型能够在 <think> 块中针对五个维度生成结构化推理，并在 <answer> 块中给出最终偏好决策。

阶段2：通过一致性筛选进行自训练 阶段1模型为剩余的8K个样本生成多条推理轨迹。我们采用基于两个标准的一致性筛选机制：偏好正确性（与真实标注的一致性）和推理连贯性（推理与决策之间的语义一致性，由高容量VLM评估）。一致的轨迹将用于后续一轮监督微调，以强化可靠的推理模式。

阶段 3：通过 GRPO 提升鲁棒性 对于不一致的样本，我们采用 GRPO [12] 来提升鲁棒性。模型探索多样的推理路径，并通过结果奖励（偏好准确率）和过程奖励（推理连贯性，由 DeBERTa-V3 [14] 等轻量级验证器评估）进行评估。这一强化阶段增强了模型生成稳定且符合人类判断的审美判断的能力。

3.3. 使用动态路径引导的BeautyGRPO

我们使用训练好的奖励模型作为反馈，对流匹配图像编辑策略模型 π_θ进行对齐。我们在 FlowGRPO 框架 [22]的基础上扩展出一种新颖的动态路径引导（DPG）机制，以稳定高保真修图任务中的基于 SDE 的探索。

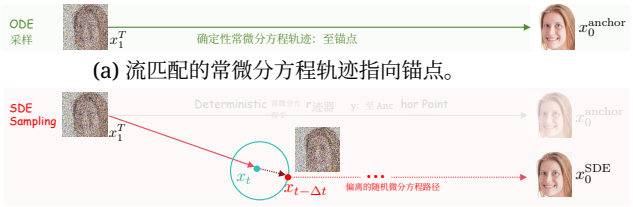
3.3.1. 预备知识：保真度与探索之间的冲突

流匹配模型 [21, 25] 学习一个向量场 $v_\theta(x_t, t)$，通过确定性常微分方程轨迹将噪声分布 $p_1(x)$传输至数据分布 $p_0(x)$。为支持在线强化学习中的探索，FlowGRPO [22]通过注入噪声将该常微分方程转换为随机微分方程形式：

$$dx_t = \left[v_\theta(x_t, t) + \frac{\sigma_t^2}{2t} \left(x_t + (1-t) v_\theta(x_t, t) \right) \right] dt + \sigma_t d\omega, \tag{1}$$

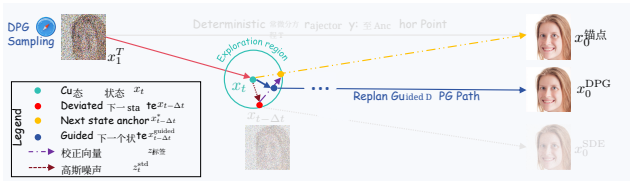
其中，$d\omega$ 表示标准布朗运动，$\sigma_t = \eta \sqrt{\frac{t}{1-t}}$ 控制噪点水平。该随机项会产生多样化的采样轨迹，使策略能够获得更丰富、区分度更高的奖励信号。具体而言，给定G 生成的样本 $\{x_0^i\}_{i=1}^G$ 及其对应的奖励$R(x_0^i)$，FlowGRPO 会将这些奖励归一化为标准化优势：

$$\hat{A}^i = \frac{R(x_0^i) - \text{mean}(\{R(x_0^i)\}_{i=1}^G)}{\text{std}(\{R(x_0^i)\}_{i=1}^G)}. \tag{2}$$



(a) 流匹配的常微分方程轨迹指向锚点。

(b) FlowGRPO的随机微分方程采样轨迹。



(c) 使用DPG采样的BeautyGRPO轨迹。

图3. 不同采样轨迹概览。(a) 用于监督微调的标准流匹配常微分方程轨迹。(b) FlowGRPO中不受控的随机微分方程轨迹，它逐渐偏离高保真锚点，并引入噪声伪影。(c) 提出的采用动态路径引导（DPG）的BeautyGRPO。在每个时间步，DPG都会动态计算一条基于锚点的常微分方程路径，并通过将校正向量与标准高斯噪声进行线性混合，重新规划一条引导轨迹，从而校正随机漂移，同时保持受控探索。

策略通过GRPO目标函数进行优化：

$$\mathcal{L}(\theta) = \mathbb{E}_{c\sim\mathcal{C},\{x^i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\min\left(r_t^i(\theta)\hat{A}_t^i,\right.\right.$$
$$\left.\left.\text{clip}(r_t^i(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t^i\right)\right], \tag{3}$$

其中，$r_t^i(\theta) = \frac{p_\theta(x_{t-1}^i|x_t^i)}{p_{\theta_{\text{old}}}(x_{t-1}^i|x_t^i)}$ 表示旧策略模型 $\pi_{\theta_{\text{old}}}$ 与更新后的模型 π_θ之间逐步的似然比。

尽管随机项 $\sigma_t\sqrt{dt}\epsilon$ 在公式 1 中能够增强一般文本到图像生成的探索能力，但它与人脸修饰对高保真度的要求相冲突。累积噪点会逐渐将采样轨迹推离高保真流形，导致不受控的漂移和明显的视觉伪影。

3.3.2. 锚定探索的动态路径引导（DPG）

为协调探索与保真度之间的冲突，我们提出了动态路径引导（DPG）。DPG 通过稳定高保真流形附近的随机微分方程轨迹来进行锚定探索，同时允许奖励信号引导轨迹偏离。

稳定性锚点。我们使用 FRPref-10K 数据集中高偏好样本作为稳定性锚点，记作 by x_0^{anchor}。锚点仅在采样过程中引入，不像监督微调（SFT）那样被视为监督目标。它的作用是将随机轨迹约束在一个高保真度流形内，同时策略仍可在其中探索并优化，以获得奖励更高的修饰结果。

随机微分方程公式与漂移。FlowGRPO 的随机微分方程更新（式1）可简洁地表示为：

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}}\, z_t, \qquad z_t \sim \mathcal{N}(0, I), \tag{4}$$

其中 μ_t 是模型预测的漂移项：

$$\mu_t = x_t + \left[v_\theta(x_t,t) + \frac{\sigma_t^2}{2t}\big(x_t + (1-t)v_\theta(x_t,t)\big)\right]\Delta t, \tag{5}$$

而 $\sigma_{step} = \sigma_t\sqrt{\Delta t}$ 是逐步标准差。如图 3b所示，重复应用此更新会逐渐累积随机漂移，推动轨迹偏离高保真流形。

动态轨迹重规划。为抵消漂移累积，DPG 会在每个时间步动态重规划一条朝向锚点的引导轨迹。如图 3c所示，给定当前状态 x_t，我们重规划一条连接 x_t 与锚点 x_0^{anchor}的直线常微分方程轨迹。该锚点引导路径上下一时间步 $s = t - \Delta t$ 对应的目标状态为：

$$x_{t-\Delta t}^* = \left(\frac{\Delta t}{t}\right)x_0^{\text{anchor}} + \left(1 - \frac{\Delta t}{t}\right)x_t. \tag{6}$$

受控随机性引导校正。我们通过引入校正向量 z_t^{anchor}，将随机微分方程更新引导至 $x_{t-\Delta t}^*$，使预测均值 μ_t 与该锚点引导目标对齐：

$$z_t^{\text{anchor}} = \frac{x_{t-\Delta t}^* - \mu_t}{\sigma_{\text{step}}}. \tag{7}$$

该向量指示将轨迹拉回锚点引导路径的方向。

然而，严格遵循 z_t^{anchor} 会使探索退化为确定性模仿，从而阻碍强化学习超越锚点的质量。为避免这种情况，我们将校正向量与标准高斯噪声 $z_t^{std} \sim \mathcal{N}$ 线性组合，构造混合噪声向量 z_t^{mix}（$0, I$）：

$$z_t^{\text{mix}} = \lambda(t)\, z_t^{\text{anchor}} + \big(1 - \lambda(t)\big)\, z_t^{\text{std}}, \tag{8}$$

其中 $\lambda(t) = \frac{t}{\max(1,T-1)}$ 是沿反向时间轨迹递减的时间相关系数。因此，较早的反向时间步采用更强的锚点引导以校正结构偏差，而较晚的时间步则更多依赖随机噪声以促进细粒度探索。

代入 z_t^{mix}式4中的随机微分方程更新规则，可得 DPG 公式：

$$\begin{aligned} x_{t-\Delta t}^{guided} &= \mu_t + \lambda\,\sigma_{\text{step}}\, z_t^{\text{anchor}} + (1-\lambda)\,\sigma_{\text{step}}\, z_t^{\text{std}} \\ &= (1-\lambda)\,\mu_t + \lambda\, x_{t-\Delta t}^* + (1-\lambda)\,\sigma_{\text{step}}\, z_t^{\text{std}}. \end{aligned} \tag{9}$$

表1. 在FFHQR和真实场景数据集上对最先进方法进行定量评估，利用多种无参考指标评估感知质量和美学吸引力。我们的BeautyGRPO在几乎所有基准测试中均显著优于专用修图模型、通用编辑基线方法以及其他强化学习策略。这些结果验证了我们的偏好对齐框架在实现高质量、自然修图效果方面的有效性，且能很好地泛化到多样化场景。最佳结果以**粗体**标出。

