MoRE-OPD: Mixture of Reward Experts with On-Policy Distillation for Image Editing

Anonymous authors
Paper under double-blind review

Abstract

Reliable evaluation is critical for image editing, supporting data curation, benchmarking, and reinforcement learning. However, diverse editing tasks require distinct criteria, leading to potential criterion entanglement in a shared evaluator. To address this challenge, we propose **MoRE-OPD**, a reward modeling framework that combines a **M**ixture **o**f **R**eward **E**xperts with privileged **O**n-**P**olicy **D**istillation. MoRE-OPD first learns category-specialized reward experts under task-specific evaluation protocols and then distills their heterogeneous evaluation capabilities into a compact universal evaluator. During OPD, specialized experts additionally access oracle-generated rationales as privileged context, enabling more effective supervision of student-generated predictions. Furthermore, we construct **PointReward-Bench** for reliable pointwise evaluation of image editing reward models, and apply MoRE-OPD to curate approximately 3.8 million high-quality samples for downstream optimization. Extensive experiments demonstrate that MoRE-OPD consistently outperforms state-of-the-art open-source image editing reward models across pointwise evaluation, preference alignment, and downstream reinforcement learning.

1 Introduction

Instruction-guided image editing has progressed rapidly, with models such as FLUX.2 (Labs, 2025), Qwen-Image-Edit (Wu et al., 2025a), GPT-Image-2, and Nano Banana Pro producing increasingly photorealistic results. As editing capabilities continue to improve, *reliably evaluating edit success* has become increasingly important. Reliable evaluation supports three critical applications: (i) **Large-Scale Data Curation**, where automated quality filtering is required to construct high-quality training data from millions of editing samples; (ii) **Reinforcement Learning**, where inaccurate or biased reward signals can misguide policy optimization; and (iii) **Benchmarking**, where reliable evaluation is necessary to distinguish genuine improvements in editing capability.

Early attempts adopted text-to-image (T2I) metrics (Radford et al., 2021; Xu et al., 2023; Wu et al., 2023), which are poorly suited to image editing due to their limited awareness of editing instructions and non-edited region preservation. Recent works have therefore developed dedicated reward models for image editing. These efforts broadly advance along three dimensions: learning reward models for semantic and perceptual evaluation or human preference alignment (Wu et al., 2025b; Luo et al., 2026), incorporating explicit reasoning mechanisms via spatial understanding (Long et al., 2026) or multi-dimensional rationales (Wang et al., 2026b; Yang et al., 2026), and developing adaptive evaluation strategies such as dynamic rubrics (Kan et al., 2026) or difference-first evaluation (Zhao et al., 2026). Despite these advances, existing approaches largely rely on a single reward model to handle heterogeneous editing tasks, leading to two key limitations:

- **Criterion entanglement across edit types.** Different edits impose distinct and sometimes conflicting evaluation requirements. For instance, fine-grained localized modifications demand strict content preservation in unedited regions, whereas background replacements permit large-scale visual alterations. A single evaluator must therefore reconcile conflicting criteria across tasks, leading to criterion entanglement and potentially misaligned judgments.
- **Limited supervision from static SFT.** Supervised Fine-Tuning (SFT) relies on fixed teacher-generated annotations and provides no adaptive supervision on the student's own predictions. This

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

off-policy nature introduces a distributional shift, limiting the effective transfer of fine-grained evaluation capabilities from specialized teachers to a universal student.

To address these limitations, we propose **MoRE-OPD**, a framework that combines a **M**ixture **o**f **R**eward **E**xperts with **O**n-**P**olicy **D**istillation under privileged supervision. To disentangle the conflicting evaluation criteria across editing tasks, we first train eight category-specialized 27B reward experts (Qwen3.6-27B), each governed by a category-specific evaluation protocol and supervised with oracle annotations from Qwen3.7-plus. We then consolidate these specialized evaluation capabilities into a single compact 9B student (Qwen3.5-9B) through on-policy distillation. Unlike static SFT, the specialized teachers directly evaluate the student's own predictions, providing adaptive feedback on its on-policy distribution. During distillation, each teacher additionally receives oracle-generated rationales as *privileged context*, enabling more informative supervision of the student's predictions, while the student observes only the standard evaluation inputs. By minimizing the reverse KL divergence between the student and the corresponding specialized teacher, MoRE-OPD transfers heterogeneous evaluation capabilities into a universal reward model without requiring multiple experts or privileged information at inference. Extensive experiments demonstrate that MoRE-OPD consistently outperforms state-of-the-art open-source image editing reward models across pointwise evaluation, preference alignment, and downstream reinforcement learning.

The main contributions of this work are summarized as follows:

- **Mixture of Reward Experts with Privileged On-Policy Distillation.** We propose MoRE-OPD, a novel reward modeling framework that addresses criterion entanglement across heterogeneous image editing tasks. It distills category-specialized reward experts into a compact universal evaluator through privileged on-policy distillation, consolidating heterogeneous evaluation capabilities.
- **Benchmark Construction for Pointwise Evaluation.** We construct **PointReward-Bench** to assess models on independent success/failure judgments of individual edits. Covering diverse categories and representative failure modes, it provides reliable ground-truth labels established by multiple human annotators.
- **Large-Scale Data Curation.** Applying MoRE-OPD to filter public data, we curate a multi-category dataset of approximately 3.8 million editing pairs selected using explicit instruction-following, preservation, and visual-quality criteria.
- **Comprehensive Evaluation and Downstream Validation.** Extensive experiments demonstrate that MoRE-OPD outperforms state-of-the-art open-source reward models in pointwise and preference evaluation and provides stronger reward signals for downstream reinforcement learning.

