arXiv:2603.14916v1 [cs.CV] 16 Mar 2026

EditHF-1M：面向图像编辑的百万规模丰富人类偏好反馈

Zitong Xu[1], Huiyu Duan[1*]、Zhongpeng Ji[2], Xinyun Zhang[3], YutaoLiu[4],Xiongkuo Min[1], Ke Gu[5], Jian Zhang[2], ShusongXu[2], JinweiChen[2], Bo Li[2*]和 Guangtao Zhai[1**]通讯作者

[1] 上海交通大学[2] 维沃移动通信有限公司[3] 中国电子科技大学[4] 中国海洋大学[5] 北京工业大学

摘要。 近年来，文本引导图像编辑（TIE）模型取得了显著进展，但许多编辑后的图像仍然存在伪影、意外编辑和不美观内容等问题。尽管已有一些基准测试和方法被提出用于评估编辑后的图像，但可扩展的评估模型仍然缺失，这限制了图像编辑领域人类反馈奖励模型的发展。为应对这些挑战，我们首先推出 **EditHF-1M**，这是一个百万规模的图像编辑数据集，包含超过 2900 万个人类偏好对和 14.8 万个人类平均意见评分，所有数据均从三个维度进行评估，即视觉质量、指令对齐和属性保留。在 EditHF-1M 的基础上，我们提出了**EditHF**，一种基于多模态大语言模型（MLLM）的评估模型，用于提供与人类判断对齐的图像编辑反馈。最后，我们推出 **EditHF-Reward**，利用 EditHF 作为奖励信号，通过强化学习优化文本引导图像编辑模型。大量实验表明，EditHF 与人类偏好具有更优的对齐效果，并在其他数据集上展现出强大的泛化能力。此外，我们使用 EditHF-Reward 对 Qwen-Image-Edit 进行微调，实现了显著的性能提升，证明了 EditHF 作为奖励模型扩展图像编辑能力的潜力。数据集和代码将在我们的 GitHub 仓库中发布：
https://github.com/IntMeGroup/EditHF。

1 引言

图像编辑技术的快速发展，使得通过自然语言指令 [3,7,44,56,66,83]即可对图像进行精细修改。尽管一些最先进的编辑模型（如 Nano-Banana [7], SeedDream [56], Qwen-Image- Edit [66], 等）已取得令人瞩目的表现，但许多编辑后图像仍然

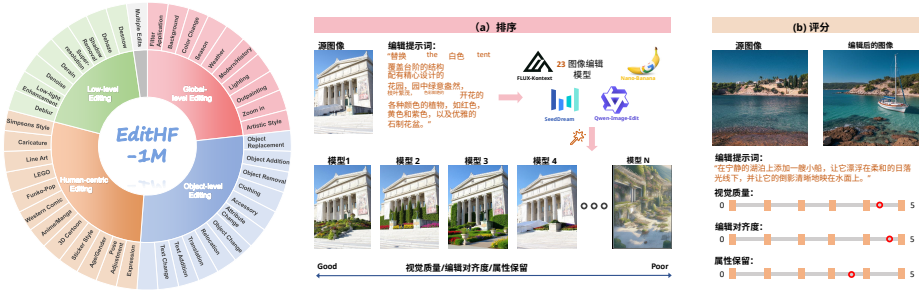


图 1： EditHF-1M 与人工标注概览：(a) 排序：将来自不同编辑模型的编辑后图像按源图像和编辑提示词分组，并进行排序，以评估相对质量；(b) 评分：对每张编辑后图像分别进行评分，以评估绝对质量。

表 1： 文本引导图像编辑模型评估基准的比较。✓ 表示公开可用，✗ 表示不适用或不可用。

数据库	已编辑源图像		标注	编辑模型	任务	提示 Type	维度	评估验证对象 模型	细化
TedBench [23]	100	400	300 对	3	4	描述	总体	✗	✗
EditVal [2]	648	5,184	10.3K 分数	8	12	指令	对齐，保留	✗	✗
EditEval [17]	150	1,050	4.2K 分数	7	19	描述，指令	质量，对齐，保留，真实感	✓	✗
AURORA-Bench [26]	400	2,000	2K 分数	5	8	指令	总体	✓	✗
ImagenHub [27]	179	1,432	2.8K 分数	8	-	描述，指令	质量，保留	✗	✗
IE-Bench [60]	301	1,505	4.5K 分数	5	11	描述，指令	质量、对应性、保留	✓	✗
EBench-18K [74]	1,080	18,360	54K 分数	17	21	描述，指令	质量、对齐、保留	✓	✗
EditScore [41]	577	2,885	3,072 对	11	13	指令	质量、对齐、保留	✓	✓
EditReward [69]	9,557	114.7K	200K 对	6	7	指令	质量、对齐	✓	✓
EditHF-1M	**44.3K**			**23**	**43**	**描述，指令**	**质量，对齐，保留**	✓	✓

存在伪影、意料之外的编辑以及内容不够美观等问题。因此，研究者推出了多个数据集和基准测试，用于评估文本引导图像编辑（TIE）模型 [2,17,41,43,60,69,74]。然而，如表1所示，现有基准测试存在若干关键局限。**(i) 内容多样性有限：**大多数现有基准测试仅涵盖有限的编辑任务，并且依赖数量较少的源图像 [17, 41, 60], ，这限制了它们在不同场景下评估灵活、通用图像编辑模型的能力。**(ii) 对编辑模型的覆盖有限：**大多数现有数据集纳入的编辑方法数量不足 [60, 69] ，且常常未涵盖近期最先进的模型或闭源模型 [2,17,43], ，因此评估并不完整且可能存在偏差，无法反映模型在实际场景中的性能。**(iii) 对模型细化的效用有限：**大多数现有数据集和基准测试主要用作静态评估工具，很少考察或展示其在细化图像编辑模型方面的效果 [17, 60, 74]。

除了数据集和基准测试外，还有一些指标被开发或用于评估图像编辑，包括图像质量评估（IQA）指标 [8,46,48,57,58,64,65,84,85], 视觉语言方法 [15,33,72], 以及基于多模态大语言模型（MLLM）的评估方法 [17, 43]。传统的图像质量评估指标主要关注自然失真和低级视觉保真度，但无法捕捉与用户指令的语义对齐。CLIPScore [15]和 HPSv2 [70] 等视觉语言方法已展现出

通过融入人类视觉反馈，在评估图像生成方面取得了显著进展 [15,33,36,70,72]，然而，这些方法主要评估文本-图像对齐，并非旨在捕捉图像编辑特有的语义变化。近来，已有多项研究探索利用多模态大语言模型评估图像编辑 [17, 41, 43, 69, 74]。然而，零样本多模态大语言模型评估 [17, 43]在复杂的编辑场景中，往往与人类偏好的对齐程度有限；而微调多模态大语言模型 [41, 69, 74]遇到未见过的编辑模型和编辑任务时，通常会受到泛化能力有限的困扰。此外，以往研究尚未考虑将分数回归与成对比较相结合来训练模型，从而同时实现有效的评估和奖励能力。

在这项工作中，我们介绍了 **EditHF-1M**，这是一个百万规模的图像编辑评估数据集，由训练有素的标注员按照严格且标准化的标注协议整理而成。EditHF-1M 包含 **44.3K** 张来自 Pico-Banana-400K [53],EBench-18K [74],和免费摄影网站的源图像，每张图像均配有编辑提示词，涵盖 **43 项编辑任务**，如图 1所示。基于这些源图像和提示词，我们使用 **23** 种最先进的图像编辑模型生成了超过 **1M**张编辑后图像，其中包括 Qwen-Image-Edit 等开源模型 [66] 和 OmniGen2 [67], ，以及 NanoBanana 等闭源模型 [7] 和 Seedream4 [56]。通过一项大规模主观研究，我们收集了超过 **29M**个人类偏好对，以及**148K** 个人类平均意见评分，涵盖感知质量、编辑对齐度和属性保留等维度。在 EditHF-1M 的基础上，我们提出了 **EditHF**，这是一种基于多模态大语言模型的图像编辑评估模型，可从多个维度进行评估。在 EditHF-1M 和其他基准测试上开展的大量实验表明，EditHF 达到了最先进的性能，并展现出强大的泛化能力。此外，我们还提出了 **EditHF-Reward**，结果表明，EditHF 可有效用作强化学习的奖励模型，通过优化 Qwen-Image-Edit 来提升图像编辑模型 [66]。

本研究的主要贡献包括：

– 我们推出了 EditHF-1M，这是一个百万规模的数据集，包含来自多种任务的 100 多万张编辑后图像，并提供全面的人类偏好排名和分数，涵盖感知质量、编辑对齐度和属性保留等维度。– 我们提出了 EditHF，这是一种通过联合训练分数预测和成对比较而构建的统一模型，可从多个维度实现与人类判断一致的图像编辑评估。– 我们推出了 EditHF-Reward，它将 EditHF 用作强化学习的奖励信号，以增强图像编辑模型的能力。– 在 EditHF-1M 和其他图像编辑基准测试上进行的大量实验表明，EditHF 达到了最先进的性能，并具有良好的泛化能力。此外，EditHF-Reward 可以借助 EditHF 有效提升图像编辑模型的性能。