	FFHQR								真实场景						
	自然图像质量评估器↓	NIMA↑	MUSIQ↑	MANIQA↑	NRQM↑	TOPIQ↑	FID↓	ArcFace↑	自然图像质量评估器↓	NIMA↑	MUSIQ↑	MANIQA↑	NRQM↑	TOPIQ↑	ArcFace↑
修图模型															
ABPN[20]	11.152	4.861	4.599	1.027	8.219	0.611	4.713	0.971	11.979	5.005	4.651	0.997	7.019	0.603	0.968
RestoreFormer[44]	12.281	4.712	4.361	0.958	7.411	0.587	15.898	0.799	12.801	4.819	4.534	0.978	7.372	0.601	0.761
RestoreFormer++[45]	12.144	4.618	4.273	0.935	7.415	0.572	15.311	0.907	12.624	4.713	4.386	0.959	7.307	0.594	0.890
VRetouchEr[54]	12.948	4.586	4.169	0.962	6.167	0.551	2.229	0.978	13.605	4.658	4.308	0.957	5.635	0.565	0.974
RetouchFormer[46]	11.153	4.723	4.465	1.036	8.175	0.605	2.275	**0.986**	12.996	4.712	4.501	0.999	6.753	0.593	**0.976**
通用编辑模型															
ICEdit [59]	12.502	4.645	4.192	0.978	7.424	0.555	2.818	0.967	13.016	4.175	4.371	0.967	6.610	0.567	0.949
SeedDream4.0[35]	15.033	4.804	4.503	0.993	7.538	0.571	12.314	0.893	16.812	4.979	4.638	0.966	6.371	0.573	0.898
NanoBanana[11]	11.301	4.919	4.681	1.009	8.065	0.621	6.804	0.889	12.043	5.131	4.783	0.989	6.935	0.632	0.881
Flux.K w/LoRA[19]	12.913	4.694	4.459	1.035	8.009	0.601	**1.735**	0.973	12.501	4.751	4.524	0.978	6.468	0.586	0.968
强化学习模型															
Flux.K+使用 FlowGRPO 的 LoRA[22]	15.024	4.573	4.271	0.935	7.132	0.571	15.657	0.882	13.513	4.687	4.391	0.958	5.844	0.572	0.879
Flux.K+LoRA（使用我们的方法）	**10.831**	**5.123**	**4.906**	**1.079**	**8.401**	**0.676**	4.054	0.952	**11.821**	**5.357**	**4.982**	**1.052**	**7.526**	**0.674**	0.944

这稳定了在线探索，并使策略能够发现比锚点本身获得更高奖励且美学效果更好的修饰结果。

策略优化的对数概率。对于 GRPO 优化（式3），我们需要引导更新规则下的转移概率$p_\theta(x_{t-\Delta t}|x_t)$。根据式9，该条件分布服从高斯分布 $\mathcal{N}(\mu_{\text{new}},\ \sigma_{\text{new}}^2 I)$，其中 $\mu_{\text{new}} = (1-\lambda)\,\mu_t + \lambda\, x^*_{t-\Delta t}$ 且 $\sigma_{\text{new}} = (1-\lambda)\sigma_{\text{step}}$。相应的对数似然为：

$$\log p_\theta(x_{t-\Delta t} \mid x_t) = -\frac{\|x_{t-\Delta t} - \mu_{\text{new}}\|_2^2}{2\,\sigma_{\text{new}}^2} - C, \qquad (10)$$

其中 C 表示常数归一化项。这些逐步概率用于计算策略更新的似然比$r_t(\theta)$。为减少计算量，我们将完整的采样轨迹划分为 $K = 3$ 个片段，在每个片段中随机选取一个时间步应用 DPG 更新，并对其余时间步执行常微分方程更新。

4. 实验

本节通过定量、定性和用户研究比较，将 BeautyGRPO 与最先进的方法进行对比验证，随后对奖励模型和动态路径引导（DPG）机制开展消融研究。受篇幅限制，更多实验细节、设置和结果见补充材料。

4.1. 实验设置

训练设置。对于所有开源基线方法，我们使用由 FRPref-10K 数据集构建的编辑三元组 ⟨prompt、输入、编辑结果⟩ 进行 LoRA [15] 训练。主实验中，我们采用 FluxKontext [19]作为骨干模型，并在这些三元组上进行相同的 LoRA 微调。从经过 LoRA 适配的骨干模型出发，我们分别使用 FlowGRPO 和 BeautyGRPO 进行强化学习，所用三元组相同以保证两者一致。两种情况下，均使用修饰奖励模型预测的奖励来优化式3 中的 GRPO目标函数。测试集包含来自 FFHQR 的 1,000 张肖像，以及从互联网收集的 1,000 张真实场景图像。所有实验均在 8 张 H20 GPU 上运行。

评估指标。人脸修饰具有内在的主观性，因此我们采用无参考感知指标和审美指标来近似人类判断。我们使用 NIQE [29] 和 NRQM [27] 评估整体自然度，使用 NIMA（AVA） [38], 、MUSIQ（AVA） [16], 、MANIQA [57],和 TOPIQ（GFIQA） [6] 衡量审美质量，并使用 ArcFace 分数评估面部质量和身份保持效果。此外，我们在 FFHQR 上报告 FID，以量化分布真实感。有关指标选择的详细依据和具体评估细节，请参见补充材料D.1。

4.2. 主要结果

4.2.1. 定量结果

表1 报告了在 FFHQR 和真实场景测试集上的定量比较。对于人脸修饰，结果的质量最终取决于比较输入图像与修饰后图像时人类的审美感知，而非与有限参考标签集的像素级匹配。PSNR、SSIM 和学习感知图像块相似度等全参考指标主要衡量图像与这些参考图像之间的像素保真度，因此无法可靠地反映感知到的自然度或视觉吸引力。因此，我们重点关注无参考感知指标和美学指标，这些指标能更好地捕捉自然度、真实感和审美质量。如表1所示，BeautyGRPO 在两个数据集上的表现均优于所有基线方法。在 FFHQR 上，它在全部六项无参考指标上都取得了最佳成绩，在 NIMA、MUSIQ、MANIQA、NRQM 和 TOPIQ 上均有明显提升，且 NIQE 显著更低。此外，我们还在一个包含 1000 张互联网人像图像的真实场景数据集上评估了本方法。在该数据集上取得的类似改进表明，这些优势能够泛化到经过整理的数据集之外。尽管 BeautyGRPO 的 FID 略高于表现最佳的有监督基线，但这是预期之中的。通过显式优化人类偏好，BeautyGRPO 探索了更广泛的合理修图风格，并突破了 FID 所依赖的特定数据集标签分布，从而发现比训练标签所呈现的结果更符合人类审美判断的方案。虽然提升审美效果不可避免地需要调整纹理和结构，进而对身份相似度造成轻微影响，但本方法仍保持较高的 ArcFace 得分（FFHQR 上为 0.952，真实场景下为 0.944），确保了稳健的身份保持。

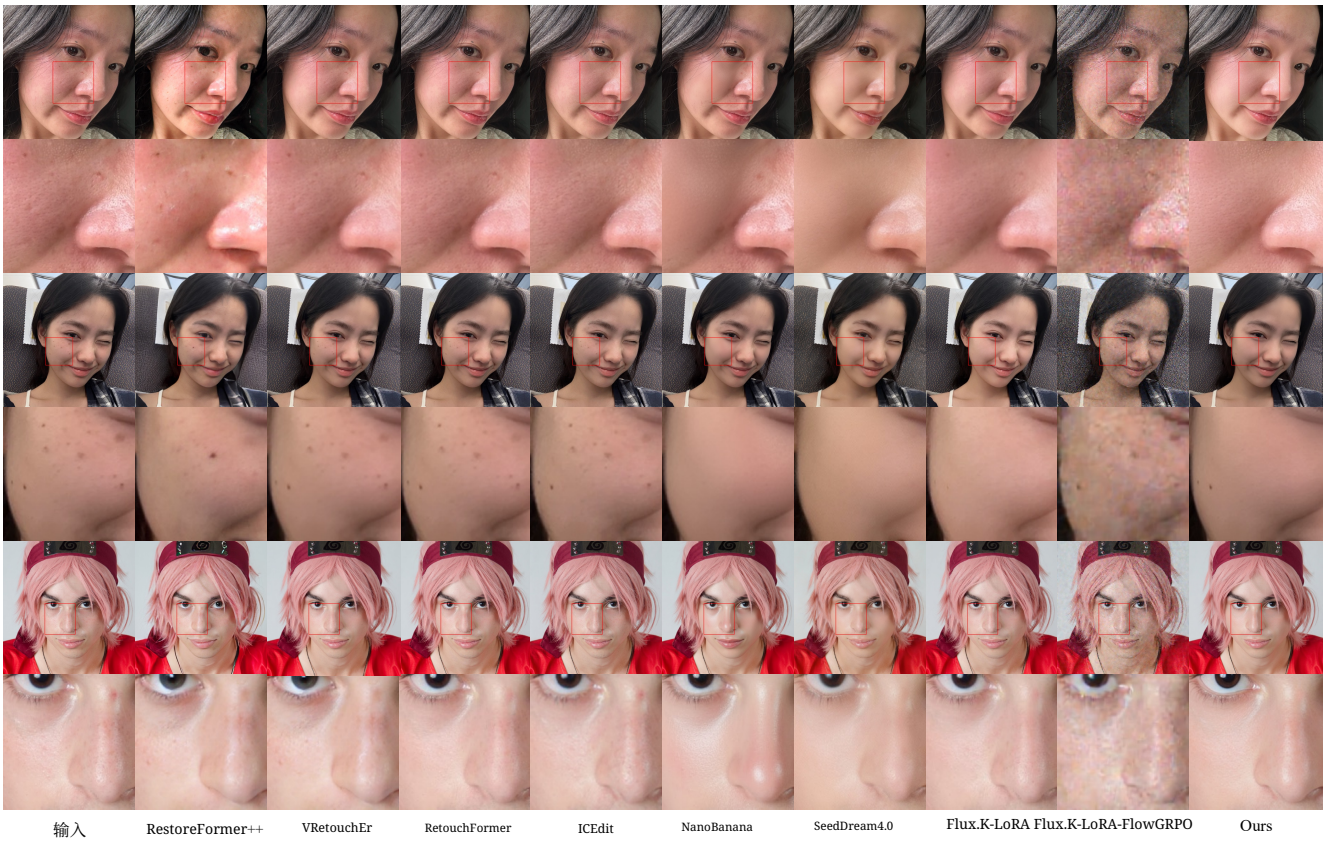


图4. 不同方法在人脸修饰方面于 FFHQR 和真实场景数据集上的结果对比，其中 Flux.K 表示 FluxKontext。与现有模型会出现瑕疵去除不彻底、过度平滑或外观不自然的问题不同，我们的 BeautyGRPO 能够干净地去除瑕疵，同时保留自然纹理、皮肤光泽和痣。更多结果请参阅补充材料。

这些优势能够泛化到经过整理的数据集之外。尽管 BeautyGRPO 的 FID 略高于表现最佳的有监督基线，但这是预期之中的。通过显式优化人类偏好，BeautyGRPO 探索了更广泛的合理修图风格，并突破了 FID 所依赖的特定数据集标签分布，从而发现比训练标签所呈现的结果更符合人类审美判断的方案。虽然提升审美效果不可避免地需要调整纹理和结构，进而对身份相似度造成轻微影响，但本方法仍保持较高的 ArcFace 得分（FFHQR 上为 0.952，真实场景下为 0.944），确保了稳健的身份保持。