2 Related Work

2.1 Image Editing Reward Models

Early image editing evaluation largely relied on generic perceptual metrics such as CLIP (Radford et al., 2021), or text-to-image reward models such as ImageReward (Xu et al., 2023) and HPS v2 (Wu et al., 2023), which do not jointly assess instruction following and non-edited content preservation. Recent reward models explore preference-based approaches, including human preference alignment (Wu et al., 2025b), rationale distillation (Wang et al., 2026b), reasoning internalization (Yang et al., 2026), and dynamic rubrics (Kan et al., 2026), as well as pointwise approaches based on semantic and perceptual quality (Luo et al., 2026), spatial reasoning (Long et al., 2026), difference-first evaluation (Zhao et al., 2026), and principle-based verification (Guo et al., 2026). EditHF (Xu et al., 2026) further combines preference pairs with absolute ratings. Despite these advances, existing methods generally rely on a single reward model across heterogeneous editing tasks, potentially entangling distinct evaluation requirements. In contrast, we distill category-specialized reward experts into a universal evaluator through privileged on-policy distillation.

2.2 On-Policy Distillation

On-policy distillation (OPD) enables students to learn from teacher feedback on their own generated outputs, reducing the distribution mismatch between training and inference. MiniLLM (Gu et al., 2024) introduces reverse-KL-based on-policy distillation for generative language models,



Figure 1: **Criterion Entanglement in Universal Image Editing Evaluation.** Universal criteria may incorrectly accept localized edits with unintended changes while rejecting valid global edits. Category-specific criteria provide appropriate judgment boundaries for different editing types.

while GKD (Agarwal et al., 2024) generalizes on-policy distillation with flexible divergence objectives. Recent works further extend OPD to richer supervision settings. OPCD (Ye et al., 2026b) distills context-conditioned knowledge through student-generated trajectories, while privileged information distillation (Penaloza et al., 2026) allows the teacher to access additional training-time information unavailable to the student. MiMo-MOPD (Ma et al., 2026b) further extends OPD to multiple domain-specialized teachers through teacher routing. However, existing OPD methods primarily focus on distilling generative capabilities of language models. How to effectively adapt OPD to reward modeling and consolidate heterogeneous evaluation capabilities remains underexplored.

3 METHOD

3.1 MOTIVATION

Image editing evaluation is inherently heterogeneous, for example, localized edits require strict preservation of unrelated regions, and global edits such as background replacement permit broader appearance changes. As illustrated in Figure 1, these differences create conflicting requirements for a universal evaluator: relaxed preservation constraints may falsely accept unintended changes in localized edits, whereas overly strict constraints may falsely reject legitimate global changes. We term this conflict *criterion entanglement*, motivating category-specialized reward experts with task-specific evaluation criteria. However, deploying multiple large experts is costly for data curation, reinforcement learning, and benchmarking. We therefore consolidate their evaluation capabilities into a compact universal evaluator through privileged on-policy distillation. Unlike conventional SFT, which learns from fixed oracle-generated annotations, our student generates its own evaluation predictions and receives on-policy supervision from the corresponding category-specialized expert. This allows the supervision to adapt to the student's evolving prediction distribution and directly address its current evaluation behavior. Moreover, each expert is provided with oracle-generated rationales as privileged context, enabling more informative feedback without exposing such privileged information to the student. Together, these mechanisms facilitate the transfer of heterogeneous evaluation capabilities from specialized experts into a single universal evaluator.

Method Overview. Figure 2 illustrates the overall framework of MoRE-OPD, which consists of three stages. First, we use a proprietary oracle MLLM to generate annotations under category-specific evaluation protocols, and initialize specialized reward experts accordingly and a universal student under a shared protocol through SFT. Second, we distill the specialized experts into the universal student through privileged on-policy distillation, where experts supervise student-generated predictions with oracle rationales as privileged context. Finally, we augment the distilled evaluator with scalar reward heads and adapt it using LoRA with pairwise preference supervision for fine-grained reward discrimination.

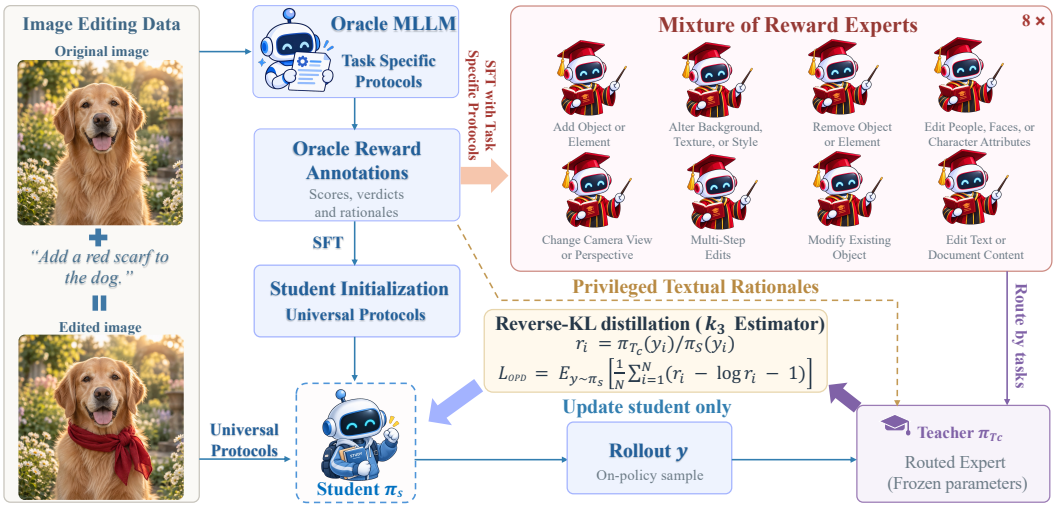


Figure 2: Overview of the proposed method.