2 相关工作

2.1 图像编辑

随着 Stable Diffusion 等生成模型的发展 [54]和 FLUX [10], , 人们提出了众多图像编辑方法 [17]。根据编辑提示词的类型，这些方法可分为基于描述的方法（例如，“一只鹿在草地上奔跑” → “一匹马在草地上奔跑”）和基于指令的方法（例如，“将鹿换成马”） [17, 74]。基于描述的方法主要采用基于反演的修改 [11, 21, 28,50, 73], , 先将图像反演为潜在表示，再根据新的提示生成编辑结果。早期的基于指令的方法 [3, 35, 83]在指令-图像配对数据集上训练模型。较新的方法 [7, 56,66, 67, 71] 则统一理解与生成，解读编辑指令，并实现灵活且高度逼真的图像编辑。

2.2 图像编辑基准

现有的图像编辑基准，以及与我们的 EditHF- 1M 的比较，总结于表 1中。TedBench [23]是首个图像编辑基准，仅包含带有提示的 100 张源图像。EditVal [2] 涵盖了更多任务和方法，但由于数据源的原因，生成的图像通常存在分辨率低和模糊的问题。EditEval [17]包含更多图像编辑任务，但其分数直接由多模态大语言模型生成，未能与人类感知保持一致。IE-Bench [60] 和 EBench-18K [74]提供平均意见分（MOS），但图像规模有限。EditScore [41] 和 EditReward [69] 纳入了人类偏好对，并展现出作为奖励模型的潜力，但其编辑任务和源内容的覆盖范围仍然有限，因此不太适合评估灵活的编辑模型。相比之下，EditHF-1M 包含超过一百万张编辑后图像，涵盖 43 个编辑任务和 23 个编辑模型，可实现全面评估。

2.3 图像编辑评估指标

传统的图像质量评估（IQA）模型包括全参考（FR）IQA [12, 30, 64, 76, 84, 85] 和无参考（NR）IQA [22, 46, 47, 55, 59, 79] 指标。这些方法主要评估低层次视觉质量和自然失真，但无法捕捉编辑指令与编辑后的图像之间的关系。为纳入文本提示，研究者提出了视觉语言指标[15, 25, 33, 36, 72] ，用于评估基于文本的图像生成中的对齐程度。尽管这些指标能够有效衡量图像与文本之间的对应性，但通常无法评估编辑指令与编辑后的图像之间的语义对齐。随着多模态大语言模型的发展，许多多模态大语言模型展现出描述图像质量的有效性 [62,75]。例如， [41,69,74] 等工作基于人类偏好数据对多模态大语言模型进行微调，以提供更准确的评估。 [41,69]

进一步研究表明，学习型评估模型可以充当奖励函数，通过强化学习优化图像编辑模型，从而弥合评估与生成之间的差距。然而，这些学习型评估器通常仅涵盖范围有限的编辑任务和源内容，限制了其在现代、大规模且多样化的图像编辑系统中的适用性。为此，我们提出了 EditHF，以提供与人类判断对齐的反馈，并提出 EditHF-Reward，利用 EditHF 改进图像编辑模型。

3 EditHF-1M

本节介绍 **EditHF-1M**，这是首个百万规模的图像编辑基准，兼具人类偏好对和细粒度评分。该基准包含 44.3K 张高质量源图像，涵盖 43 项编辑任务的指令型和基于描述的提示；由 23 个图像编辑模型生成的逾百万张编辑后图像；以及大规模人工标注，包括 29.1M 个偏好对和 148K 项评分，涉及感知质量、编辑对齐度和属性保留。凭借丰富多样的图像内容和全面的标注，EditHF-1M 为评估图像编辑模型提供了综合性资源。

3.1 编辑任务的设计

综合考虑实际相关性和现实世界中的使用情况，我们选取了 43 项图像编辑任务，并将其划分为**全局级任务**、**对象级任务**、**以人为中心的任务**和**低层任务**，如图1所示。具体而言，全局级任务涉及对整幅图像进行修改，例如整体风格迁移或全局颜色调整；对象级任务侧重于编辑特定对象或区域，例如添加、移除和属性修改；以人为中心的任务针对与人物主体相关的编辑，例如表情变化或姿势调整；低层任务则指旨在提升图像质量的像素级变换，例如去噪、去模糊和超分辨率。所有编辑任务的详细描述见补充材料。

3.2 数据收集

对于高级任务，我们从公开的摄影网站收集了约 20K 张源图像，每张图像都附有文本描述。根据图像内容和预定义的编辑任务，我们利用先进的多模态大语言模型 InternVL3.5 [63] 生成相应的编辑提示词，随后进行仔细的人工检查和细化。为适应不同类型的编辑模型，生成的提示既包括基于指令的形式，也包括基于描述的形式。此外，我们还从 Pico-Banana-400K [53], 中抽取了 26K 张图像，该数据集提供基于指令的编辑提示。

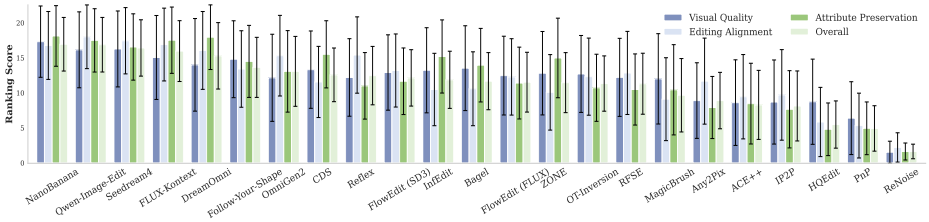


图2： 使用 EditHF-1M 对图像编辑模型进行比较。排名分数根据两两组间比较中的获胜次数得出。

对于这些图像，我们进一步使用 InternVL3。5 [63]根据图像内容和原始指令生成补充的基于描述的提示，并再次进行人工验证和修订。高级任务中的所有源图像的最低分辨率为 1024×1024，达到或超过了大多数现有图像编辑模型所要求的最高输入分辨率。对于低层任务，源图像及其对应的编辑提示均直接取自 EBench-18K [74]。

接下来，我们选取了 23 个图像编辑模型，涵盖基于描述和基于指令的方法，包括开源和闭源方法，以及多种骨干架构。这些模型的详细描述见补充材料。需要指出的是，部分模型无法针对某些源图像和提示生成有效的编辑后图像，而是产生异常输出（例如噪声图像或黑色图像）。我们已过滤掉这些图像，同时确保测试集中每项编辑任务和每个模型对应的图像相同，以便进行进一步的公平比较。最后，我们使用源图像和编辑提示生成了总计 1.01M 张编辑后图像。

3.3 主观实验

为了评估编辑后的图像，我们基于 EditHF-1M 开展了一项主观质量评估实验，旨在捕捉人类对编辑后图像的偏好，并确保其与真实世界的人类感知保持对齐。与现有基准测试通常只标注分数或排序不同，我们同时收集了组内排序和绝对分数，如图1所示。

对于同一张源图像和编辑提示词下不同模型生成的结果，组内排序尤其适合进行比较，因为它可以减少个体评分偏差，并提供更可靠的相对判断。因此，排序非常适合用于特定编辑案例中的模型层面比较。然而，排序仅适用于局部比较，无法直接汇总不同编辑任务或图像内容的结果。相比之下，绝对评分支持跨任务和跨内容比较，使我们能够分析模型在多种编辑场景下的性能，并研究模型在特定编辑任务中的优势与不足。通过结合利用排序和分数，我们的评估方案兼具两者的优势，从而能够可靠地评估图像编辑模型。为了全面评估图像编辑，我们从三个维度进行主观评估：

– **感知质量：** 衡量编辑后的图像的总体视觉质量，例如真实感、无伪影程度、结构完整性、色彩一致性以及视觉细节的丰富程度。– **编辑对齐度：** 评估编辑后的图像正确遵循给定编辑指令的程度，考察预期修改是否得到准确且完整的应用。– **属性保留：** 评估编辑后的图像在预期编辑之外保留源图像关键属性的程度，包括主体身份、关键视觉特征和情境一致性的保留。

实验使用基于 Python 的图形用户界面进行，该界面显示在经过校准、分辨率为 3840×2160 的 LED显示器上，图像以 1024×1024的分辨率呈现。标注员在受控环境中就座，距离显示器约 2 英尺。每个图像排序组由三名专业标注员进行标注，而每个分数则根据五名专业标注员提供的评分获得，以确保可靠性。为进一步确保测试集的可靠性，其分数根据十五名专业标注员提供的评分获得。所有标注员均按照标准化流程接受严格训练。通过预试验评估参与者对标准的理解程度及其与标准示例的一致性。未达到规定准确率阈值的参与者将被排除在后续参与之外。