4.2.2. 可视化结果

图 4 展示了不同方法在人脸修饰结果上的视觉对比。RetouchFormer 和 VRetouchEr 等专用修图模型通常会出现瑕疵检测不准确、瑕疵去除不彻底，以及过度平滑导致皮肤纹理被抹除等问题。NanoBanana 和 SeedDream4.0 等通用编辑模型则容易过度修饰面部，将痣误判为缺陷，并产

表2. 各种人脸修饰方法的用户偏好胜率对比。

	VRetouchEr	RetouchFormer	NanoBanana	Kontext LoRA	我们的方法
胜率	6.50	8.50	9.75	12.00	**63.25**

生不自然、光泽过强的塑料感外观。相比之下，我们的 BeautyGRPO 能干净地去除面部瑕疵，同时保留皮肤纹理、毛孔、自然光泽和均衡的肤色，并保留痣等独特特征，从而呈现自然且精致的效果，同时具有出色的身份一致性。

4.3. 用户研究

我们邀请100名不同年龄和编辑技能水平的参与者，对20个随机选取的测试案例进行用户研究。研究包含两份问卷。在问卷I中，每道题展示一张原始肖像和采用不同方法得到的五张修饰结果，参与者选择自己最偏好的图像。在问卷II中，每

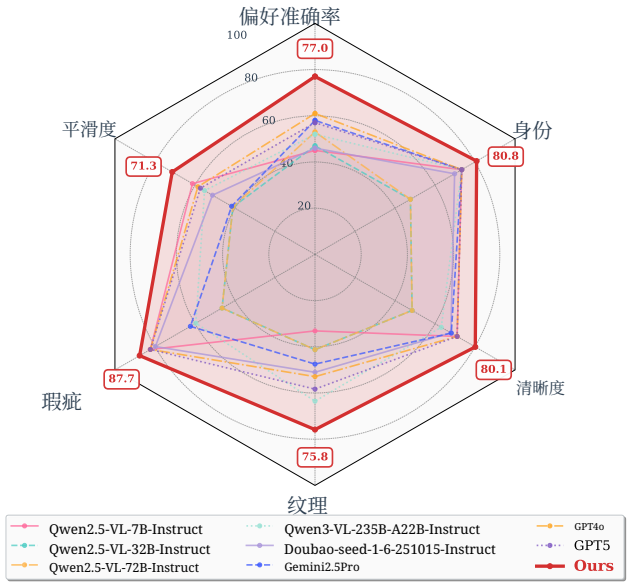


图5. 雷达图比较了我们的方法的奖励模型与各种视觉语言模型在人类判断准确率方面的表现。

道题展示一张原始图像和一张修饰结果，参与者使用五点评分量表，从五个感知维度对修饰进行评分，分别为皮肤平滑、瑕疵去除、纹理质量、清晰度和身份保持。如表2所示，我们的方法取得了显著更高的胜率，表明其与人类审美偏好的契合度更高。图5 比较了我们的奖励模型与其他视觉语言模型在五个维度及最终偏好分数上的评分准确率，结果表明，我们的奖励模型能够更可靠地捕捉细粒度的修饰偏好。用户研究设置的更多细节见补充材料。

4.4. 消融研究

4.4.1. 奖励模型

我们在 FluxKontext-LoRA 上将 BeautyGRPO 与各种编辑奖励模型结合使用，结果如表3所示。现有奖励模型主要评估图像与指令之间的粗粒度对齐程度以及全局感知质量，因此不太适用于人脸修饰。相比之下，我们的奖励模型针对五个关键维度进行细粒度奖励建模，从而在各项指标上均取得稳定提升。

4.4.2. Qwen-Image-Edit 上的 BeautyGRPO

我们进一步通过将 BeautyGRPO 应用于 Qwen-Image-Edit [47]来评估骨干网络的泛化能力。如表 4所示，我们通过 BeautyGRPO 训练的变体在所有指标上均优于官方预训练模型和 LoRA 微调模型。如图 6所示，预训练模型生成的面部不够真实，且存在身份偏移；LoRA 微调模型



图6. 不同训练范式下 Qwen-Image-Edit 的人脸修饰结果。

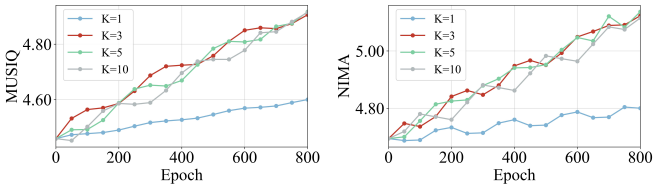


图7. FFHQR 上 DPG 步骤的消融实验 K。

表3. BeautyGRPO 在 FFHQR 上使用各种奖励模型的定量评估。

	NIMA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
EditReward[48]	4.754	4.462	1.028	0.613
EditScore[26]	4.737	4.576	1.045	0.621
UnifiedReward-Edit[43]	4.756	4.571	1.037	0.625
我们的奖励模型	**5.123**	**4.906**	**1.079**	**0.676**

表4. Qwen-Image-Edit 各变体的定量比较，展示了集成 Beauty-GRPO 后的优异性能。

	NIMA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
Qwen-Image-Edit	4.571	4.535	0.991	0.563
Qwen-Image-Edit+LoRA	4.824	4.745	1.028	0.595
Qwen-Image-Edit+LoRA w/我们的方法	**5.351**	**5.133**	**1.075**	**0.664**

仍会残留瑕疵，并过度平滑皮肤；而 BeautyGRPO 能更彻底地去除瑕疵，并保留更自然的皮肤纹理，从而生成更真实、视觉效果更佳的结果。

4.4.3. 不同的引导随机微分方程步长 K

我们对动态路径引导（DPG）步数进行了消融研究 K，考察其在采样过程中的使用情况。如图 7 所示，MUSIQ 和 NIMA 分数在 $K = 3$ 处与 $K = 5$ 处相当，同时所需的 DPG 更新更少，从而实现更快的采样。因此，我们采用 $K = 3$作为默认设置。

5. 结论

我们提出 BeautyGRPO，这是一种基于强化学习的方法，旨在高保真约束下使人脸修饰与人类审美偏好对齐。我们构建了 FRPref-10K 数据集，并训练了一个奖励模型，对人脸修饰的美化质量进行细粒度奖励建模。此外，我们引入 DPG，以协调随机探索与稳定的高保真采样轨迹。大量实验和用户研究表明，BeautyGRPO 能够生成高质量的修饰结果，且更符合人类审美偏好。

致谢

作者谨此感谢广东省信息安全技术重点实验室（资助编号：2023B1212060026）、广东省智能信息处理重点实验室和深圳市媒体安全重点实验室（资助编号：2023B1212060076）提供的支持。本工作在第一作者于中国 vivo 移动通信有限公司 vivo BlueImage Lab 实习期间完成。

参考文献

[1] Kaoru Arakawa。用于多媒体系统中人脸图像美化的非线性数字滤波器。载于 2004 IEEE 国际电路与系统研讨会（IEEE 编号：04CH37512），第 V–V 页。IEEE，2004 年。3 [2] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。Qwen2. 5-vl 技术报告。arXiv 预印本 arXiv:2502.13923，2025 年。3、4、5[3] Nazre Batool 和 Rama Chellappa。利用纹理方向场和马尔可夫随机场建模检测并修复面部皱纹。IEEE 图像处理汇刊，23(9):3773–3788，2014 年。3[4] Kevin Black、Michael Janner、Yilun Du、Ilya Kostrikov 和 Sergey Levine。使用强化学习训练扩散模型。载于国际学习表征会议论文集，2024 年。2、3[5] Yochai Blau 和 Tomer Michaeli。感知-失真权衡。载于 IEEE 计算机视觉与模式识别会议论文集，第 6228–6237 页，2018 年。4[6] Chaofeng Chen、Jiadi Mo、Jingwen Hou、Haoning Wu、Liang Liao、Wenxiu Sun、Qiong Yan 和 Weisi Lin。TOPIQ：一种从语义到失真的自顶向下图像质量评估方法。IEEE 图像处理汇刊，33:2404–2418，2024 年。6、4[7] Kevin Clark、Paul Vicol、Kevin Swersky 和 David J Fleet。在可微奖励上直接微调扩散模型。载于国际学习表征会议论文集，2025 年。2[8] Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、Dan Zhang、Evan Rosen 等。Gemini 2.5：推动具备高级推理、多模态、长上下文以及下一代智能体能力的前沿。arXiv预印本 arXiv:2507.06261，2025年。3[9] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim En- tezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel等。扩展整流流Transformer以实现高分辨率图像合成。见国际机器学习会议论文集，2024年。2[10]Ying Fan、Olivia Watkins、Yuqing Du、Hao Liu、Moonkyung Ryu、Craig Boutilier、Pieter Abbeel、Moham- mad Ghavamzadeh、Kangwook Lee和Kimin Lee。DPOK：用于微调文生图扩散模型的强化学习。《神经信息处理系统进展》，36:79858–79885，2024年。2, 3[11] 谷歌。https : / / aistudio . google . com / ?model = gemini - 2 . 5 - flash - image - preview，2025年。2, 3, 6[12] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi等。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。arXiv预印本 arXiv:2501.12948，2025年。3, 4[13] Ziyu Guo、Renrui Zhang、Chengzhuo Tong、Zhizheng Zhao、Peng Gao、Hongsheng Li和Pheng-Ann Heng。我们能否借助思维链生成图像？让我们逐步验证并强化图像生成。见IEEE/CVF计算机视觉与模式识别会议论文集，2025年。2[14] Pengcheng He、Xiaodong Liu、Jianfeng Gao和Weizhu Chen。DeBERTaV3：使用带梯度解耦嵌入共享的ELECTRA式预训练改进DeBERTa。见国际学习表征会议论文集，2023年。4[15] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen等。LoRA：大语言模型的低秩适配。国际学习表征会议论文集，1(2):3，2022年。6[16] Junjie Ke、Qifei Wang、Yilin Wang、Peyman Milanfar和Feng Yang。MUSIQ：多尺度图像质量变换器。见IEEE/CVF国际计算机视觉会议论文集，第5148–5157页，2021年。6, 4[17] Yuval Kirstain、Adam Polyak、Uriel Singer、Shahbuland Ma- tiana、Joe Penna和Omer Levy。Pick-a-Pic：一个用于文生图的用户偏好开放数据集。《神经信息处理系统进展》，36 36663，2023年 36652 3 : – 。[18] Black Forest Labs。FLUX。https://github.com/ black-forest-labs/flux，2023年。2[19] Black Forest Labs。FLUX.1-kontext-dev。https : / / huggingface.co/black-forest-labs/FLUX.1- Kontext-dev，2023年。3, 6[20] Biwen Lei、Xiefan Guo、Hongyu Yang、Miaomiao Cui、Xu- ansong Xie和Di Huang。ABPN：用于超高分辨率照片实时局部润饰的自适应混合金字塔网络。见IEEE/CVF Com-会议