3.2 Reward Model Initialization

Evaluation Protocols. To accommodate the heterogeneous evaluation requirements across editing tasks, we partition image editing into $C = 8$ categories.[1] Together, these categories encompass the major forms of image editing operations and provide broad coverage of diverse editing tasks. We construct a category-specific evaluation protocol R_c for each category. All protocols share the same evaluation schema, consisting of instruction following (IF), non-edited region consistency (NC), visual quality (VQ), and an overall verdict, but differ in task definitions, preservation requirements, failure modes, deduction rules, and fatal conditions. We additionally construct a universal protocol R_{uni} that integrates category-dependent rules into a shared evaluation interface for the universal student while retaining the same dimensions, scoring scheme, and output format. Complete protocols are provided in the supplementary material.

Specialized Reward Experts. For each editing category c, we query a stronger proprietary MLLM as the oracle under the corresponding protocol R_c to generate evaluation annotations consisting of scores, textual rationales, and overall verdicts. The resulting annotated dataset $\mathcal{D}_c$ is used to finetune the corresponding expert: $\pi_{T_c} = \text{SFT}(\mathcal{D}_c; R_c), c = 1, \ldots, C$. Here, π_{T_c} denotes the specialized reward expert for category c. Each expert is optimized exclusively for its corresponding editing category, allowing it to specialize in the evaluation criteria and failure modes encoded by R_c.

Universal Student Initialization. We additionally initialize a universal student by aggregating the category-wise oracle annotations as $\mathcal{D} = \bigcup_{c=1}^{C} \mathcal{D}_c$ and performing SFT under the universal protocol R_{uni}: $\pi_S^{(0)} = \text{SFT}(\mathcal{D}, R_{\text{uni}})$, where $\pi_S^{(0)}$ serves as the initialization for subsequent OPD.

3.3 Privileged Multi-Teacher On-Policy Distillation

Given the initialized specialized experts $\{\pi_{T_c}\}_{c=1}^{C}$ and universal student $\pi_S^{(0)}$, we further consolidate the complementary evaluation capabilities of the specialized experts into the student through privileged on-policy distillation. For simplicity, we denote the student as π_S hereafter.

Privileged On-Policy Distillation. For each training instance x from category c, the student first generates an on-policy evaluation response y under the universal evaluation protocol, i.e., $y \sim \pi_S(\cdot \mid x, R_{\text{uni}})$, where x contains the editing instruction, source image, and edited image. The

[1] The eight categories are (1) *Add Object or Element*, (2) *Remove Object or Element*, (3) *Alter Background, Texture, or Style*, (4) *Modify Existing Object*, (5) *Multi-Step Edits*, (6) *Edit People, Faces, or Character Attributes*, (7) *Edit Text or Document Content*, and (8) *Change Camera View or Perspective*.

response is then routed to the corresponding expert π_{T_c} from the pool of category-specialized teachers $\{\pi_{T_c}\}_{c=1}^{C}$ for evaluation. This category-aware routing enables the student to acquire category-specific evaluation capabilities from the corresponding specialized teachers. During distillation, the expert operates under its category-specific protocol R_c and is additionally conditioned on the oracle textual rationale z. As privileged information, the rationale provides additional reasoning evidence that guides the expert toward more reliable judgments and reduces the propagation of erroneous supervision to the student. Importantly, only the textual rationale is provided to the expert, while the oracle scores and overall verdict are withheld.

Unlike static SFT supervision, the on-policy responses evolve with the student throughout training, allowing the specialized experts to provide dynamic supervision tailored to the student's current behavior. Moreover, the student learns from privileged expert supervision without directly accessing the oracle rationale. Together, these designs enable the universal student to progressively integrate the category-specific evaluation capabilities of multiple specialized experts into a unified evaluator.

Reverse-KL Optimization. To transfer the routed expert's evaluation capability to the student, we perform token-level distribution matching on student-generated responses. Following prior on-policy distillation methods (Agarwal et al., 2024; Ma et al., 2026b), we adopt reverse-KL divergence to align the student with the routed expert over the response distribution explored by the student:

$$\mathcal{L}_{\mathrm{RKL}} = D_{\mathrm{KL}}(\pi_S \,\|\, \pi_{T_c}) = \mathbb{E}_{y\sim\pi_S}\left[\log \frac{\pi_S(y \mid x, R_{\mathrm{uni}})}{\pi_{T_c}(y \mid x, R_c, z)}\right]. \tag{1}$$