对于排序标注，EditHF-1M 被划分为 44.3K 个图像组，每组包含由不同编辑模型根据同一源图像和编辑指令生成的编辑后图像。要求标注员分别沿三个评估维度对编辑后图像进行排序。每组的排序结果由三名标注员提供，并评估标注员间一致性。对于一致性较低的组，将由另外三名标注员重新标注，以确保排序结果可靠且一致。每个图像组的最终排序结果取三名标注员排序名次的平均值。我们进一步根据排序结果生成人类偏好对；对于包含 M 张图像的组，共生成 C_M^2 个偏好对。

为进行评分标注，我们从 EditHF-1M 中选取了涵盖所有编辑任务和模型的 49.4K 张编辑后图像，并确保这些样本的源图像内容和编辑提示词具有多样性。参与者会随机看到一张源图像、一张编辑后的图像及其对应的编辑提示词，并被要求从三个方面按照 5点连续量表对编辑后的图像进行评分。按照 [19], 那些与均值相差超过 2 个标准差的评分被视为离群值并予以剔除；离群评分超过 5 % 的参与者也会被移除。剩余评分会转换为 Z分数，并线性缩放至 [0,100]。每张图像的最终分数计算如下：

$$z_{ij} = \frac{s_{ij} - \mu_i}{\sigma_i},\ z_j = \frac{1}{N_j}\sum_{i=1}^{N_j} z_{ij},\ \text{Score}_j = \frac{100(z_j + 3)}{6} \tag{1}$$

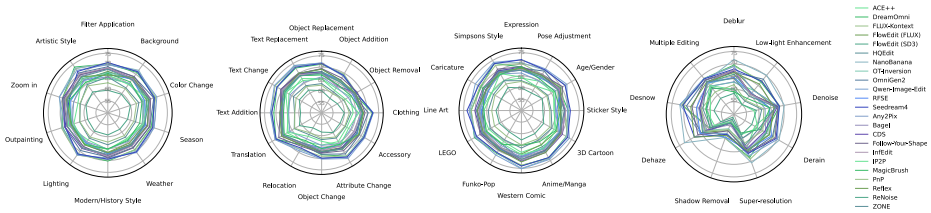


图3： 不同编辑任务中图像编辑模型的对比。

其中， s_{ij} 是第 i 位受试者对第 j 张图像给出的原始分数， μ_i 是评分均值， σ_i 是第 i 位受试者提供的标准差， N_j是第 j 张图像的评分数量。我们还会进一步比较分数和偏好对，并剔除分数与偏好对不一致的标注。标注协议和流程的详细描述见补充材料第4节。最后，我们从三个评估维度获得了 29.1M 对和 148K 个分数。

3.4 数据分析

为评测编辑模型，我们采用 EditHF-1M 中的组内排序。每张图像根据其在这些成对比较中的胜出次数获得排序分数。随后，按照以下公式，将视觉质量分数 Score_v、编辑对齐分数 Score_e 和属性保留分数 Score_p 结合，得出总分：

$$\text{Score}_{all} = \text{Score}_v^{0.3} \times \text{Score}_e^{0.4} \times \text{Score}_p^{0.3} \tag{2}$$

编辑对齐分数的权重更高，以强调满足编辑指令的重要性。图 2展示了 EditHF-1M 中各编辑模型的结果。Qwen-Image-Edit [77] 在编辑对齐度方面表现最佳，而 NanoBanana [7] 在其他评估维度上表现最佳。

我们进一步评测了各模型在不同编辑任务上的表现。由于组内排序无法反映单次编辑的绝对质量，我们采用了 EditHF-1M 中提供明确分数标注的部分。对于每项任务，通过对该任务中所有编辑样本的分数取平均，计算任务级分数。如图 3所示，模型在高级编辑任务中的表现通常更好，表明其处理语义级修改的能力较强。在高级任务中，姿势调整和物体移动的表现较弱，说明精准的几何控制和空间一致性仍具挑战。低级编辑任务的总体结果较差，反映出模型在精细像素级编辑方面存在局限。阴影去除在所有任务中表现最差，凸显了建模复杂光照交互并保持视觉真实感的难度。

4 EditHF

本节介绍 EditHF，这是一种统一的图像编辑评估模型，同时支持细粒度质量分数预测和成对偏好

比较，可根据人类感知共同评估感知质量、编辑对齐度和属性保留。

4.1 模型架构

受多模态大语言模型在多模态推理方面成功的启发，我们采用多模态大语言模型作为 EditHF 的骨干网络。该模型联合推理编辑后的图像I_e、对应的源图像 I_s 和编辑提示词 T，并在三个维度上生成细粒度评分，如图4所示。

首先，使用基于 InternViT 的预训练视觉编码器 E_v，从 I_e 和 I_s 中提取视觉特征 [6]。随后，通过带有参数 φ的可训练投影器 P_φ 将这些特征投影到语言嵌入空间中，得到视觉词元

$$T_e = P_\varphi(E_v(I_e)), \qquad T_s = P_\varphi(E_v(I_s)). \tag{3}$$

与此同时，编辑提示词被分词为文本词元 T_p。视觉词元和文本词元拼接后输入先进的多模态大语言模型 InternVL3.5[63], ，以执行多模态特征融合与联合推理。将 LLM 骨干网络定义为 H_ψ，其可训练参数为 ψ，即可获得最后一层隐藏状态 $\mathbf{h} = H_\psi(T_e, T_s, T_p)$。为同时支持文本响应生成和细粒度分数预测，EditHF 采用自适应解码策略。首先，文本解码器对最后一层隐藏状态 $\mathbf{h}$ 进行解码，使模型能够初步理解内容并生成便于人类理解的响应。文本输出稳定后，再将相同的隐藏状态输入带有参数 ω的 MLP 头 S_ω，生成与三个维度对应的质量分数 s_v, s_e, s_p。整个过程可形式化表示为:

$$s_v, s_e, s_p = S_\omega(H_\psi(P_\varphi(E_v(I_s, I_e)), T_p)), \tag{4}$$

其中，参数 φ、ψ、ω 在训练过程中进行优化。输出分数用于评估感知质量、编辑对齐度和属性保留情况，也可进一步用于成对偏好比较。

4.2 训练策略

我们的训练分为三个阶段：**文本学习、逐点学习和成对学习**，每个阶段都采用针对不同目标的特定损失函数。

文本学习作为初始阶段，通过文本类- 别使模型对图像编辑质量形成粗粒度理解。连续的质量分数被转换为离散的、基于文本的质量等级，利用了多模态大语言模型通常比对原始数值具有更强文本信息理解能力这一事实。具体而言，我们将最高分 M 与最低分 m 之间的范围均匀划分为五个不同区间，并将每个分数 s 分配至相应的质量等级 $L(s)$：

$$L(s) = l_i \quad \text{if} \quad m + \frac{i-1}{5}(M-m) < s \le m + \frac{i}{5}(M-m) \tag{5}$$

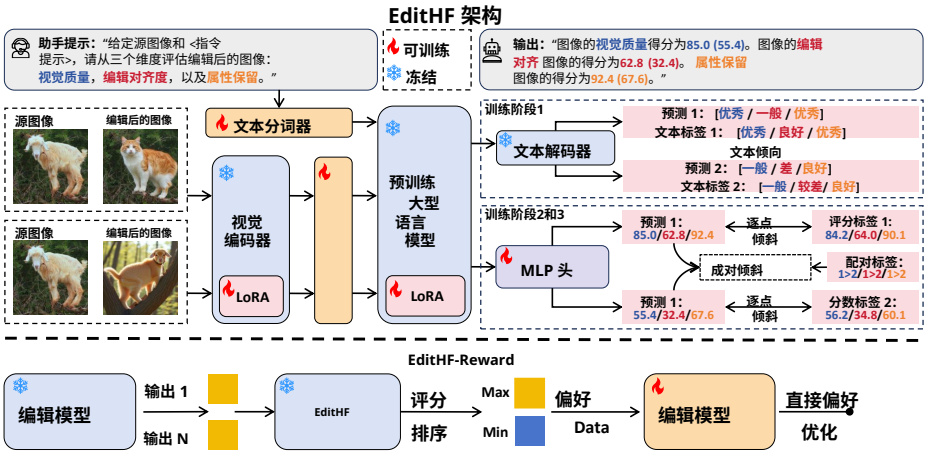


图 4： EditHF 架构概览，以及由 EditHF 引导的编辑模型优化过程。

其中， $i \in \{1,2,3,4,5\}$ 对应于五个标准质量等级， $\{l_i\}_{i=1}^{5}$ ={差、较差、一般、良好、优秀} 为标准五点评分等级。该阶段采用**交叉熵损失**进行训练。

点式学习 旨在赋予模型预测细粒度质量分数的能力。利用 EditHF-1M 中带有人工标注分数的样本，模型学习估计绝对质量值，以反映直接的感知判断。此阶段采用**均方误差损失**进行监督，促使模型尽可能准确地拟合人类提供的分数。