puter视觉与模式识别，第2108–2117页，2022年。2, 3, 6 [21] Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel和Matt Le。面向生成建模的流匹配。见国际学习表征会议论文集，2023年。2, 4[22] Jie Liu、Gongye Liu、Jiajun Liang、Yangguang Li、Jiaheng Liu、Xintao Wang、Pengfei Wan、Di Zhang和Wanli Ouyang。Flow-GRPO：通过在线强化学习训练流匹配模型。《神经信息处理系统进展》，2025年。2, 3, 4, 6, 1[23] Jiaxin Liu、Qichao Ying、Zhenxing Qian、Sheng Li、Runqi Zhang、Jian Liu和Xinpeng Zhang。MoFRR：用于人脸修饰复原的扩散模型混合。见IEEE/CVF国际计算机视觉会议论文集，第12842–12851页，2025年。3[24] Shiyu Liu、Yucheng Han、Peng Xing、Fukun Yin、Rui Wang、Wei Cheng、Jiaqi Liao、Yingming Wang、Honghao Fu、Chun- rui Han等。Step1X-Edit：通用图像编辑的实用框架。arXiv预印本 arXiv: *2504 17761*, 2025 2, 3 . . [25] Xingchao Liu、Chengyue Gong和Qiang Liu。流动直且快速：利用整流流学习生成并迁移数据。见国际学习表征会议论文集，2023年。2, 4[26] Xin Luo、Jiahao Wang、Chenyuan Wu、Shitao Xiao、Xiyan Jiang、Defu Lian、Jiajun Zhang、Dong Liu和Zheng Liu。EditScore：通过高保真度奖励建模为图像编辑解锁在线强化学习。arXiv预印本 arXiv:2509 23909，2025年 2, 3, 8 . . [27] Chao Ma、Chih-Yuan Yang、Xiaokang Yang和Ming-Hsuan Yang。学习用于单图像超分辨率的无参考质量指标。《计算机视觉与图像理解》，158:1–16，2017年。6, 4[28] Vongani H Maluleke、Neerja Thakkar、Tim Brooks、Ethan Weber、Trevor Darrell、Alexei A Efros、Angjoo Kanazawa和Devin Guillory。从种族视角研究GAN中的偏见。见欧洲计算机视觉会议，第344–360页。Springer，2022年。4[29] Anish Mittal、Anush K Moorthy和Alan C Bovik。构建一个“完全盲”的图像质量分析器。《IEEE信号处理快报》，20(3):209–212，2013年。6, 4[30] OpenAI。https://platform.openai.com/docs/ models/gpt-image-1，2025年。2, 3[31] OpenAI，:，Aaron Hurst、Adam Lerer、Adam P. Goucher、Adam Perelman、Aditya Ramesh和Aidan Clark等。GPT-4o系统卡，2024年。3[32] Mihir Prabhudesai、Anirudh Goyal、Deepak Pathak和Katerina Fragkiadaki。利用奖励反向传播对齐文生图扩散模型。2023年。2[33] Mihir Prabhudesai、Russell Mendonca、Zheyang Qin、Kate- rina Fragkiadaki和Deepak Pathak。通过奖励梯度实现视频扩散对齐。arXiv预印本 arXiv:2407 08737，2024年。[34] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher 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。2, 3 [35] Seedream 团队、Yunpeng Chen、Yu Gao、Lixue Gong、Meng Guo、Qiushan Guo、Zhiyao Guo、Xiaoxia Hou、Weilin Huang、Yixuan Huang 等。Seedream 4.0：迈向下一代多模态图像生成。arXiv 预印本 arXiv:2509.20427，2025年。2, 6[36] Alireza Shafaei、James J. Little 与 Mark Schmidt。Autoretouch：自动专业人脸修饰。见 IEEE/CVF 冬季计算机视觉应用会议（WACV）论文集，第 990–998 页，2021年。3[37] Alireza Shafaei、James J Little 与 Mark Schmidt。Autoretouch：自动专业人脸修饰。见 IEEE/CVF 冬季计算机视觉应用会议论文集，第 990–998 页，2021年。2[38] Hossein Talebi 与 Peyman Milanfar。NIMA：神经图像评估。《IEEE 图像处理汇刊》，27(8)：3998–4011，2018年。6, 4[39] Sudha Velusamy、Rishubh Parihar、Raviprasad Kini 与 Aniket Rege。Fabsoften：通过动态皮肤平滑、引导式羽化与纹理恢复实现人脸美化。见 IEEE/CVF 计算机视觉与模式识别会议研讨会论文集，第 530–531 页，2020年。3[40] Bram Wallace、Meihua Dang、Rafael Rafailov、Linqi Zhou、Aaron Lou、Senthil Purushwalkam、Stefano Ermon、Caiming Xiong、Shafiq Joty 与 Nikhil Naik。使用直接偏好优化实现扩散模型对齐。见 IEEE/CVF计算机视觉与模式识别会议论文集，第 8228–8238 页，2024年。2, 3 [41] Fu-Yun Wang、Yunhao Shui、Jingtan Piao、Keqiang Sun 与 Hongsheng Li。Diffusion-NPO：面向扩散模型更优偏好对齐生成的负偏好优化。见国际学习表征会议论文集，2025年。2[42] Yibin Wang、Zhimin Li、Yuhang Zang、Chunyu Wang、Qinglin Lu、Cheng Jin 与 Jiaqi Wang。通过强化微调构建统一多模态思维链奖励模型。《神经信息处理系统进展》，2025年。4[43] Yibin Wang、Yuhang Zang、Hao Li、Cheng Jin 与 Jiaqi Wang。面向多模态理解与生成的统一奖励模型。arXiv 预印本 arXiv:2503.05236，2025年。2, 3, 8[44] Zhouxia Wang、Jiawei Zhang、Runjian Chen、Wenping Wang 与 Ping Luo。RestoreFormer：基于未退化键值对的高质量盲人脸修复。见 IEEE/CVF计算机视觉与模式识别会议论文集，第 17512–17521 页，2022年。2, 3, 6[45] Zhouxia Wang、Jiawei Zhang、Tianshui Chen、Wenping Wang 与 Ping Luo。RestoreFormer++：迈向基于未退化键值对的真实场景盲人脸修复。《IEEE 模式分析与机器智能汇刊》，45(12)：15462–15476，2023年。3, 6 [46] Xue Wen、Lianxin Xie、Le Jiang、Tianyi Chen、Si Wu、Cheng Liu 与 Hau-San Wong。RetouchFormer：基于先验的选择性自注意力半监督高质量人脸修饰 Transformer。见 AAAI 人工智能会议论文集，第 5903–5911 页，2024年。2, 3, 6[47] Chenfei Wu、Jiahao Li、Jingren Zhou、Junyang Lin、Kaiyuan Gao、Kun Yan、Sheng ming Yin、Shuai Bai、Xiao Xu、Yilei

Chen、Yuxiang Chen、Zecheng Tang、Zekai Zhang、Zhengyi Wang、An Yang、Bowen Yu、Chen Cheng、Dayiheng Liu、Deqing Li、Hang Zhang、Hao Meng、Hu Wei、Jingyuan Ni、Kai Chen、Kuan Cao、Liang Peng、Lin Qu、Minggang Wu、Peng Wang、Shuting Yu、Tingkun Wen、Wensen Feng、Xiaoxiao Xu、Yi Wang、Yichang Zhang、Yongqiang Zhu、Yujia Wu、Yuxuan Cai 与 Zenan Liu。Qwen-image 技术报告，2025年。3, 8[48] Keming Wu、Sicong Jiang、Max Ku、Ping Nie、Minghao Liu 与 Wenhu Chen。EditReward：面向指令引导图像编辑的人类偏好对齐奖励模型。arXiv 预印本 arXiv:2509.26346，2025年。2, 3, 8[49] Xiaoshi Wu、Yiming Hao、Keqiang Sun、Yixiong Chen、Feng Zhu、Rui Zhao 与 Hongsheng Li。Human Preference Score v2：评估文生图合成人类偏好的可靠基准。arXiv 预印本 arXiv:2306.09341，2023年。2, 3[50] Lianxin Xie、Wen Xue、Zhen Xu、Si Wu、Zhiwen Yu 与 Hau San Wong。在有限配对数据下的瑕疵感知渐进式人脸修饰。见 IEEE/CVF计算机视觉与模式识别会议论文集，第 5599–5608 页，2023年。3[51] Jiazheng Xu、Xiao Liu、Yuchen Wu、Yuxuan Tong、Qinkai Li、Ming Ding、Jie Tang 与 Yuxiao Dong。ImageReward：学习并评估文生图的人类偏好。见第 37 届神经信息处理系统国际会议论文集，第 15903–15935 页，2023年。2, 3[52] Jiazheng Xu、Yu Huang、Jiale Cheng、Yuanming Yang、Jiajun Xu、Yuan Wang、Wenbo Duan、Shen Yang、Qunlin Jin、Shurun Li 等。Visionreward：面向图像与视频生成的细粒度多维人类偏好学习。arXiv 预印本 arXiv:2412.21059，2024年。2[53] Zhidan Xu、Xiaoqin Zhang 与 Shijian Lu。结合扩散数据生成与频谱修复的人脸修饰。见 IEEE/CVF国际计算机视觉会议论文集，第 14722–14731 页，2025年。3[54] Wen Xue、Le Jiang、Lianxin Xie、Si Wu、Yong Xu 与 Hau San Wong。VRetouchEr：利用瑕疵流学习跨帧特征相互依赖以进行视频人脸修饰。见 IEEE/CVF计算机视觉与模式识别会议论文集，第 9141–9150 页，2024年。2, 3, 6[55] Wen Xue、Chun Ding、Ruotao Xu、Si Wu、Yong Xu 与 Hau-San Wong。RetouchGPT：通过瑕疵提示实现基于大语言模型的交互式高保真人脸修饰。见 AAAI 人工智能会议论文集，第 9059–9067 页，2025年。2, 3[56] Kai Yang、Jian Tao、Jiafei Lyu、Chunjiang Ge、Jiaxin Chen、Weihan Shen、Xiaolong Zhu 与 Xiu Li。利用人类反馈在无需任何奖励模型的情况下微调扩散模型。见 IEEE/CVF计算机视觉与模式识别会议论文集，第 8941–8951 页，2024年。2[57] Sidi Yang、Tianhe Wu、Shuwei Shi、Shanshan Lao、Yuan Gong、Mingdeng Cao、Jiahao Wang 与 Yujiu Yang。MANIQA：面向无参考图像质量评估的多维注意力网络。见 IEEE/CVF计算机视觉与模式识别会议论文集，第 1191–1200 页，2022年。6, 4