Directly optimizing Eq. (1) via the log-probability ratio $\log(\pi_S/\pi_{T_c})$ (the k_1 estimator) yields an unbiased but high-variance estimate and can be negative for individual sampled tokens. Such token-wise fluctuations can introduce unstable optimization signals during on-policy distillation. We therefore use the k_3 estimator (Schulman, 2020), which provides a non-negative, lower-variance estimate of the reverse KL. Concretely, we approximate $\mathcal{L}_{\mathrm{RKL}}$ with the token-level on-policy objective:

$$\mathcal{L}_{\mathrm{OPD}} = \mathbb{E}_{y\sim\pi_S}\left[\frac{1}{N}\sum_{i=1}^{N}(r_i - \log r_i - 1)\right], \tag{2}$$

where N denotes the number of tokens in the response and $r_i = \frac{\pi_{T_c}(y_i|x,y_{<i},R_c,z)}{\pi_S(y_i|x,y_{<i},R_{\mathrm{uni}})}$. The specialized experts remain frozen throughout distillation, and only the universal student is optimized. Consequently, the category-specific evaluation capabilities of the specialized experts are consolidated into a single student that operates solely under the universal protocol at inference time.

3.4 Preference-Based Reward Adaptation

The preceding stages equip the distilled evaluator with absolute evaluation capabilities, but do not explicitly train it to distinguish subtle quality differences among plausible edits. We therefore adapt the distilled checkpoint with preference supervision via LoRA, while introducing lightweight scalar reward heads for the overall judgment and the three evaluation dimensions. Given an original image I, an editing instruction q, and an edited image $\tilde{I}$, let $h(I, q, \tilde{I})$ denote the final-layer hidden state at the last non-padding token of the evaluator. For each dimension $d \in \{\mathrm{Overall}, \mathrm{IF}, \mathrm{NC}, \mathrm{VQ}\}$, the scalar reward is computed as

$$s_d(I, q, \tilde{I}) = \mathbf{w}_d^{\top} h(I, q, \tilde{I}) + b_d, \tag{3}$$

where $\mathbf{w}_d$ and b_d are the learnable weight and bias of the reward head for dimension d, respectively.

Given two edited candidates $(\tilde{I}^A, \tilde{I}^B)$ for the same original image and editing instruction, the two candidates are encoded independently with shared weights. Preference supervision is applied only to their score difference through a masked Bradley–Terry loss:

$$\mathcal{L}_d = -\frac{1}{\sum_j m_{d,j}} \sum_j m_{d,j} \log \sigma\Big((2p_{d,j} - 1)\left[s_d(I_j, q_j, \tilde{I}_j^A) - s_d(I_j, q_j, \tilde{I}_j^B)\right]\Big), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, $p_{d,j} = 1$ denotes preference for A and 0 for B, and $m_{d,j}$ masks pairs without a valid preference label for dimension d. The overall adaptation objective is

$$\mathcal{L}_{\mathrm{pref}} = \sum_d \lambda_d \mathcal{L}_d, \tag{5}$$

Table 1: Comparison with existing reward models on pointwise and preference benchmarks. Best and second-best results are shown in bold and underlined, respectively.

Method	PointReward-Bench					EditScore-Bench				EditReward-Bench				MMRB2
	Acc.	TP	FP	TN	FN	PF	Cons.	Ovrl.	Avg.	K2	K3	K4	Avg.	Acc.
EditScore	68.25	346	215	200	39	58.37	56.85	67.77	61.00	69.17	24.00	2.03	30.39	58.00
SpatialReward	71.50	323	166	249	62	64.62	63.37	74.88	67.62	71.43	31.33	7.43	35.50	62.30
EditReward	71.62	286	128	287	99	<u>85.81</u>	66.07	79.64	77.17	76.69	41.33	10.81	41.76	66.10
RationalRewards	<u>75.12</u>	318	132	283	67	84.64	<u>73.03</u>	80.94	79.54	75.19	38.67	8.78	39.68	71.90
RubricRM	73.38	299	127	288	86	84.75	72.47	<u>82.39</u>	<u>79.87</u>	**80.45**	<u>44.00</u>	**17.57**	<u>46.17</u>	<u>72.85</u>
Ours	**82.62**	295	49	366	90	**87.82**	**87.42**	**84.73**	**86.46**	<u>79.70</u>	**46.67**	<u>16.22</u>	**46.40**	**73.80**

where λ_d controls the contribution of each evaluation dimension. This preference-based adaptation enables the scalar rewards to capture fine-grained relative quality differences while retaining independent scoring of individual edited candidates at inference time. We use the generative evaluator obtained after OPD for pointwise evaluation and data filtering. Its preference-adapted variant with scalar reward heads is used for preference evaluation and downstream RL.

4 Pointwise Evaluation

We first assess the pointwise evaluation capability of MoRE-OPD, i.e., its ability to determine the success or failure of an individual edit.

4.1 PointReward-Bench Construction

Existing image editing reward benchmarks primarily evaluate relative preferences between candidate edits, rather than whether an individual edit succeeds or fails. To fill this gap, we construct **PointReward-Bench**, a human-annotated benchmark containing 800 samples across the 8 categories defined in Section 3.2, with 100 samples per category. A human annotator selects successful edits and representative failures from multiple editing models, including Klein (Labs, 2025), HunyuanImage-3.0 (Cao et al., 2025), and Nano Banana Pro. Three additional annotators independently label each sample as successful or failed based on instruction following, non-edited content preservation, and visual quality, with the final ground-truth label determined by majority vote.