成对学习 侧重于对具有相同源图像和编辑指令的图像进行相对质量评估。模型使用从 EditHF-1M 组内排序中得出的偏好对进行训练。训练时优先处理排名差异较大的对，再处理差异较小的对，从而逐步形成由粗到细的相对质量理解。模型通过以下**成对损失**进行优化：

$$\mathcal{L}_{\text{pairwise}} = \frac{1}{N}\sum_{i=1}^{N} \log\Big(1 + \exp(s_{\text{neg},i} - s_{\text{pos},i})\Big), \tag{6}$$

其中， $s_{\text{pos},\ i}$ 和 $s_{\text{neg},\ i}$ 分别表示第 i 对中偏好图像和较不受偏好图像的预测分数。此损失促使模型为感知质量更高的图像分配更高分数，从而提高模型对相对图像编辑质量的敏感度。

虽然点式训练使模型能够预测反映人类直接感知的绝对质量分数，但它本质上难以感知由相同源图像和编辑指令生成的图像之间细微的相对差异。相反，成对训练侧重于对组内相对偏好进行建模，捕捉相似编辑之间的细粒度差异，但无法生成可在不同源图像或

表2：从视觉质量、编辑对齐度和属性保留三个方面比较各质量评估方法在 EditHF-1M 上的表现。$SRCC_{global}$ 和 $PLCC_{global}$ 用于衡量整个数据集上的相关性。$SRCC_{group}$ 表示在所有图像组上计算的平均 SRCC。准确率表示成对比较预测的准确率。 ♠ 传统全参考图像质量评估指标， ♡传统无参考图像质量评估指标， ♣ 基于深度学习的全参考图像质量评估方法， ◇基于深度学习的无参考图像质量评估方法，★ 视觉语言方法， ☆ 基于多模态大语言模型的模型。微调结果以 *标记。最佳结果以**红色**突出显示，次佳结果以**蓝色**突出显示。

维度	视觉质量				编辑对齐度				属性保留			
方法/指标	$SRCC_{global}$	$PLCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$PLCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$PLCC_{global}$	$SRCC_{group}$	Acc
♡DIIVINE [49]	0.2062	0.3162	0.1795	0.5633	0.0618	0.1947	0.0962	0.5332	0.1641	0.2686	0.1506	0.5522
♡BRISQUE [46]	0.2446	0.2996	0.2020	0.4286	0.0494	0.0975	0.0495	0.4827	0.2019	0.2457	0.1654	0.4426
♡NIQE [47]	0.2816	0.3145	0.2386	0.4142	0.0635	0.1432	0.0811	0.4723	0.2623	0.2778	0.2405	0.4157
♠MSE	0.1421	0.1418	0.2900	0.3968	0.0600	0.0601	0.1730	0.4412	0.3108	0.1893	0.4736	0.3268
♠峰值信噪比	0.1270	0.0232	0.2844	0.6011	0.0572	0.0134	0.1694	0.5575	0.2978	0.1851	0.4690	0.6708
♠梯度幅值相似性偏差 [76]	0.0176	0.0222	0.2715	0.4019	0.0527	0.0486	0.1903	0.4339	0.1926	0.1398	0.4750	0.3265
♠多尺度结构相似性 [65]	0.1023	0.1041	0.2360	0.5821	0.0894	0.0915	0.1768	0.5597	0.2914	0.2490	0.4563	0.6628
♠结构相似性 [64]	0.0807	0.0762	0.2454	0.5868	0.0688	0.0646	0.1641	0.5557	0.2448	0.2107	0.4480	0.6605
♠结构内容相似性 [14]	0.0398	0.0226	0.0867	0.5283	0.0337	0.0329	0.0567	0.5187	0.1140	0.1054	0.3029	0.6053
♠视觉信息保真度 [57]	0.1769	0.0134	0.3703	0.6343	0.2009	0.0922	0.3217	0.6133	0.3367	0.1167	0.5625	0.7087
♠特征相似性 [84]	0.1777	0.1840	0.3357	0.6219	0.1350	0.1461	0.2461	0.5846	0.3683	0.3271	0.5387	0.6993
◇卷积神经网络图像质量评估* [24]	0.6438	0.5824	0.6394	0.7609	0.1927	0.2066	0.1719	0.5605	0.3229	0.2153	0.3086	0.6149
◇MANIQA* [79]	0.8150	0.8104	0.7893	0.8152	0.3068	0.3381	0.2569	0.5907	0.5090	0.5192	0.4968	0.6853
◇TOPIQ* [4]	0.7926	0.8007	0.7359	0.7823	0.2977	0.3175	0.2630	0.5926	0.6006	0.6403	0.5314	0.6936
◇Q-align* [68]	**0.8285**	**0.8403**	0.7659	0.7948	0.3322	0.3520	0.3340	0.6189	0.6401	0.7023	0.5639	0.7049
♣LPIPS [85]	0.2760	0.2629	0.3664	0.3666	0.1783	0.1690	0.2920	0.3988	0.4783	0.4192	0.6007	0.2758
♣ST-LPIPS [12]	0.2896	0.2953	0.3763	0.3611	0.1825	0.1986	0.2903	0.3992	0.4919	0.4376	0.6082	0.2710
♣CVRKD* [82]	0.7649	0.7777	0.7430	0.7917	0.3282	0.3344	0.2943	0.6039	0.7831	0.7989	0.7931	0.8155
♣AHIQ* [30]	0.8172	0.8249	0.8219	0.8326	0.3538	0.3680	0.3254	0.6157	0.8110	0.8210	0.8223	0.8308
★CLIPScore [15]	0.0348	0.0487	0.0879	0.4659	0.1430	0.1585	0.1085	0.5377	0.0422	0.0569	0.1035	0.4597
★BLIPScore [33]	0.1158	0.1235	0.0171	0.4921	0.2562	0.2580	0.2185	0.5805	0.1330	0.1376	0.0345	0.4861
★ImageReward [72]	0.0973	0.1019	0.0555	0.5188	0.2017	0.1894	0.2340	0.5853	0.1092	0.1195	0.0269	0.5062
★HPSv2 [70]	0.1128	0.1291	0.0280	0.5098	0.2912	0.3216	0.2429	0.5851	0.2278	0.2484	0.1643	0.5571
★VQAScore [31]	0.0904	0.0930	0.0029	0.4999	0.2763	0.2850	0.2536	0.5900	0.3573	0.3754	0.2910	0.6032
☆LLaVA-1.5（7B） [38]	0.0700	0.0942	0.0729	0.1677	0.0253	0.0271	0.0163	0.0396	0.1463	0.1548	0.1122	0.1100
☆mPLUG-Owl3（7B） [81]	0.1435	0.0801	0.1242	0.5338	0.1173	0.0645	0.1601	0.5606	0.1962	0.0331	0.2368	0.5932
☆MiniCPM-V2.6（8B） [80]	0.1482	0.0913	0.1224	0.3648	0.2283	0.0220	0.2473	0.4098	0.2765	0.0671	0.2479	0.4029
☆Ovis2.5（8B） [40]	0.2011	0.0984	0.1858	0.4255	0.2274	0.0377	0.3004	0.4647	0.2467	0.0716	0.2385	0.4376
☆LLama3.2-V（11B） [45]	0.4004	0.2283	0.3115	0.2813	0.3743	0.3334	0.3337	0.2071	0.2181	0.1944	0.0945	0.1526
☆DeepSeekVL2（小型） [37]	0.1634	0.1081	0.0824	0.3440	0.1751	0.0663	0.1491	0.3335	0.2562	0.1232	0.1781	0.3408
☆Qwen2.5-VL（8B） [1]	0.2818	0.3065	0.3362	0.4033	0.4528	0.4572	0.6193	0.5195	0.4664	0.4278	0.5871	0.4869
☆Qwen3-VL（8B） [77]	0.3256	0.2263	0.2982	0.5096	0.4578	0.4606	0.6237	0.4990	0.4791	0.4415	0.6296	0.4884
☆InternVL3（8B） [86]	0.2462	0.2995	0.2947	0.4710	0.4649	0.4354	0.5853	0.5851	0.3604	0.3095	0.4335	0.5512
☆InternVL3.5 (8B) [63]	0.3607	0.3055	0.3976	0.3148	0.4671	0.3700	0.6051	0.4103	0.5065	0.3324	0.5887	0.4077
☆EditScore (Qwen3) [41]	0.3678	0.3297	0.3730	0.6440	0.3700	0.2657	0.4149	0.6753	0.4110	0.3054	0.4639	0.6863
☆EditReward（MiMo） [69]	0.3940	0.3320	0.4321	0.3412	0.3410	0.3290	0.3705	0.3657	0.3941	0.3544	0.4266	0.3449
☆DeepSeekVL2（小型）* [37]	0.7931	0.7771	0.8111	0.8330	0.8176	0.8077	0.8273	0.8341	0.7790	0.7984	0.8870	0.8693
☆Qwen3-VL（8B）* [77]	0.7751	0.7673	**0.8674**	0.8588	0.8246	0.8137	0.8498	0.8472	**0.8173**	**0.8342**	**0.9167**	**0.8904**
☆InternVL3（8B）* [86]	0.8044	0.7852	0.8284	**0.8817**	**0.8248**	**0.8156**	**0.8540**	**0.8498**	0.8120	0.8297	0.9137	0.8883
EditHF（我们的方法）*	**0.8548**	**0.8422**	**0.9223**	**0.9152**	**0.8488**	**0.8348**	**0.9319**	**0.9298**	**0.8475**	**0.8442**	**0.9251**	**0.9197**