[58] Huizhuo Yuan、Zixiang Chen、Kaixuan Ji 和 Quanquan Gu。用于文生图的扩散模型自博弈微调。 神经信息处理系统进展，37:73366–73398，2024年。2[59] Zechuan Zhang、Ji Xie、Yu Lu、Zongxin Yang 和 Yi Yang。上下文内编辑：在大规模扩散变换器中通过上下文内生成实现指令式图像编辑。 *arXiv* 预印本 *arXiv:2504.20690*，2025年。2，3，6[60] Huaisheng Zhu、Teng Xiao 和 VasantG Honavar。DSPO：用于扩散模型对齐的直接分数偏好优化。载于 第十三届国际学习表征会议，2025年。2

BeautyGRPO：通过动态路径引导和细粒度偏好建模实现人脸修饰的审美对齐

补充材料

A. 动态路径引导的数学推导

A.1. 从流匹配常微分方程到随机微分方程（Flow-GRPO）

流匹配常微分方程。在流匹配生成模型中，使用确定性常微分方程（ODE）沿连续轨迹，将时间$t = 1$ 时的简单噪声分布 $p_1(x)$ 输运至 $t = 0$ 时的数据分布 $p_0(x)$。模型学习一个依赖时间的向量场 $v_\theta(x,\ t)$，使得对其进行积分

$$dx_t = v_\theta(x_t, t)\, dt \tag{11}$$

会使 x_t 从 $t = 1$ 时的纯噪声演变为 $t = 0$时的逼真样本x_0。严格遵循学习到的常微分方程轨迹，通常能够生成高保真度图像。然而，确定性路径与强化学习（RL）场景中的探索并不相容：我们需要随机性来探索其他可能的结果，并获得多样化的奖励信号。

ODE 到 SDE 的转换（FlowGRPO）。FlowGRPO [22]将确定性常微分方程转换为随机微分方程（SDE），同时保留原始常微分方程在每个时间 t步的边缘分布。具体而言，它会在探索过程中每个采样步骤注入高斯噪声：

$$dx_t = \left[v_\theta(x_t, t) + \frac{\sigma_t^2}{2t}\big(x_t + (1-t)v_\theta(x_t, t)\big)\right] dt + \sigma_t\, d\omega_t, \tag{12}$$

其中，$d\omega_t$ 表示标准维纳过程的增量，可近似为 $d\omega_t \approx \sqrt{dt}\,\epsilon_t$，其中 $\{\epsilon_t\}$在各时间步上独立同分布，服从 $\mathcal{N}(0,\ I)$，而 $\sigma_t = \eta\sqrt{\frac{t}{1-t}}$控制时间 t处的噪声幅度。

此处，式（12）中的随机微分方程是以反向时间（从 $t = 1$到 $t = 0$）形式定义的。对于较小的时间步 Δt，式（12）的欧拉–丸山离散化给出的条件均值为

$$\mu_t = x_t + \left[v_\theta(x_t, t) + \frac{\sigma_t^2}{2t}\big(x_t + (1-t)v_\theta(x_t, t)\big)\right]\Delta t, \tag{13}$$

以及逐步标准差

$$\sigma_{\text{step}} = \sigma_t\sqrt{\Delta t}. \tag{14}$$

因此，单步更新可以写为

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t, \qquad z_t \sim \mathcal{N}(0, I), \tag{15}$$

其中 z_t 是一个标准正态随机向量。公式（13）–（15）表明，每次更新都以模型预测 μ_t为中心，并叠加各向同性高斯噪声，因此，迭代这一随机更新会产生可能的最终状态 x_0 分布，而非单一的确定性结果。这使策略能够在相同条件下探索多种图像变体，并在强化学习设置中获得多样化的奖励反馈。

A.2. 保真度–探索冲突

尽管 FlowGRPO 中的随机微分方程为在线强化学习提供了所需的随机性，但它与人脸修饰对高保真度的要求相冲突。根据公式15中的离散更新，每个噪声样本 z_t 都会使状态偏离模型预测均值 μ_t，而累积项

$$\sum_t \sigma_{\text{step}}(t)\, z_t$$

经过 T 个采样步（我们将探索过程离散为 T 步）后，会逐渐使轨迹 $\{x_t\}$ 偏离高质量图像流形。这里，$\sigma_{\text{step}}(t) = \sigma_t\sqrt{\Delta t}$ 表示时间 t处逐步变化的标准差。在通用的文本到图像生成中，这种漂移可能是可以接受的，甚至有助于提升多样性；但在人脸修饰中，它主要表现为噪声伪影。减小 σ_t 可以缓解漂移，但也会抑制探索，从而限制强化学习寻找更高奖励解的能力。因此，我们引入了一种机制，使轨迹在保持受控随机探索能力的同时，始终接近高保真流形。

A.3. 动态路径引导（DPG）

为解决保真度–探索冲突，我们提出动态路径引导（DPG），通过重塑 FlowGRPO 随机微分方程中的噪点来进行锚定探索，使采样轨迹在保持随机性的同时，仍贴近高保真参考路径。

锚点引导目标。我们假设可以获取一个高质量的锚点样本 x_0^{anchor}，它位于期望的高保真度流形上（例如，来自我们的偏好数据集的一张评分最高的修饰后人脸）。在采样时间 t，给定当前状态 x_t，我们构造一条直线路径，在剩余的采样时间范围内连接 x_t 和 x_0^{anchor}。该锚点引导路径上下一时间步 $t - \Delta t$ 对应的目标定义为

$$x^*_{t-\Delta t} = \Big(1 - \frac{\Delta t}{t}\Big)x_t + \frac{\Delta t}{t}\, x_0^{\text{anchor}}, \tag{16}$$

这是一种线性插值，使状态从 x_t 朝 x_0^{anchor}移动距离的 $\Delta t/t$。从几何上看， $x^*_{t-\Delta t}$位于连接 x_t 和 x_0^{anchor}的线段上，并为下一次更新指定一个锚点一致的参考方向。在实际应用中，探索过程中的 t 被离散为 T 个采样时间步，并且我们仅对 $t > 0$应用锚点引导更新，因此除以 t 是有定义的，而 $t = 0$ 处的最后一步则被视为终止状态。

锚点校正噪声。回顾一下，式 (15) 中的 FlowGRPO 离散化形式可写为

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t, \qquad z_t \sim \mathcal{N}(0, I), \tag{17}$$

其中 μ_t 由式 (13) 给出，且 $\sigma_{\text{step}} = \sigma_t\sqrt{\Delta t}$。给定 x_t，这会产生如下高斯转移

$$x_{t-\Delta t} \mid x_t \sim \mathcal{N}\big(\mu_t,\, \sigma_{\text{step}}^2 I\big). \tag{18}$$

将下一状态精确映射到式 (16) 中锚点引导目标 $x^*_{t-\Delta t}$ 的噪点，可通过施加以下条件得到

$$x^*_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t, \tag{19}$$

由此得到锚点校正向量

$$z_t^{\text{anchor}} = \frac{x^*_{t-\Delta t} - \mu_t}{\sigma_{\text{step}}}. \tag{20}$$

式 (20) 中的向量 z_t^{anchor}可将下一状态精确映射到锚点引导目标 $x^*_{t-\Delta t}$。然而，如果在每一步都确定性地应用此校正，即$z_t = z_t^{\text{anchor}}$对所有 t，那么 $x_{t-\Delta t} = x^*_{t-\Delta t}$ 将在整个采样过程中成立，导致轨迹收敛到式 (16) 中的确定性锚点路径，并失去随机探索的空间。

混合噪点与时变引导。为保留随机性，DPG 将原始标准高斯分布 z_t 替换为锚点校正与独立标准正态样本的时变混合。令 $z_t^{\text{std}} \sim \mathcal{N}(\mathbf{0}, I)$ 与 x_t 相互独立，并定义

$$z_t^{\text{mix}} = \lambda_t\, z_t^{\text{anchor}} + \big(1 - \lambda_t\big)\, z_t^{\text{std}}, \quad 0 \le \lambda_t \le 1, \tag{21}$$

其中 λ_t 是时变引导权重。实践中，我们将反向时间采样过程离散为 T 个步骤，并用整数 $t \in \{0, \ldots, T-1\}$ 为其编索引，其中 T 表示探索过程中步骤的总数。我们采用与正文相同的线性调度，

$$\lambda_t = \frac{t}{\max(1, T-1)}. \tag{22}$$

采样按反向时间进行，从 $t = T-1$ （早期、高噪声步骤）到 $t = 0$（后期细化），因此，有效引导权重 λ_t 在早期采样时间步较大，并在接近最终细化步骤时逐渐降至 0。为简洁起见，我们在下文的推导中继续用 λ_t 表示这一随时间变化的权重。然后，我们将上文离散化中的 z_t 替换为z_t^{mix}，得到 DPG 转移

$$x_{t-\Delta t} = \mu_t + \sigma_{\text{step}} z_t^{\text{mix}}. \tag{23}$$

在给定 x_t 的条件下，锚点修正 z_t^{anchor} 是 $(x_t,\, x_0^{\text{anchor}})$ 的确定性函数，唯一的随机成分是 z_t^{std}。对 z_t^{std} 求条件期望，并利用 $\mathbb{E}[z_t^{\text{std}}] = 0$，可得

$$\begin{aligned}
\mathbb{E}[x_{t-\Delta t} \mid x_t] &= \mu_t + \sigma_{\text{step}} \mathbb{E}[z_t^{\text{mix}} \mid x_t] && (24)\\
&= \mu_t + \sigma_{\text{step}} \lambda_t z_t^{\text{anchor}} && (25)\\
&= \mu_t + \lambda_t \big(x^*_{t-\Delta t} - \mu_t\big) && (26)\\
&= \big(1 - \lambda_t\big)\mu_t + \lambda_t x^*_{t-\Delta t}. && (27)
\end{aligned}$$