4.2 Experimental Setup

We construct a 10K-pair training set covering the eight editing categories defined in Section 3.2. Qwen3.7-Plus serves as the oracle MLLM. MoRE-OPD is implemented with *verl* (Sheng et al., 2024), using Qwen3.6-27B as the eight specialized experts and Qwen3.5-9B as the universal student. During OPD, we use a batch size of 96, a per-GPU micro-batch size of 1, and a learning rate of 1×10^{-6}. Training uses 64 NVIDIA H20 GPUs, with 48 for the student and 16 for the frozen experts. We compare against five recent state-of-the-art open-source image editing reward models: EditScore (Luo et al., 2026), EditReward (Wu et al., 2025b), SpatialReward (Long et al., 2026), RationalRewards (Wang et al., 2026b), and RubricRM (Kan et al., 2026). We report classification accuracy with true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

4.3 Results

Table 1 reports the pointwise results on PointReward-Bench. For competing models with continuous outputs, we exhaustively search the decision threshold and report the best accuracy. Since RubricRM requires paired candidates, we use the source image as the reference. MoRE-OPD achieves the highest accuracy of 82.62%, outperforming the recent state-of-the-art baselines RationalRewards and RubricRM by 7.50 and 9.24 percentage points, respectively. This demonstrates that MoRE-OPD effectively consolidates the evaluation capabilities of category-specialized teachers into a compact universal evaluator. EditScore and SpatialReward yield high true positive counts but also more false positives, resulting in relatively low overall accuracy. EditReward and RubricRM reduce false

Table 2: Ablation study on PointReward-Bench.

Method	Acc. (%)	TP	FP	TN	FN
Qwen3.7-Plus (Oracle)	83.75	323	68	347	62
Specialized Teacher	82.12	331	89	326	54
Initial Student	80.38	319	91	324	66
MoRE-OPD w/o Privilege	75.75	351	160	255	34
MoRE-OPD	82.62	295	49	366	90

positives at the cost of fewer true positives. RationalRewards provides the best balance among the baselines. In comparison, MoRE-OPD reduces false positives to only 49 while correctly rejecting 366 unsuccessful edits, demonstrating stronger failure detection for large-scale data curation.

4.4 Application in High-Quality Training Data Construction

High-quality training data is essential for improving the performance and reliability of image editing models. To construct a large-scale and reliable training set, we apply MoRE-OPD to publicly available image editing datasets (Liu et al., 2026; Chen et al., 2026; Wang et al., 2025; Chang et al., 2025; Ye et al., 2026a; Ma et al., 2026a; Kuprashevich et al., 2025) for automatic quality filtering. Specifically, MoRE-OPD retains samples satisfying $\mathrm{IF} \geq 9$, $\mathrm{NC} \geq 8$, and $\mathrm{VQ} \geq 6$. This process retains approximately 3.8 million editing samples that satisfy the model-based success criteria, forming a large-scale filtered data resource for downstream optimization.

4.5 Ablation Studies

To investigate the effectiveness of the proposed privileged OPD, we compare MoRE-OPD with its variant without oracle-generated rationales (w/o Privilege) and the initial student. As shown in Table 2, MoRE-OPD improves the initial student by 2.24 percentage points and substantially outperforms its variant without privileged information. Both the specialized teacher and initial student exhibit a relatively permissive evaluation tendency, producing more false positives than false negatives. Without privileged information, OPD further amplifies this tendency, increasing the number of false positives to 160. By incorporating oracle-generated textual rationales, MoRE-OPD effectively suppresses this bias, reducing false positives to 49. Finally, the resulting 9B universal student achieves performance comparable to the specialized 27B teacher and approaches the oracle performance. These results demonstrate that privileged OPD mitigates permissive evaluation while effectively consolidating specialized evaluation capabilities into a compact universal evaluator.

5 Preference Evaluation

We further evaluate the preference discrimination capability of MoRE-OPD against EditScore, EditReward, SpatialReward, RationalRewards, and RubricRM. This evaluation examines whether a reward model can reliably distinguish better edits from worse ones.

5.1 Experimental Setup

For preference-based reward adaptation, we use the EditReward training set (Wu et al., 2025b), which provides human-annotated preferences for overall quality, instruction following (IF), and visual quality (VQ). Since EditReward does not provide annotations for non-edited region consistency (NC), we additionally query Qwen3.7-Plus to obtain pairwise NC preferences for these samples. We initialize from the distilled MoRE-OPD checkpoint and apply LoRA with rank 64 and $\alpha = 128$. The loss weights for Overall, IF, NC, and VQ are set to 1.0, 0.5, 0.5, and 0.5, respectively. We use learning rates of 5×10^{-6} and 5×10^{-5} for the LoRA parameters and scalar reward heads, respectively, and train for two epochs on four nodes, each equipped with eight NVIDIA H20 GPUs. For preference evaluation, we employ three commonly used benchmarks: EditScore-Bench (Luo et al., 2026), EditReward-Bench (Wu et al., 2025b), and MMRB2-Edit (Hu et al., 2026). To avoid positional bias, we randomly shuffle the order of the two candidates in each preference pair.

Table 3: Downstream reinforcement learning results on GEdit-Bench and ImgEdit-Bench. All methods use the same base editing model and RL algorithm, differing only in the reward model. Higher is better. Best and second-best results are shown in bold and underlined, respectively.