编辑指令之间进行比较的绝对分数。通过结合这两种训练范式，我们的模型同时具备绝对分数预测和相对质量比较的能力。这种协同训练策略确保模型既能作为可靠的评估基准，也能作为奖励模型，用于图像编辑中以偏好为导向的优化。

5 实验

在本节中，我们通过大量实验评估 EditHF 的性能。

5.1 实验设置

我们在 EditHF-1M 上训练 EditHF，该数据集按 5:1:1 的比例划分为训练集、验证集和测试集。测试集中的每张图像都包含排序和

评分标注。我们采用 LoRA [16]进行微调，并将 LoRA 模块应用于视觉模型和语言模型。模型使用 PyTorch 实现，并在 2 张 NVIDIA RTX A6000 GPU 上进行训练，批大小为 16。

5.2 在 EditHF-1M 上的评估

表 2 展示了 EditHF 与传统及基于深度学习的全参考图像质量评估和无参考图像质量评估方法、视觉语言模型以及基于多模态大语言模型的方法的性能比较。我们报告整个测试集上的斯皮尔曼等级相关系数（SRCC）和皮尔逊线性相关系数（PLCC），以评估总分预测，分别记为 SRCC^{global} 和PLCC^{global}。为评估成对比较，我们报告共享相同源图像和编辑提示词的图像组的平均 SRCC 和比较准确率，分别记为 SRCC^{group} 和准确率。

传统图像质量评估方法表现较差，因为其特征主要用于捕捉低层次失真，难以应对高层次语义变化。基于深度学习的图像质量评估方法在视觉质量评估方面表现更好，但仍难以评估编辑对齐度。视觉语言预训练模型的效果有限，因为它们侧重于文本-图像对齐，无法捕捉编辑语义。基于多模态大语言模型的方法在零样本条件下表现也较差，这是由于它们对编辑评估任务理解不足。EditScore [41] 和 EditReward [69], 虽在其他图像编辑数据集上进行训练，但表现仍然有限，因为其训练数据仅涵盖有限范围的图像内容、编辑任务和编辑模型，导致泛化能力不足。我们的 EditHF 在所有评估维度上与人类感知的对齐效果最佳。表2 进一步表明，在我们的 EditHF-1M 上微调的多模态大语言模型取得了显著提升，证明了该数据集对于推动图像编辑质量评估发展的价值。

表3 证明了我们的模型在对图像编辑模型进行基准评测方面的有效性。我们的模型与人类分数的 SRCC 最高，展现出其在评测图像编辑模型方面的强大能力。此外，结果显示，NanoBanana [7],Seedream4 [56], 和 FLUX-Kontext [29],等闭源图像编辑方法与开源模型之间存在明显的性能差距，这促使我们进一步利用源自闭源方法的数据来改进开源模型。

5.3 消融研究

我们进行了全面的消融研究，结果如表 4所示。第 1–2 行说明了 MLP 头对于精确分数回归的重要性。第 2–4 行表明，结合逐点训练和成对训练能够取得最佳性能。第 4–8 行比较了参数规模相近的不同多模态大语言模型骨干网络，其中 InternVL3.5 的表现最佳。

表 3： 比较不同评估方法与人类感知在评估图像编辑模型时的对齐程度。最佳结果以**红色**标出，次佳结果以**蓝色**标出。 * 表示经过微调的模型。

维度	感知质量				编辑对齐度				属性保留				总体分数				总排名	
模型/指标	人类	我们	* Q-align	EditReward	人类	我们	* Qwen3-VL	EditReward	人类	我们	AHIQ	EditScore	人类	我们	EditScore	EditReward	人类	我们
NanoBanana [7]	17.54	17.61	16.20	5.863	17.01	17.97	17.24	6.746	18.56	18.43	18.65	18.85	17.68	17.70	5.608	1.810	1	1
Qwen-Image-Edit [66]	15.26	16.53	14.51	6.453	18.17	18.26	17.38	6.952	17.51	17.86	19.95	18.19	17.15	17.33	5.463	1.799	2	2
Seedream4 [56]	15.94	16.77	14.85	6.377	17.31	18.06	17.31	7.013	16.57	17.38	19.34	17.85	16.74	17.19	5.352	1.799	3	4
FLUX-Kontext [29]	15.12	16.37	14.87	6.602	17.12	18.04	17.13	7.126	17.40	18.06	20.13	17.05	16.65	17.25	5.259	1.794	4	3
DreamOmni [71]	14.33	15.91	13.57	6.658	16.14	17.45	16.69	7.109	18.08	18.22	20.06	16.59	16.19	16.93	5.234	1.794	5	5
Follow-Your-Shape [39]	14.85	16.41	14.81	10.29	12.94	16.53	14.81	11.01	14.24	15.80	16.20	13.78	13.92	16.01	4.664	1.695	6	6
OmniGen2 [78]	11.89	15.16	12.60	10.21	15.25	16.86	16.33	10.32	13.26	15.67	18.39	13.41	13.66	15.73	4.581	1.704	7	7
Reflex [24]	12.03	15.47	13.03	10.87	14.83	16.59	15.89	11.08	11.11	14.48	18.07	11.99	12.89	15.36	4.413	1.686	8	8
CDS [50]	13.11	15.59	13.29	11.57	11.07	14.71	14.15	11.95	14.65	16.50	12.47	10.59	12.71	15.23	4.370	1.665	9	9
InfEdit [73]	13.10	15.49	13.52	10.98	10.73	14.30	13.79	10.86	14.80	16.25	14.40	12.78	12.60	14.96	4.552	1.687	10	10
FlowEdit (SD3) [28]	12.07	15.38	12.56	11.22	13.04	14.82	12.86	11.45	11.43	15.20	14.46	13.37	12.30	14.85	4.542	1.676	11	11
Bagel [9]	13.33	15.67	13.23	11.30	10.46	14.26	13.61	11.83	13.41	15.76	15.46	8.310	12.16	14.86	4.040	1.670	12	12
RFSE [61]	12.44	15.62	12.91	11.11	12.86	15.11	14.63	11.40	10.81	14.10	17.27	11.58	12.15	14.71	4.100	1.678	13	14
FlowEdit (FLUX) [28]	12.21	15.46	12.94	10.70	12.26	15.01	14.33	10.99	11.63	14.50	13.97	12.17	12.09	14.74	4.461	1.689	14	13
ZONE [34]	12.17	15.19	12.54	10.95	9.934	14.17	13.28	11.59	14.30	16.21	12.42	12.62	11.84	14.81	4.472	1.678	15	15
OT-Inversion [42]	12.48	15.62	12.44	12.26	12.14	14.91	14.32	12.58	10.39	14.06	16.54	10.47	11.74	14.61	3.948	1.648	16	16
Any2Pix [35]	9.270	14.12	10.06	14.93	11.64	14.41	13.89	14.79	8.384	12.49	17.32	8.400	9.951	13.51	3.459	1.583	17	17
MagicBrush [83]	11.69	14.02	12.36	11.41	8.504	10.77	10.31	10.75	9.693	13.10	12.22	12.34	9.772	12.16	4.388	1.682	18	18
ACE++ [44]	8.925	13.52	9.703	13.86	9.781	11.86	12.67	13.76	8.905	9.961	11.79	8.760	9.278	11.54	3.832	1.610	19	19
IP2P [3]	8.557	12.93	9.904	11.33	8.637	10.10	9.751	11.15	6.968	9.844	11.38	13.35	8.116	10.63	4.499	1.678	20	20
HQEdit [18]	8.834	14.32	10.05	20.20	6.175	8.588	8.420	20.20	5.275	9.011	7.398	4.040	6.641	10.08	2.655	1.441	21	21
PnP [21]	6.608	12.23	8.570	11.90	5.917	9.502	9.171	11.21	5.416	9.431	9.591	12.97	5.959	10.07	4.523	1.669	22	22
ReNoise [11]	1.544	1.402	3.446	12.52	2.367	2.193	2.400	11.19	1.719	1.572	1.377	12.68	1.892	1.736	4.361	1.661	23	23
SRCC 至人类 ↑		**0.976**	**0.951**	0.766		**0.987**	**0.952**	0.632		**0.996**	**0.734**	0.612		**0.998**	0.755	**0.833**		**0.998**