因此，从期望意义上讲，DPG 在原始 FlowGRPO 均值 μ_t 与锚点引导目标 $x^*_{t-\Delta t}$ 之间对转移均值进行插值，插值权重为 λ_t。

类似地，仍在给定 x_t 的条件下，并将 z_t^{anchor} 视为确定量， z_t^{mix} 的协方差为

$$\text{Cov}\big[z_t^{\text{mix}} \mid x_t\big] = (1 - \lambda_t)^2 I, \tag{28}$$

因为 z_t^{anchor} 的条件方差为零，且 $\text{Cov}[z_t^{\text{std}}] = I$。这意味着

$$x_{t-\Delta t} \mid x_t \sim \mathcal{N}\Big(\big(1-\lambda_t\big)\mu_t + \lambda_t x^*_{t-\Delta t},\, \big((1-\lambda_t)\sigma_{\text{step}}\big)^2 I\Big). \tag{29}$$

当 $\lambda_t = 0$时，式（29）精确退化为原始 FlowGRPO 核，其均值为 μ_t，方差为 $\sigma_{\text{step}}^2 I$。当$\lambda_t = 1$时，方差坍缩为零，更新变为到达 $x^*_{t-\Delta t}$的确定性步骤。对于中间值$0 < \lambda_t < 1$，轨迹偏向锚定路径，同时保留减小但非零的方差。

对保真度与探索的影响。式（27）与（29）表明，DPG 同时（i）将均值转移拉向与锚点一致的方向，以及（ii）将步长方差按因子 $(1 - \lambda_t)^2$缩小。通过令 λ_t 在 t 较大且状态仍高度含噪时取较大值，探索过程中的采样轨迹在早期结构形成阶段被紧密约束在高保真锚定流形附近。随着 $t \to 0$与 $\lambda_t \to 0$，方差逐渐恢复至其原始水平，从而可在已经结构良好的人脸之上进行更精细的随机探索。由此，DPG 将人像修饰的保真度要求与在线强化学习的探索需求相协调，且仅通过对 FlowGRPO 转移核进行采样时修改，而不改变底层流匹配策略 v_θ。

B. 采用标准 GRPO 优化的锚定探索

需要特别强调的是，锚定图像仅用于基于采样的探索

阶段，以生成用于 GRPO 的轨迹，绝不会在优化目标中用作监督目标。在探索阶段，对于 FRPref-10K 中的每张输入肖像，我们选择一张高质量修饰图像作为锚点 x_0^{anchor}，仅用于构建 DPG 引导的采样路径。具体而言，在每个采样时间步 t，DPG 会计算锚点引导目标 $x^*_{t-\Delta t}$ 及相应的校正噪声 z_t^{anchor}（公式（16）–（20）），形成公式（ z_t^{mix} 21）中的混合噪声 $x_{t-\Delta t}$，并根据公式（ $x_{t-\Delta t}$ 23）更新 $x_{t-\Delta t}$。这种锚定探索只会改变强化学习轨迹采样期间 $\{x_t\}$ 轨迹的采样方式，不会在损失中增加任何朝向锚点的重建项或回归项。

在优化阶段，我们将标准 GRPO目标函数应用于策略模型 π_θ。奖励模型对最终的修饰后输出 x_0进行评分，并将这些分数转换为优势，通过截断GRPO损失优化策略。这一优化过程完全独立于锚定图像。具体而言，锚定图像 x_0^{anchor}从未被用作像素级或特征级目标，策略输出与锚定图像之间也不存在监督对齐。在推理时，DPG会被完全弃用，我们采用标准的常微分方程采样。这确保推理过程无需参考图像，也不会产生额外的计算成本。

C. 与参考引导方法、无分类器引导和路径正则化的关系

尽管 DPG 在探索阶段使用锚定图像，但其作用和实现方式与典型的参考引导编辑、无分类器引导（CFG）或路径正则化存在根本差异。

与参考引导编辑的区别。 传统的基于参考的方法会明确地将参考图像作为额外的条件信号注入模型架构中（例如通过拼接、交叉注意力或风格编码），通常还会训练网络通过监督损失严格模仿参考图像的外观。相比之下，我们的锚定图像从未作为输入条件经过修饰网络，也从未出现在优化目标中。相反，它仅作为几何锚点，用于在潜空间中定义基于锚点的ODE路径，并在采样器中计算校正噪声 z_t^{anchor}。策略模型 π_θ 始终仅以原始输入人像和文本提示为条件，通过奖励信号学习审美对齐，而不是通过针对锚定图像的直接监督来学习。

与无分类器引导（CFG）的区别。CFG 通过添加一个引导项来改变采样动态，该项与条件模型输出和无条件模型输出之间的差值成正比。这实际上会引导生成过程趋向条件下似然更高的区域。相比之下，DPG 无需外部分类器或双重前向传播。它并不改变

表 5. FFHQR 上失真与感知质量指标的对比。

	方法	峰值信噪比↑	结构相似性↑	学习感知图像块相似度↓	MUSIQ↑	MANIQA↑	NRQM↑
Supervised	RetouchFormer	42.088	**0.997**	0.024	4.465	1.036	8.175
	Xu（2025年）[53]	45.480	0.993	0.008	–	–	–
	RetouchGPT	**46.210**	–	**0.002**	–	–	–
Generative	ICEdit	32.463	0.882	0.152	4.192	0.978	7.424
	SeedDream	25.598	0.739	0.170	4.503	0.993	7.538
	NanoBanana	28.841	0.824	0.106	4.681	1.009	8.065
	Flux.Kw/LoRA	36.001	0.956	0.063	4.459	1.035	8.009
	Ours	33.697		0.102	**4.906**	**1.079**	**8.401**

向量场，而是通过一个随时间变化的权重 λ_t，将确定性锚点校正与标准高斯噪声线性混合，从而重塑随机转移核。这样既使转移均值偏向锚点引导的方向（式（27）），又缩小了方差（式（29）），使轨迹在早期采样阶段保持在高保真流形上，同时仍允许探索。所有这些都无需辅助引导网络，也无需修改漂移预测。

与路径正则化的区别。 路径正则化方法通常通过在训练损失中添加显式正则化项，惩罚整个轨迹上的偏离，从而强制学习到的向量场或采样轨迹接近预定义的参考路径。我们的 DPG 不会在损失函数中添加任何路径惩罚项或一致性项。基于锚点的 ODE 路径仅在采样时用于构造 $x^*_{t-\Delta t}$ 和 z_t^{anchor}，并且锚点的影响还会受到调度函数λ_t 的进一步调节，该调度函数沿采样时间步逐渐衰减。因此，DPG 充当的是对探索的采样时约束，而不是训练时的路径正则化器。它对 GRPO 收集经验所依据的展开分布进行正则化，而 GRPO目标函数本身保持不变。

总结。 DPG 与这些方法的区别在于：(i) 锚定图像既不是监督目标，也不是显式条件输入；(ii) 不使用辅助引导网络或分类器；(iii) 不会向训练损失中添加正则化项。相反，DPG 是一种轻量级的探索时机制，利用锚点诱导的几何结构重塑 SDE 采样核，在保持推理时采样与基于标准 ODE 的流匹配完全一致的同时，实现高保真且具有探索性的强化学习更新。

D. 评估指标

D.1. 为什么不使用全参考指标

以往的人脸修饰工作 [20, 44, 46, 54]通常采用经典的全参考（FR）指标，如峰值信噪比、结构相似性和学习感知图像块相似度，来评估修饰后输出与真实参考图像之间的像素级保真度。然而，我们认为，全参考指标根本不适用于

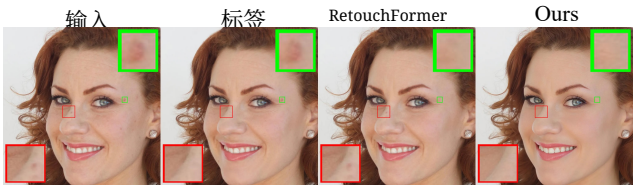


图 8. FFHQR 数据集中不完美真实参考图像的视觉示例，这些参考图像被我们的方法主动校正。

评估人脸修饰这类美学增强任务，这一观点有理论和实证证据支持：

• **理论冲突（感知-失真权衡）**。正如 Blau 和 Michaeli [5],所论证的那样，在最大化感知质量与最小化像素级失真（例如，峰值信噪比）之间存在固有的数学冲突。具体而言，严格最小化失真的模型往往会平均掉高频细节，常常导致过度平滑、感知上平坦的皮肤纹理。相反，合成逼真的面部细节（如自然毛孔和光照变化）不可避免地会引入相对于参考图像的逐像素偏差。这一权衡在表 5：中清晰可见：与我们的有监督基线（Flux.K w/LoRA）相比，经强化学习优化的 BeautyGRPO 以像素保真度（峰值信噪比从 36.00 降至 33.70 dB）换取显著的美学增益（MUSIQ 从 4.46 提升至 4.91）。严格针对全参考指标进行优化，从根本上限制了模型达到峰值感知质量的能力。

• **不完美的真实标注与缺陷复现**。 全参考指标默认目标参考图像毫无瑕疵。实际上，现有数据集（如 FFHQR）中的真实参考图像往往仍有残留瑕疵，或审美质量欠佳（见图 8）。因此，为了获得较高的峰值信噪比而优化的模型，会被迫复现这些缺陷，以严格匹配参考图像。相比之下，Beauty-GRPO 会主动纠正这些问题，突破数据集固有的局限，从而提升自然度和视觉吸引力。因此，全参考指标会因为我们的模型偏离这些有缺陷的参考图像而对其不公正地予以惩罚。

因此，我们重点关注无参考指标，这类指标直接评估自然度、真实感和审美质量，能够更准确、可靠地反映人类在人脸修饰中的真实偏好。

D.2. 无参考指标

基于这些原因，我们在评估中转而关注无参考的感知指标和审美指标，因为它们更能贴近人类的判断。

• **自然图像质量评估器 [29] (↓).** 自然图像质量评估器对高质量自然图像的统计规律进行建模，并衡量测试图像偏离这些规律的程度。对于人脸修饰，它
主要反映整体感知自然度，包括