Reward	GEdit-Bench			ImgEdit-Bench
	Q_SC	Q_PQ	Q_O	Overall
None (klein)	7.58	6.61	6.89	3.44
EditScore	7.66	6.72	7.01	<u>3.47</u>
EditReward	7.64	6.66	6.95	<u>3.47</u>
SpatialReward	<u>7.70</u>	<u>6.79</u>	<u>7.03</u>	3.44
RubricRM	7.65	6.62	6.94	<u>3.47</u>
RationalRewards	7.64	6.73	6.99	3.46
Ours	**7.85**	**6.84**	**7.13**	**3.52**

5.2 Results

As reported in Table 1, MoRE-OPD demonstrates consistently strong preference discrimination across all three benchmarks. Notably, such strong preference performance is achieved with only lightweight LoRA-based adaptation, further indicating that MoRE-OPD effectively integrates the evaluation capabilities of specialized teachers and the oracle into the compact universal reward model. On EditScore-Bench, MoRE-OPD achieves the best performance across all reported dimensions, particularly in consistency evaluation, suggesting a stronger ability to capture preservation-related differences between candidate edits. On EditReward-Bench, MoRE-OPD achieves competitive performance with the state-of-the-art baseline RubricRM, despite evaluating each candidate independently, whereas RubricRM jointly observes both candidates during comparison. MoRE-OPD further achieves the highest accuracy of 73.80% on MMRB2-Edit, surpassing RubricRM and RationalRewards at 72.85% and 71.90%, respectively. These results suggest that MoRE-OPD can provide reliable reward signals for downstream reinforcement learning.

6 Downstream Reinforcement Learning

6.1 Experimental Setup

To further evaluate the effectiveness of MoRE-OPD as a reward signal, we sample approximately 20K source image–instruction pairs from the data pool in Section 4.4 and perform online RL on the four-step distilled FLUX.2 [klein] model (Labs, 2025) using Flow GRPO (Wang et al., 2026a). Its strong editing capability provides a challenging setting for evaluating reward effectiveness. Each rollout uses 32 image–instruction pairs with 16 candidates per pair, resulting in 512 trajectories. Images are generated at 512×512 resolution using four denoising steps with a guidance scale of 1.0. MoRE-OPD remains frozen, while the editing model is optimized with LoRA using a rank of 16, a scaling factor of 32, and a learning rate of 5×10^{-5}. For a fair comparison, all competing reward models use the same training data, base editing model, and RL configuration as MoRE-OPD, differing only in the reward model used for optimization. We use GEdit-Bench (Liu et al., 2025) and ImgEdit-Bench (Ye et al., 2026c) to evaluate the resulting editing models. We follow the official benchmark protocols, replacing the default judge with Qwen3.7-Plus for all methods.

6.2 Results

Quantitative Results. As shown in Table 3, MoRE-OPD achieves the highest scores across all reported metrics on both benchmarks. Despite the strong performance of the four-step distilled FLUX.2 [klein] baseline, optimizing with MoRE-OPD improves GEdit-Bench Q_{SC}, Q_{PQ}, and Q_{O} from 7.58, 6.61, and 6.89 to 7.85, 6.84, and 7.13, respectively, and increases the ImgEdit-Bench score from 3.44 to 3.52. Under identical RL training settings, it also outperforms all compared reward models, with gains of 0.15, 0.05, and 0.10 over the best competing scores on the three GEdit-Bench metrics. These results support the effectiveness of MoRE-OPD as a reward signal for further improving a strong editing model.



Instruction: "Replace the bird in the image with a squirrel sitting on the branch."



Instruction: "Change the bicycle's color to red."

Figure 3: Qualitative comparison of editing models optimized with different reward models. The order is Original, MoRE-OPD, RationalRewards, RubricRM, SpatialReward, EditReward, EditScore, and FLUX.2 [klein], from left to right and top to bottom.

Qualitative Results. Figure 3 provides qualitative comparisons of editing models optimized with different reward models. MoRE-OPD better follows instruction in the first example, producing a naturally seated squirrel on the branch. In the second example, MoRE-OPD produces a more natural color that harmonizes with the scene, avoiding the overly saturated recoloring of several competing methods. These results demonstrate that MoRE-OPD provides an effective reward signal for improving overall editing performance. Additional qualitative results are provided in the Appendix.

7 Conclusion

In this work, we propose MoRE-OPD, a reward modeling framework for reliable image editing evaluation. To overcome the challenge of criterion entanglement across heterogeneous editing tasks, MoRE-OPD first trains category-specialized reward experts under task-specific protocols. It then consolidates their distinct capabilities into a compact, universal evaluator via privileged on-policy distillation. By equipping these specialized teachers with oracle-generated rationales as privileged context, MoRE-OPD facilitates highly effective knowledge transfer while requiring only standard inputs at inference time. Furthermore, we introduce PointReward-Bench for rigorous pointwise evaluation, and curate a large-scale dataset of approximately 3.8 million high-quality editing samples. Extensive experiments demonstrate that MoRE-OPD achieves leading pointwise accuracy, strong preference alignment, and consistent downstream RL improvements.

AI USE STATEMENT

In this work, we used generative AI tools to generate training data and supervision signals, as described in the paper. We have not used generative AI tools to develop theoretical models, formulate mathematical claims, assist with proofs, or conduct qualitative or thematic data analysis, and these tasks are otherwise not applicable to this work. Additionally, we used generative AI tools to draft parts of the manuscript, and to edit the language for improved readability. All AI-generated training data and supervision signals were subject to the quality control and evaluation procedures described in the paper, and all AI-assisted manuscript content was reviewed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv e-prints*, pp. arXiv–2509, 2025.