表 4： 不同骨干网络和训练策略的消融研究。

骨干网络与策略				感知质量			编辑对齐度			属性保留		
骨干网络	MLP 头	逐点训练	成对训练	$SRCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$SRCC_{group}$	Acc
InternVL3.5 [63]		✓		0.7184	0.7215	0.7904	0.7025	0.7312	0.7750	0.7241	0.7506	0.7862
InternVL3.5 [63]	✓	✓		0.7762	0.7829	0.7826	0.7806	0.7964	0.7825	0.7752	0.7864	0.7821
InternVL3.5 [63]	✓		✓	0.8018	0.8792	0.8663	0.8213	0.8894	0.8762	0.8173	0.8726	0.8765
InternVL3.5 [63]	✓	✓	✓	**0.8548**	**0.9223**	**0.9152**	**0.8488**	**0.9319**	**0.9298**	**0.8475**	**0.9251**	**0.9197**
InternVL3 [5]	✓	✓	✓	0.8316	0.8964	0.8826	0.8155	0.9109	0.9035	0.8141	0.9022	0.8943
Qwen3-VL [77]	✓	✓	✓	0.8417	0.9032	0.8985	0.8231	0.9127	0.9046	0.8240	0.9068	0.9074
DeepSeekVL2 [37]	✓	✓	✓	0.8274	0.8823	0.8970	0.8285	0.8963	0.8856	0.8106	0.8769	0.8811
LLaVA-NeXT [32]	✓	✓	✓	0.8164	0.8723	0.8657	0.8108	0.8814	0.8741	0.8037	0.8806	0.8794

5.4 在其他基准测试上的零样本性能

表 5 展示了 EditHF 与其他先进的闭源和开源多模态大语言模型在多个图像编辑基准上的性能。对于具有成对标注的基准，包括 GenAI-Bench [20], AROURA-Bench [26], 和 EditScore-Bench [41],，我们报告预测准确率。对于具有逐点评注的基准，包括 ImagenHub [27],EBench-18K [74],和 EditReward-Bench [69], ，我们报告与人工评分的 SRCC。对于提供多个评估维度的基准，包括 ImagenHub [27],EBench-18K [74], EditScore-Bench [41], 和 EditReward-Bench [69], ，我们报告不同维度的平均结果。结果表明，EditHF 在各种基准上始终表现出色，展现了强大的泛化能力。

表 5： 在其他公开图像编辑基准上的对比结果。最佳结果以**红色**标示，次佳结果以**蓝色**标示。**“-”**表示该模型在此数据集上进行了训练，因此未报告其结果，以确保跨数据集评估的公平性。

方法/基准测试	GenAI-Bench [20]	AROURA-Bench [26]	ImagenHub [27]	EBench-18K [74]	EditScore-Bench [41]	EditReward-Bench [69]
GPT-4o [51]	0.5354	0.5081	0.3821	0.4108	0.6730	0.2831
GPT-5 [52]	0.5961	0.4727	0.4085	0.4260	**0.7524**	0.3781
Gemini-2.0-Flash [13]	0.5332	0.4431	0.2369	0.3892	0.6958	0.3347
Gemini-2.5-Flash [7]	0.5701	0.4763	**0.4162**	**0.4337**	0.7026	0.3802
Qwen3-VL（7B） [77]	0.4458	0.3428	0.2426	0.3726	0.4251	0.2731
InternVL3.5（7B） [63]	0.4517	0.3692	0.2071	0.3511	0.4518	0.2406
EditScore（Qwen3） [41]	0.5024	0.4218	0.3062	0.3206	-	0.2590
EditReward（MiMo） [69]	**0.6572**	0.6362	0.3520	0.3718	0.7125	-
LMM4Edit [74]	0.6327	**0.6521**	0.3726	-	0.7064	**0.3958**
EditHF（我们的方法）	**0.6792**	**0.6608**	**0.4526**	**0.4627**	**0.7750**	**0.4325**

表 6： Qwen-Image-Edit 的结果 [66] 使用我们的 EditHF 在不同基准测试上进行优化。报告了人工评估和我们的 EditHF 得出的总分。

基准测试	EditHF-1M		AROURA-Bench [26]		ImagenHub [27]		EBench-18K [74]		EditScore-Bench [41]		EditReward-Bench [69]	
方法/指标	人类	Ours	人类	Ours	人类	Ours	人类	Ours	人类	Ours	人类	Ours
Qwen-Image-Edit	75.39	78.06	83.21	68.47	75.64	73.26	73.25	78.29	70.17	75.22	80.15	77.03
Qwen-Image-Edit（自偏好直接优化）	78.36	81.55	84.28	69.26	76.55	75.14	73.80	79.05	72.18	76.55	81.17	78.46
提升	+3.94%	+4.47%	+1.29%	+1.15%	+1.20%					+1.77%	+1.27%	+1.86%
Qwen-Image-Edit（Global-DPO）	84.26	86.15	85.46	69.12	79.12	78.06	74.18	79.60	75.23	78.41	82.60	78.12
提升	+11.77%			+0.95%	+4.60%					+4.24%	+3.06%	+1.42%

5.5 EditHF 奖励建模

为验证 EditHF 作为奖励模型优化图像编辑模型的有效性，我们将其用于优化 Qwen-Image-Edit [66]，该模型在所有开源编辑模型中，在 EditHF-1M 上取得了最佳性能。

我们首先从 EditHF-1M 训练集中涵盖的所有编辑任务中，选取源图像及其对应的指令提示词。对于每条指令，我们使用 Qwen-Image-Edit 通过 20 个不同的随机种子生成编辑后图像，随后使用 EditHF 对所有生成结果进行评估。在每个随机种子组内，我们将所有评估维度上的平均分数最高的样本指定为选中样本，并将排名最低的样本指定为拒绝样本。为避免从过于模糊或容易失败的案例中构建偏好，我们排除编辑结果始终获得较低总分的指令。此外，在每个编辑任务内，我们进一步根据选中样本的总分对所得偏好对进行排序，并舍弃排名后半部分，从而确保仅保留相对高质量的编辑结果用于构建偏好。最终，我们获得了 18K 个高质量偏好对。

随后，我们采用直接偏好优化（DPO），使 Qwen-Image-Edit 的输出与人类偏好保持一致。具体而言，我们冻结 VAE模块，并通过比较原始预训练 Qwen-Image-Edit（作为参考模型）的流预测误差来优化 Transformer，如下所示：

$$\begin{aligned} \mathcal{L}(\theta) &= -\mathbb{E}_{(x_0^c, x_0^r)\sim\mathcal{D}_{\text{Gen}},\, t\sim\mathcal{U}(0,T)}\Big[\log\sigma\big(\Delta_\theta(x_t^c, t) - \Delta_\theta(x_t^r, t)\big)\Big], \\ \Delta_\theta(x_t, t) &= \beta_g\Big(\|(\epsilon - x_0) - v_\theta(x_t, t)\|_2^2 - \|(\epsilon - x_0) - v_{\text{ref}}(x_t, t)\|_2^2\Big). \end{aligned} \tag{7}$$

其中，ϵ 表示从标准正态分布中采样的高斯噪声，用于扰动干净图像 x_0。x_t^c 和 x_t^r 分别表示时间步 t处选中样本和拒绝样本的含噪潜变量。$v_\theta(\cdot)$和 $v_{\text{ref}}(\cdot)$分别表示微调后的编辑模型和参考编辑模型的流预测。β_g 是控制优化强度的温度参数，$\sigma(\cdot)$表示逻辑函数。该损失促使微调后的编辑模型降低选中样本的去噪误差，同时提高拒绝样本的去噪误差，并以微调前的编辑模型为参照，从而增强训练稳定性。

依赖模型生成自身的奖励信号，可能会强化其固有偏见，并忽视其他高质量输出。为了获得更多样且可靠的监督信号，我们使用 EditHF-1M 中多个编辑模型的输出构建选中样本和拒绝样本。具体而言，对于每个包含不同模型在相同