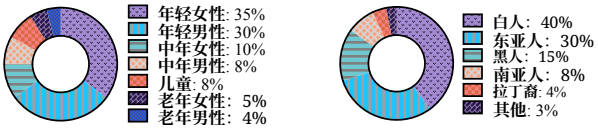


图9. FRPref-10K数据集的年龄与种族分布。

皮肤纹理是否显得不自然、过于平滑或油腻，以及是否存在明显的人工智能伪影。**较低**的分数表示感知自然度更高。

• **NIMA [38]**（↑）。NIMA预测人类意见评分的分布，并将期望评分作为审美质量的度量。经AVA训练的模型强调整体视觉吸引力，包括构图、色彩和谐和审美偏好。**更高**的分数表示更好的审美质量。

• **MUSIQ [16]**（↑）。MUSIQ通过综合考虑纹理保真度、清晰度和整体审美印象，对感知质量进行整体评估。经AVA训练的模型对全局平衡以及细粒度视觉质量较为敏感，因此适合用于评估人像增强。**更高**的分数表示更好的感知质量。

• **NRQM [27]**（↑）。NRQM预测从主观研究中学习得到的单一感知质量分数。在人脸修图中，它反映感知锐度、纹理真实感和局部结构保真度，尤其是在面部区域。**更高**的分数表示更好的感知质量。

• **MANIQA [57] (↑)**。MANIQA 旨在检测复杂失真和生成伪影。它对纹理完整性、精细面部细节以及修饰过程中引入的细微伪影尤为敏感。**更高**的分数表示感知质量和纹理质量更好。

• **TOPIQ [6] (↑)**。TOPIQ 采用自上而下的流程，着重关注语义重要区域中的失真。它使用经 GFIQA 训练的人脸模型，重点评估面部区域清晰度、纹理自然度和身份保持。**更高**的分数表示面部特定感知质量更好。

E. FRPref-10K 的构成与统计

为确保用于审美对齐的训练语料库具有多样性和代表性，我们精心构建了 FRPref-10K 数据集。该数据集由多种来源的图像组成：70% 的图像取自 FFHQR 数据集的官方训练集划分，其余 30% 则来自专有的高分辨率人像集。这种混合方式旨在帮助缓解 [28] FFHQR 数据集中存在的种族和人口统计偏差。为全面了解该数据集的多样性，我们在图9中详细展示了其人口统计分布。

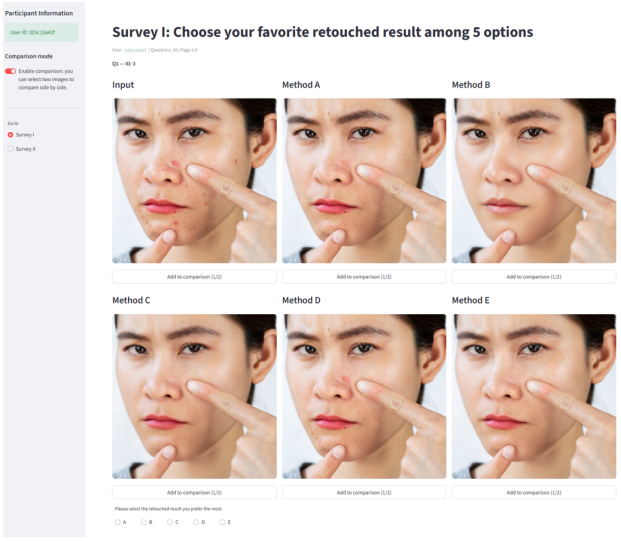


图10. 问卷I的用户界面。

F. 用户研究详情

F.1. 研究方案

我们设计了两份互补的问卷，用于评估用户的感知偏好，并考察不同奖励模型与人类判断的契合程度。

问卷I：总体偏好。第一份问卷包含一组固定的25道题。每道题都会展示一张原始输入人像，以及由不同方法（包括我们的方法和具有代表性的基线方法）生成的五张修饰结果。参与者需要针对每道题选出一张自己最喜欢的修饰图像。我们不施加任何额外指示或限制，并鼓励用户遵循第一印象，根据自己的主观审美偏好进行选择。25道题涵盖了不同年龄、性别和族裔，且每道题中的候选图像既包括差异明显的案例，也包括纹理和细节上较为细微的差异。为减少位置偏差，每道题中五张候选图像的顺序均经过随机排列。问卷I的用户界面如图10所示。

问卷II：用于奖励评估的分维度评分。第二份问卷旨在衡量不同奖励模型与人类在细粒度感知维度上的判断一致性。问卷选取了额外的20组输入-输出对。每位参与者会从中随机抽取5道题。每道题展示一张原始肖像和一张修饰结果，用户需从五个维度对结果评分：皮肤平滑、瑕疵去除、皮肤纹理质量、清晰度和身份保持。对于每个维

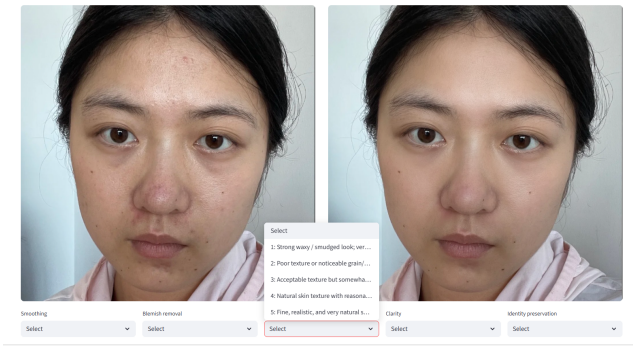


图11. 问卷II的用户界面。

度，参与者需给出0–5范围内的整数分数，数值越大表示感知质量越好。我们为每个维度的各个评分等级（从0 到5）提供了详细的文字说明和具体示例，以便用户在评分时遵循清晰一致的准则。随后，我们将这些人类评分与不同奖励模型的输出进行比较，以量化它们与人类感知的一致程度。问卷II的用户界面如图11所示。

数据收集与汇总。我们总共从匿名参与者处获得了100份有效问卷。对于问卷I，我们将某方法在所有问题和参与者的回答中被选为偏好结果的比例，计算为该方法的胜率。对于问卷II，我们对所有参与者和问题的评分取平均值，汇总各维度的人类评分，然后将这些平均值与不同奖励模型预测的相应评分进行比较。这使我们能够评估奖励模型与人类判断在各维度上的一致性。

F.2. 用户偏好的人口统计学分析

图 12 展示了问卷I中按用户年龄组和性别划分的胜率。我们的方法在每个人口统计群体中都取得了最高胜率，始终获得至少50%的用户偏好，在18至24岁的女性中更是达到约75%，以及 45+ 岁。其他方法获得的份额要小得多，在任何群体中的占比均未超过25%，也没有任何方法在某个人口统计群体中表现出持续优势。这些结果表明，我们的方法在偏好方面的优势在不同年龄和性别群体中都很稳健，并非由某个特定的子群体所驱动。

G. 训练细节

G.1. 奖励模型训练细节

我们采用 Qwen2.5-VL-7B-Instruct [2]作为基础奖励模型。对于监督微调的两个阶段（阶段1和阶段2），训练

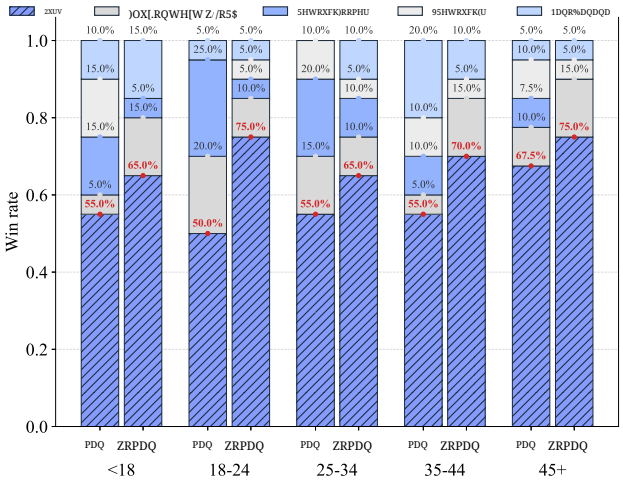


图12. 不同年龄和性别群体对各种人脸修饰方法的用户偏好胜率比较。

采用单设备批大小为 1、8 个梯度累积步数、学习率为 5×10^{-5}、预热比例为 0.1 的设置。对于 GRPO 阶段，我们采用单设备批大小为 2、4 个梯度累积步数、学习率为 1×10^{-6}、预热比例为 0.05 的设置，并将 KL惩罚系数设为 $\beta = 0.04$。生成的响应数量 N设为 8。

G.2. 修饰模型训练配置

我们采用 Flux.1 Kontext [19] 作为基础修饰模型，并首先在人脸修图语料库上进行 LoRA 适配。模型的训练分辨率为 1024×1024，LoRA 秩为 32 ，并且 $\alpha = 32$，使用 AdamW，学习率为 1×10^{-5}，采用余弦学习率衰减，500 个预热步数，权重衰减为 0.01，单设备批大小为4，以及 2 个梯度累积步数。对于 BeautyGRPO，我们从经过 LoRA 适配的 Flux.1 Kontext 初始化，并以 1024×1024 分辨率运行在线强化学习，单设备批大小为 2 ，每个提示生成 8 张图像，使用 $T = 10$个采样步数，噪声水平为 $\eta = 0.7$。

H. 奖励模型的思维链推理示例

我们在图中提供了奖励模型思维链（CoT）推理过程的定性示例13。

I. 更多视觉结果

图14–16 展示了不同人脸修饰方法之间更多的定性比较，进一步凸显了我们提出的 BeautyGRPO 在感知方面的优势。

J. 讨论与未来工作

J.1. 锚点选择与质量

在主要实验中，我们为每个输入采用排名第一的高偏好样本作为锚点 x_0^{anchor}，这样既能简化实现，也能确保锚点位于高保真流形上。从概念上讲，DPG 只要求锚点是一个质量较好的示例，能够将探索限制在高质量邻域内，同时仍允许策略在奖励信号表明存在更优解时超越该锚点。对其他锚点选择策略（例如从 top-k 集合中采样、采用多样性感知锚点或组合多个锚点）进行系统研究，是未来值得探索的方向。