Di Chang, Mingdeng Cao, Yichun Shi, Bo Liu, Shengqu Cai, Shijie Zhou, Weilin Huang, Gordon Wetzstein, Mohammad Soleymani, and Peng Wang. Bytemorph: Benchmarking instruction-guided image editing with non-rigid motions. *arXiv preprint arXiv:2506.03107*, 2025.

Guanzhou Chen, Erfei Cui, Changyao Tian, Danni Yang, Ganlin Yang, Yu Qiao, Hongsheng Li, Gen Luo, and Hongjie Zhang. Scaleedit-12m: Scaling open-source image editing data generation via multi-agent framework. *arXiv preprint arXiv:2603.20644*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Hanzhong Guo, Jie Wu, Jie Liu, Yu Gao, Zilyu Ye, Linxiao Yuan, Xionghui Wang, Yizhou Yu, and Weilin Huang. Leveraging verifier-based reinforcement learning in image editing. *arXiv preprint arXiv:2604.27505*, 2026.

Yushi Hu, Reyhane Askari-Hemmat, Melissa Hall, Emily Dinan, Luke Zettlemoyer, and Marjan Ghazvininejad. Multimodal rewardbench 2: Evaluating omni reward models for interleaved text and image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36904–36915, 2026.

Zijian Kan, Wei Wang, Long Luo, Bing Zhao, Xuan Ren, Weixu Qiao, Wenbo Li, Hu Wei, and Lin Qu. Rubricrm: Generative reward modeling via dynamic rubrics for image generation and editing. *arXiv preprint arXiv:2608.26956*, 2026.

Maksim Kuprashevich, Grigorii Alekseenko, Irina Tolstykh, Georgii Fedorov, Bulat Suleimanov, Vladimir Dokholyan, and Aleksandr Gordeev. NoHumansRequired: Autonomous High-Quality Image Editing Triplet Mining. *arXiv preprint arXiv:2507.14119*, 2025. URL https://arxiv.org/abs/2507.14119.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Delong Liu, Haotian Hou, Zhaohui Hou, Zhiyuan Huang, Shihao Han, Mingjie Zhan, Zhicheng Zhao, and Fei Su. Inter-edit: First benchmark for interactive instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37290–37300, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Yancheng Long, Yankai Yang, Hongyang Wei, Wei Chen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Jiankang Chen, Kaiyu Tang, Bin Wen, et al. Spatialreward: Bridging the perception gap in online rl for image editing via explicit spatial reasoning. *arXiv preprint arXiv:2602.07458*, 2026.

Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. In *International Conference on Learning Representations*, volume 2026, pp. 33027–33056, 2026.

Jian Ma, Xujie Zhu, Zihao Pan, Qirong Peng, Xu Guo, Chen Chen, and Haonan Lu. X2edit: Revisiting arbitrary-instruction image editing through self-constructed data and task-aware representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7764–7772, 2026a.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026b.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

John Schulman. Approximating kl divergence, 2020. URL `http://joschu.net/blog/kl-approx.html`.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Haonan Wang, Linyu Wu, Qian Qiu, Lewei Jin, Bowen Ping, Jianghai Chen, Yiheng Du, Guangxin He, Yu Shi, Yongguang Lin, Zhuoxin Zhou, Zhanchao Zhou, Keming Wu, Rizhen Hu, Xuefei Ning, Lvfang Tao, Feiyu Hu, Xiangyan Liu, Siqi Kou, Jiarui Yao, Xiangxin Zhou, Liefeng Bo, Wenxi Zhu, and Tianyu Pang. UniRL: A Reinforcement Learning Framework for Unified Multimodal Models. `https://github.com/Tencent-Hunyuan/UniRL`, 2026a.

Haozhe Wang, Cong Wei, Weiming Ren, Jiaming Liu, Fangzhen Lin, and Wenhu Chen. Rationalrewards: Reasoning rewards scale visual generation both training and test time. *arXiv preprint arXiv:2604.11626*, 2026b.

Yuhan Wang, Siwei Yang, Bingchen Zhao, Letian Zhang, Qing Liu, Yuyin Zhou, and Cihang Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. *arXiv preprint arXiv:2507.21033*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025b.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2096–2105. IEEE, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Zitong Xu, Huiyu Duan, Zhongpeng Ji, Xinyun Zhang, Yutao Liu, Xiongkuo Min, Ke Gu, Jian Zhang, Shusong Xu, Jinwei Chen, et al. Edithf-1m: A million-scale rich human preference feedback for image editing. In *European Conference on Computer Vision*, pp. 278–298. Springer, 2026.

Yankai Yang, Yancheng Long, Hongyang Wei, Wei Chen, Tianke Zhang, Kaiyu Jiang, Haonan Fan, Changyi Liu, Jiankang Chen, Kaiyu Tang, et al. Joint reward modeling: Internalizing chain-of-thought for efficient visual reward models. *arXiv preprint arXiv:2602.07533*, 2026.

Keming Ye, Zhipeng Huang, Canmiao Fu, Qingyang Liu, Jiani Cai, Zheqi Lv, Chen Li, Jing Lyu, Zhou Zhao, and Shengyu Zhang. Unicedit-10m: A dataset and benchmark breaking the scale-quality barrier via unified verification for reasoning-enriched edits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37279–37289, 2026a.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026b.

Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026c.