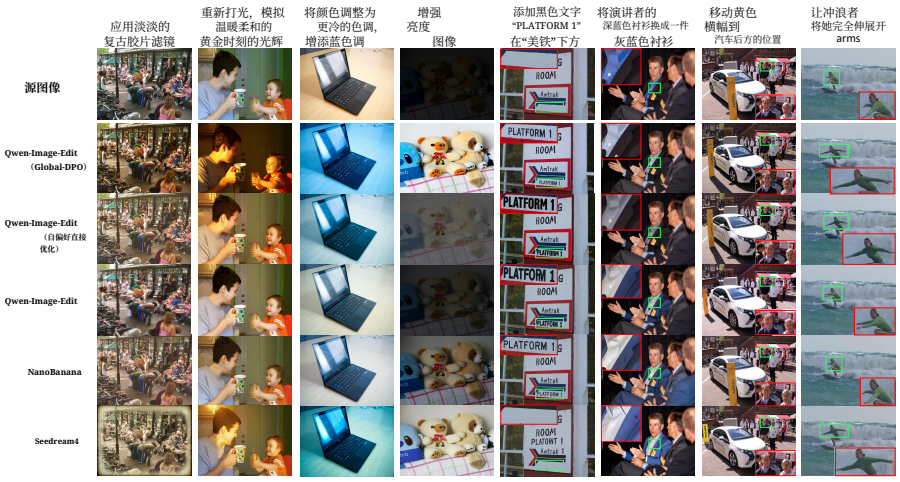


图 5：由我们的 EditHF 和其他具有竞争力的编辑模型优化后的 Qwen-Image-Edit 示例。

源图像和指令下生成的编辑结果的图像组，我们将排名最高的图像指定为选中样本，将排名最低的图像指定为拒绝样本。为避免难度过高或含义模糊的情况，我们排除所有编辑模型总体分数较低的指令。通过这一流程，我们构建了 22 K 个偏好对，并采用上述相同的训练策略对 Qwen-Image-Edit 进行微调。

随后，我们在多个图像编辑基准上，将原始 Qwen-Image-Edit 与使用其自身输出（Self-DPO）以及使用多个模型的输出（Global-DPO）进行优化的版本进行比较。我们按照第 3.3 节所述的评分标注相同流程开展主观研究。表 6 中的结果表明，使用我们的 EditHF 优化后的 Qwen-Image-Edit 在这些基准上的人类评估和 EditHF 指标方面均持续取得性能提升，其中使用 Global-DPO 时提升更为显著。图 5 展示了优化后的 Qwen-Image-Edit 生成的代表性编辑示例，以及其他具有竞争力的编辑模型生成的示例，表明我们优化后的模型能够出色地处理细粒度编辑任务，并具有更高的保真度和与给定指令更好的对齐。

6 结论

本文介绍了 EditHF-1M，这是首个用于图像编辑多维评估的百万规模基准，包含偏好对和通过严格人工标注获得的细粒度评分。基于这一基准，我们提出了 EditHF，可有效支持逐点评分和成对比较。大量实验表明，EditHF 优于现有方法，与人类偏好的对齐程度更高，且泛化能力更强。

此外，我们证明 EditHF 可作为有效的奖励模型，用于改进图像编辑模型。

参考文献

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., et al.：Qwen2.5-vl 技术报告。arXiv 预印本 arXiv:2502 11 2 13923 2025。（）. Basu, S., Saberi, M., Bhardwaj, S., Chegini, A.M., Massiceti, D., Sanjabi, M., et al.：EditVal：扩散式文本引导图像编辑方法基准测试。arXiv 预印本 arXiv:2310 2, 4 3 02426 2023。（）. Brooks, T., Holynski, A., 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（CVPR）论文集。第 18392 1, 4, 13 4 18402 页，2023–（）. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.：TOPIQ：一种从语义到失真的图像质量评估自顶向下方法。IEEE图像处理汇刊（TIP）33，2404 11 5 2418，2024–（）. Chen, Z., Wang, W., Cao, Y., Liu, Y., 高 Z., Cui, E., et al.：通过模型、数据和测试时扩展，拓展开源多模态模型的性能边界。arXiv 预印本 arXiv:2412 13 6 05271 2025。（）. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., et al.：InternVL：扩展视觉基础模型，并使其适配通用视觉语言任务。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第 24185 9 7 24198 页，2024–（）. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., et al.：Gemini 2.5：以高级推理、多模态、长上下文和新一代智能体能力推动前沿发展。arXiv 预印本 arXiv:2507 1, 3, 4, 8, 12, 13 8 06261 2025。（）. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.：基于退化模型的图像质量评估。IEEE图像处理汇刊（TIP）9 2 9 4 636 650 2000（），–（）. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., et al.：统一多模态预训练中的新兴特性。arXiv 预印本 arXiv:2505 13 10 14683 2025。（）. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., et al.：扩展整流流变换器，实现高分辨率图像合成。载于：国际机器学习会议（ICML）论文集（2024 4 11）. Garibi, D., Patashnik, O., Voynov, A., Averbuch-Elor, H., Cohen-Or, D.：ReNoise：通过迭代加噪实现真实图像反演。载于：欧洲计算机视觉会议（ECCV）。第 395 4, 13 12 413 页，2024–（）. Ghildyal, A., Liu, F.：平移容忍感知相似度指标。载于：欧洲计算机视觉会议（ECCV）论文集。第 91 4, 11 13 107 页，2022 2 0 2024–（）. Google：Gemini。。https://deepmind.google/technologies/gemini/（）13 14. Gu, K., Zhai, G., Yang, X., Zhang, W.：一种基于结构补偿的改进全参考图像质量指标。载于：亚太信号与信息处理协会（APSIPA）会议论文集。第 1–6 页（2012）11 15. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.：CLIPScore：一种用于图像描述生成的无参考评估指标。arXiv 预印本 arXiv:2104.08718（2021）2, 3, 4, 11 16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al.：LoRA：大语言模型的低秩适配。载于：国际学习表征会议（ICLR）论文集（2022 12 17）. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., et al.：基于扩散模型的图像编辑：综述。IEEE模式分析与机器智能汇刊（TPAMI）PP（2024）2, 3, 4

18. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., et al.：HQ-Edit：用于基于指令的图像编辑的高质量数据集。arXiv 预印本 arXiv:2404 13 19 09990 2024。（）. (ITU), I.T.U.：电视画面质量主观评估方法。技术报告 Rec. ITU-R BT.500-13，国际电信联盟（ITU）（2012年1月 7 20）. Jiang, D., Ku, M., Li, T., Ni, Y., 孙 S., Fan, R., et al.：Genai arena：生成模型的开放评估平台。arXiv 预印本 arXiv:2406.04485（2024） 13 21. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.：PnP反演：用三行代码增强基于扩散的编辑。载于：国际学习表征会议（ICLR）论文集（2024 4, 13 22）. Kang, L., Ye, P., Li, Y., Doermann, D.：用于无参考图像质量评估的卷积神经网络。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2014 4, 11 23）. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., et al.：Imagic：使用扩散模型进行真实图像的文本编辑。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2023 2, 4 24）. Kim, J., Park, J., Song, Y., Kwak, N., Rhee, W.：Reflex：通过中间步骤特征提取和注意力自适应，在整流流中进行真实图像的文本引导编辑。arXiv 预印本 arXiv:2507 13 25 01496 2025。（）. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.：Pick-a-pic：一个用于文本到图像生成的用户偏好开源数据集。载于：神经信息处理系统进展（NeurIPS）论文集。第 36652–36663 页 4 26 2023（）. Krojer, B., Vattikonda, D., Lara, L., Jampani, V., Portelance, E., Pal, C., et al.：从视频和模拟中学习以动作和推理为中心的图像编辑。载于：神经信息处理系统进展（NeurIPS）论文集（2024 2, 13, 14 27）. Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., et al.：ImagenHub：规范化条件图像生成模型的评估。载于：国际学习表征会议（ICLR）论文集（2024）2, 13, 14 28. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.：Flowedit：使用预训练流模型进行无反演的基于文本编辑。载于：IEEE/CVF国际计算机视觉会议（ICCV）论文集。第 19721–19730 页 4, 13 29 2025（）. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., et al.：Flux.1 kontext：在潜在空间中进行上下文内图像生成与编辑的流匹配。arXiv 预印本（2025 12, 13 30）. Lao, S., Gong, Y., Shi, S., Yang, S., Wu, T., Wang, J., et al.：注意力帮助CNN更好地视觉感知：基于注意力的混合图像质量评估网络。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集。第 1140 4, 11 31 1149 页，2022–（）. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., et al.：评估并改进组合式文本到视觉生成。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2024 11 32）. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., et al.：Llava-next-interleave：解决大型多模态模型中的多图像、视频和3D问题。arXiv 预印本 arXiv:2407.07895（2024） 13

33. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP：用于统一视觉语言理解与生成的自举式语言-图像预训练。载于：国际机器学习会议（ICML）论文集。第12888–12900页（2022）2, 3, 4, 11 34. Li, S., Zeng, B., Feng, Y., Gao, S., Liu, X., Liu, J., 等：ZONE：零样本指令引导局部编辑。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第6254–6263页（2024）2, 13 35. Li, S., Singh, H., Grover, A.: Instructany pix：通过多模态指令跟随实现灵活的视觉编辑。arXiv预印本 arXiv:2312.06738（2023）4, 13 36. Li, Y., Liu, H., Cai, M., Li, Y., Shechtman, E., Lin, Z., 等：消除描述中的分布差异可改善图像-文本对齐。arXiv预印本 arXiv:2410.00905（2024）3, 4 37. Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., 等：DeepSeek-V2：一种强大、经济且高效的混合专家语言模型。arXiv预印本 arXiv:2405.04434（2024）11, 13 38. Liu, H., Li, C., Li, Y., Lee, Y.J.: 通过视觉指令微调改进基线。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第26296–26306页（2024）11 39. Long, Z., Zheng, M., Feng, K., Zhang, X., Liu, H., Yang, H., 等：Follow-your-shape：通过轨迹引导区域控制实现形状感知图像编辑。arXiv预印本 arXiv:2508.08134（2025）13 40. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., 等：Ovis2.5技术报告。arXiv:2508.11737（2025）11 41. Luo, X., Wang, J., Wu, C., Xiao, S., Jiang, X., Lian, D., 等：EditScore：通过高保真奖励建模解锁图像编辑的在线强化学习。arXiv预印本 arXiv:2509.23909（2025）2, 3, 4, 11, 12, 13, 14 42. Lupascu, M., Stupariu, M.S.: 面向整流流图像编辑的最优传输：统一基于反演的方法与直接方法。arXiv预印本 arXiv:2508.02363（2025）13 43. Ma, Y., Ji, J., Ye, K., Lin, W., Zheng, Y., Zhou, Q., 等：I2EBench：一个全面的基于指令的图像编辑基准。载于：神经信息处理系统进展（NeurPIS）论文集（2024）2, 3 44. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++：通过上下文感知内容填充实现基于指令的图像生成与编辑。arXiv预印本 arXiv:2501.02487（2025）1, 13 45. Meta, A.: Llama：以开放、可定制的模型革新边缘人工智能与视觉。Meta人工智能博客。访问于2024年12月（2024）11 46. Mittal, A., Moorthy, A.K., Bovik, A.C.: 盲式/无参考图像空间质量评估器。载于：阿西洛马信号、系统与计算机会议（ACSSC）论文集。第723–727页（2011）2, 4, 11 47. Mittal, A., Soundararajan, R., Bovik, A.C.: 构建“完全盲式”图像质量分析器。IEEE信号处理快报（SPL）20(3)，209–212（2013）4, 11 48. Moorthy, A.K., Bovik, A.C.: 构建盲式通用质量指标的模块化框架。IEEE信号处理快报（SPL）（2009）2 49. Moorthy, A.K., Bovik, A.C.: 盲图像质量评估：从自然场景统计到感知质量。IEEE图像处理汇刊（TIP）20(11)，3350–3364（2011）12 50. Nam, H., Kwon, G., Park, G.Y., Ye, J.C.: 用于文本引导潜在扩散图像编辑的对比去噪评分。载于：IEEE/CVF计算机视觉与模式识别会议

（CVPR）论文集。第9192–9201页（2024年6月）4, 13 51. OpenAI：ChatGPT-4o：先进的多模态聊天模型（2025），https://openai.com/chatgpt 13 52. OpenAI：GPT-5。https://www.openai.com（2025）13 53. Qian, Y., Bocek-Rivele, E., Song, L., Tong, J., Yang, Y., Lu, J., 等：Pico-banana-400k：一个用于文本引导图像编辑的大规模数据集。arXiv预印本 arXiv:2510.19808（2025）3, 5 54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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（CVPR）论文集。第10684–10695页（2022）4 55. Saad, M.A., Bovik, A.C., Charrier, C.: 盲图像质量评估：DCT域中的自然场景统计方法。IEEE图像处理汇刊（TIP）21(4)，3339–3352（2012）8 56. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., 等：Seedream 4.0：迈向新一代多模态图像生成。arXiv预印本（2025）1, 3, 4, 12, 13 57. Sheikh, H.R., Bovik, A.C.: 图像信息与视觉质量。IEEE图像处理汇刊（TIP）15(2)，430–444（2006）11 58. Sheikh, H.R., Bovik, A.C., Veciana, G.D.: 一种利用自然场景统计评估图像质量的信息保真度准则。IEEE图像处理汇刊（TIP）14(12)，2117–2128（2005）2 59. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., 等：在自适应超网络的引导下评估真实场景中的图像质量。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2020年6月）4 60. Sun, S., Qu, B., Liang, X., Fan, S., Gao, W.: IE-Bench：推进文本驱动图像编辑测量，以实现人类感知对齐。arXiv预印本 arXiv:2501.09927（2025）2, 4 61. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., 等：驾驭整流流以实现反演与编辑。arXiv预印本 arXiv:2411.04746（2024）13 62. Wang, P., Sun, W., Zhang, Z., Jia, J., Jiang, Y., Zhang, Z., 等：由人工智能生成图像质量评估辅助的大型多模态模型。载于：ACM多媒体国际会议（ACM MM）论文集。第7803–7812页（2024）4 63. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., 等：InternVL3.5：提升开源多模态模型在通用性、推理与效率方面的表现。arXiv预印本 arXiv:2508.18265（2025）5, 6, 9, 11, 13 64. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: 图像质量评估：从误差可见性到结构相似性。IEEE图像处理汇刊（TIP）13(4)，600–612（2004）2, 11 65. Wang, Z., Simoncelli, E.P., Bovik, A.C.: 用于图像质量评估的多尺度结构相似性。载于：阿西洛玛信号、系统与计算机会议（ACSSC）论文集。第2卷，第1398–1402页（2003）2, 11 66. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., 等：Qwen-Image技术报告。arXiv预印本 arXiv:2508.02324（2025）1, 3, 4, 13, 14 67. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., 等：OmniGen2：探索先进的多模态生成。arXiv预印本 arXiv:2506.18871（2025）3, 4

68. Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., 等：Q-align：通过离散文本定义等级教授多模态大语言模型进行视觉评分。arXiv 预印本 arXiv:2312.17090（2023）11 69. Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., Chen, W.：EditReward：一种与人类偏好对齐的奖励模型，用于指令引导图像编辑。arXiv 预印本 arXiv:2509.26346（2025）2, 3, 4, 11, 12, 13, 14 70. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.：人类偏好分数：更好地使文生图模型与人类偏好对齐。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）（2023）2, 3, 11 71. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., 等：DreamOmni2：基于多模态指令的编辑与生成。arXiv 预印本 arXiv:2510.06679（2025）4, 13 72. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., 等：ImageReward：学习并评估文本到图像生成的人类偏好。载于：神经信息处理系统国际会议（NeurIPS）论文集。第15903–15935页（2023）2, 3, 4, 11 73. Xu, S., Huang, Y., Pan, J., Ma, Z., Chai, J.：利用自然语言进行无反演图像编辑。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第9192–9201页（2024）4, 13 74. Xu, Z., Duan, H., Liu, B., Ma, G., Wang, J., Yang, L., 等：LMM4Edit：使用多模态大语言模型对多模态图像编辑进行基准测试与评估。载于：ACM多媒体国际会议（ACM MM）论文集。第6908–6917页（2025年10月）2, 3, 4, 6, 13, 14 75. Xu, Z., Duan, H., Ma, G., Yang, L., Wang, J., Wu, Q., 等：HarmonyIQA：开创性的图像协调质量评估基准与模型。载于：IEEE国际多媒体与博览会会议（ICME）论文集。第1–6页（2025）4 76. Xue, W., Zhang, L., Mou, X., Bovik, A.C.：梯度幅值相似性偏差：一种高效的感知图像质量指数。IEEE图像处理汇刊（TIP）23(2), 684–695（2013）4, 11 77. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., 等：Qwen3技术报告。arXiv 预印本 arXiv:2505.09388（2025）8, 11, 13 78. Yang, L., Duan, H., Zhu, Y., Liu, X., Liu, L., Xu, Z., Ma, G., Min, X., Zhai, G., Callet, P.L.：Omni^2：在全景模型中统一全景图像生成与编辑（2025）13 79. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., 等：MANIQA：用于无参考图像质量评估的多维注意力网络。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集。第1191–1200页（2022）4, 11 80. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., 等：Minicpm-v：手机上的GPT-4V级多模态大语言模型。arXiv 预印本 arXiv:2408.01800（2024）11 81. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., 等：mPLUG-Owl3：迈向多模态大语言模型中的长图像序列理解。载于：国际学习表征会议（ICLR）论文集（2024）11 82. Yin, G., Wang, W., Yuan, Z., Han, C., Ji, W., 孙 S., 等：通过知识蒸馏进行内容可变的参考图像质量评估。载于：人工智能促进协会（AAAI）会议论文集。第36卷，第3134–3142页（2022）11

83. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.：MagicBrush：一个用于指令引导图像编辑的人工标注数据集。载于：神经信息处理系统进展（NeurIPS）论文集（2023, 4, 13 84）。Zhang, L., Zhang, L., Mou, X., Zhang, D.：特征相似性：用于图像质量评估的特征相似度指数。IEEE图像处理汇刊（TIP）**20**(8),23782, 4, 11 85 2386 2011–（）. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.：深层特征作为感知度量的惊人效力。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（20182, 4, 11 86 3）. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., 等：InternVL：探索开源多模态模型的先进训练与测试时方案。arXiv 预印本 arXiv:2504.10479（2025）11