J.2. 无显式锚点的场景

在这项工作中，我们聚焦于一种实用的人脸修饰场景：每张输入肖像都至少配有一张高质量的修饰后示例图像（例如 FRPref-10K 中经过筛选或获得高评分的图像），这些图像可自然地为 DPG 提供锚点。在仅有原始输入、没有此类示例图像的场景中，则可以利用强大的奖励模型、自蒸馏 EMA 策略或其他生成先验来构建伪锚点，然后应用相同的锚定探索原则。探索此类无锚点或弱锚定场景，是 DPG 一个有趣的拓展方向，我们将其留待未来工作。

J.3. 推理时的DPG

我们仅在强化学习训练的基于采样的探索阶段应用 DPG。优化完成后，标准 ODE 采样会利用改进后的向量场生成高保真结果，无需锚点，也不会在推理时增加额外计算。原则上，可以通过提供参考图像或即时合成锚点来设计推理时版本的 DPG，但这会引入额外的假设和复杂性。因此，我们在部署时采用纯 ODE 采样，并将 DPG 的推理时扩展留待未来工作。

Input



Image1

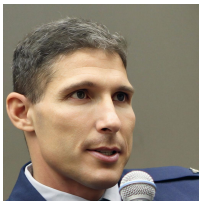
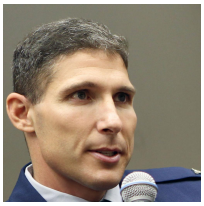


Image2



你是一位专业的人像美颜评估专家，精通图像语义理解、面部结构分析和审美判断。

你的任务是：根据用户提供的三张图像——一张原始图像（输入）和两张美化结果图像（图像1和图像2），系统地分析两张美化图像相较于输入图像所做的编辑。针对每个维度，评估编辑强度和编辑质量。

你必须放大局部区域，检查像素级细节，并特别关注以下五个维度：1. 皮肤平滑效果：检查输入图像中的面部皮肤是否经过适度平滑，合理减少额头、脸颊、鼻翼、嘴角和眼周的毛孔与细纹，同时保持自然，避免出现塑料感或毛孔完全消失。2. 瑕疵去除效果：检查输入图像中的面部瑕疵（如色斑、痤疮、粉刺或肤色不均）是否得到有效去除。3. 肤色与纹理效果：评估肤色、面部光泽和纹理，判断皮肤是否看起来更美观。4. 个人特征保留：检查是否完整保留重要的个人特征（例如男性的面部毛发、女性的妆容、老年人的皱纹、痣和雀斑），没有被删除或弱化。5. 整体清晰度：指全局图像质量，即面部特征轮廓（眼睛、嘴唇、鼻梁）是否清晰锐利，以及是否存在模糊、噪点或人工智能伪影。

对于每个维度，将图像1和图像2相对于输入进行比较，并给出1到5的整数分数（分数越高表示结果越好）。
严格使用以下输出结构：
1. 所有推理和分析必须写在以下内容之内：
<思考> ... </思考>
2. 所有评分和最终判断必须写在以下内容之内：
<答案>...</答案>
[评分部分]
1. 皮肤平滑效果：图像1 = X，图像2 = Y
2. 瑕疵去除效果：图像1 = X，图像2 = Y
3. 肤色与纹理效果：图像1 = X，图像2 = Y
4. 个人特征保留：图像1 = X，图像2 = Y
5. 整体清晰度：图像1 = X，图像2 = Y
图像1 = X.XX，图像2 = Y.YY
[偏好判断]
图像X更好
</回答>

<思考>

1. 皮肤平滑效果：- 图像1采用了轻度至中度的皮肤平滑处理。额头和眼周的细纹略有减少，但仍可见皮肤纹理，整体保持自然观感。- 图像2采用了强得多的皮肤平滑处理。额头上的纹路几乎被完全抹去，脸颊和眼下区域显得过度磨皮，形成塑料般的质感，

d

自然皮肤细节的损失。2. 瑕疵去除效果：- 图像1去除了一些轻微瑕疵，并略微均匀了肤色，但仍有些许不均匀，尤其是脸颊和下颌线周围。- 图像2去除了所有可见瑕疵和色斑，使皮肤看起来非常干净、均匀。3. 肤色与纹理效果：- 图像1保留了自然肤色，并稍微提升了亮度和暖色调，呈现出健康、真实的效果。- 图像2的色调更平淡，也更不自然。皮肤看起来像蜡一样，缺乏自然光泽和纹理，因此显得不够逼真。4. 个人特征保留：- 图像1保留了所有个人特征，例如面部结构、头发纹理和自然纹路。- 图像 2 由于过度平滑，略微改变了面部结构，尤其是眼睛和脸颊周围。一些自然轮廓也有所流失。5. 整体清晰度：- 图像 1 保持了眼睛、嘴唇和头发的锐度，画面清晰且细节丰富。- 图像2的面部特征略显模糊，尤其是眼睛和嘴部周围，这可能是过度平滑造成的。</think><answer>[评分部分]1. 皮肤平滑效果：图像1 = 4，图像2 = 2 \n2. 瑕疵去除效果：图像1 = 4，图像2 = 5 \n3. 肤色与纹理效果：图像1 = 5，图像2 = 1 \n4. 个人特征保留：图像1 = 5，图像2 = 3 \n5. 整体清晰度：图像1 = 5，图像2 = 3[平均分] 图像1 = 4.6，图像2 = 2.8[偏好判断] 图像1更好。</answer>

图13. 人脸修饰思维链奖励推理的定性示例。给定一张输入图像和两个修饰候选图像，我们的奖励模型会从五个维度进行思维链质量评估：皮肤平滑、瑕疵去除、皮肤纹理质量、清晰度和身份保持。



图 14. FFHQR 数据集上不同方法的人脸修饰结果视觉对比，其中 Flux.K 表示 FluxKontext。与现有方法存在瑕疵去除不彻底、过度平滑或视觉效果不自然等问题相比，我们的 BeautyGRPO 能够干净地去除瑕疵，同时保留身份一致的自然皮肤纹理、光泽以及皱纹等精细细节。

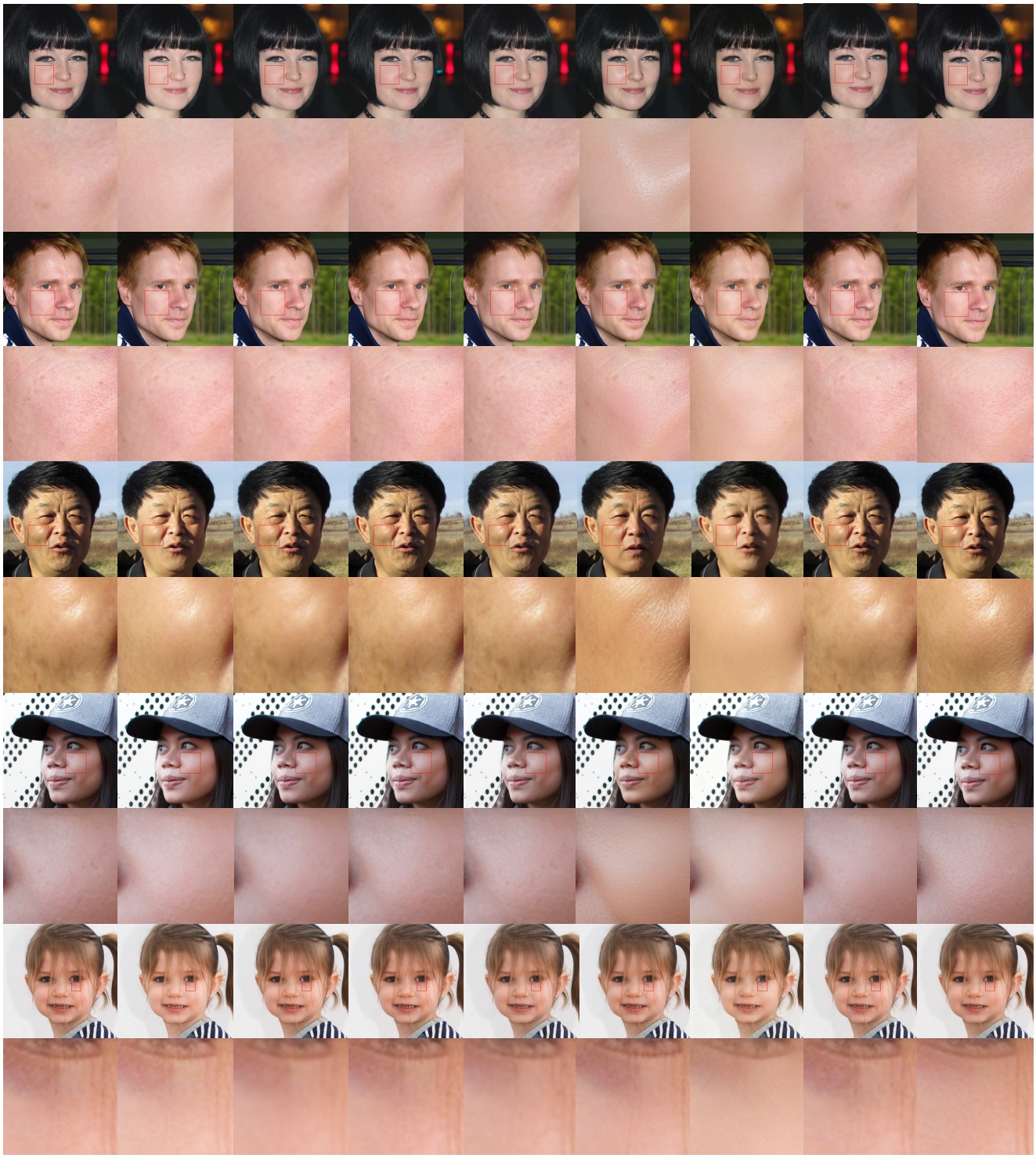


图15. FFHQR 数据集上不同方法的人脸修饰结果视觉对比，其中 Flux.K 表示 FluxKontext。与现有方法存在瑕疵去除不彻底、过度平滑或视觉效果不自然等问题相比，我们的 BeautyGRPO 能够干净地去除瑕疵，同时保留身份一致的自然皮肤纹理、光泽以及皱纹等精细细节。



图 16. 真实场景数据集上不同方法的人脸修饰结果视觉对比，其中 Flux.K 指 FluxKontext。与现有方法存在瑕疵去除不彻底、过度平滑或视觉效果不自然等问题相比，我们的 BeautyGRPO 能够干净地去除瑕疵，同时保留身份一致的自然皮肤纹理、光泽，以及痣等精细细节。