Xiangyu Zhao, Peiyuan Zhang, Junming Lin, Tianhao Liang, Yuchen Duan, Shengyuan Ding, Changyao Tian, Yuhang Zang, Junchi Yan, and Xue Yang. Trust your critic: Robust reward modeling and reinforcement learning for faithful image editing and generation. *arXiv preprint arXiv:2603.12247*, 2026.

A Evaluation Protocols

We provide additional details on the evaluation protocols used to construct the specialized reward experts and the universal student. All protocols evaluate an edited image with respect to the input image and the corresponding editing instruction along three primary dimensions: instruction following (IF), non-edited region consistency (NC), and visual quality (VQ), together with an overall judgment. Scores range from 0 to 10, and an edit is considered successful only when

$$\mathrm{IF} \geq 9, \qquad \mathrm{NC} \geq 8, \qquad \mathrm{VQ} \geq 6.$$

The complete evaluation protocols, including the universal protocol and all eight category-specific protocols, are provided in the supplementary material.

Category-Specific Protocols. We construct eight category-specific evaluation protocols R_c corresponding to the editing categories defined in Section 3.2. All protocols follow the same evaluation structure but differ in their task definitions, deduction rules, exemptions, and failure conditions. These differences capture the distinct evaluation requirements of heterogeneous editing tasks. For example, text editing imposes strict character-level correctness requirements, camera-view editing distinguishes genuine 3D viewpoint transformations from 2D transformations, multi-step editing evaluates each requested sub-task independently, and background/style editing permits appearance changes that are expected consequences of the requested global transformation.

Instruction Following. IF evaluates whether the requested edit is accurately and completely executed on the intended target. Each protocol starts from a score of 10 and applies category-dependent deductions for errors such as missing edits, incorrect targets, incorrect edit types, incomplete execution, and unintended modifications. Additional failure conditions are introduced for categories with stricter requirements. For example, text editing heavily penalizes misspelled or illegible text, while camera-view editing penalizes apparent 2D transformations that do not correspond to the requested 3D viewpoint change.

Non-Edited Region Consistency. NC evaluates whether content outside the intended editing scope is properly preserved. The protocols consider semantic content changes, geometric and structural deformation, visual artifacts, abnormal lighting or color changes, identity changes, and unintended text modifications. Importantly, the applicable preservation requirements depend on the editing category. Background replacement permits expected changes to the original background, while camera-view changes permit geometrically plausible parallax, occlusion, lighting variation, and newly revealed scene content.

Visual Quality. VQ evaluates the perceptual and physical plausibility of the edited result. Common considerations include boundary integration, lighting and material consistency, shadows, perspective, depth of field, structural correctness, and generation artifacts. Category-dependent criteria are additionally applied when necessary, such as typography and surface alignment for text editing, anatomical correctness for human editing, and cross-edit coherence for multi-step editing.

Universal Protocol. In addition to the category-specific protocols, we construct a universal evaluation protocol R_{uni} covering all eight editing categories. Given an editing instruction, R_{uni} first infers the corresponding edit type and decomposes compound instructions into individual sub-tasks when necessary. It then evaluates the edited image using the shared IF, NC, and VQ dimensions while applying conditional rules for category-dependent cases. The universal protocol is used by the student model, whereas each specialized teacher uses its corresponding category-specific protocol.

B Additional Qualitative Results

We provide additional qualitative comparisons across four representative image editing tasks, including text removal, brightness adjustment, background replacement, and object color modification. All examples follow the same model ordering as the qualitative comparison in the main paper. As shown in Figures 4–7, these examples cover both localized edits that require strict preservation of non-target content and global edits that permit substantial appearance changes. For text removal, the

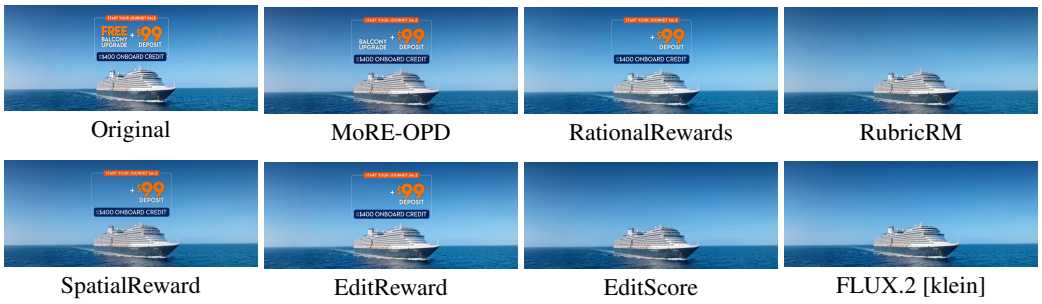


Figure 4: Additional qualitative comparison. **Instruction:** "Remove the text 'FREE' ".



Figure 5: Additional qualitative comparison. **Instruction:** "Make the image brighter."

compared reward models differ in their ability to completely remove the target text while preserving the surrounding content. For brightness adjustment, the competing methods tend to excessively increase the overall brightness, resulting in noticeable overexposure and loss of scene details, whereas MoRE-OPD produces a more controlled brightness increase while better preserving the original scene details. The background replacement example further examines whether large-scale scene changes can be performed while preserving the foreground subject and composition. Finally, for object color modification, all methods largely achieve the requested color change, while MoRE-OPD produces a more natural appearance that better harmonizes with the original lighting, material, and scene. Overall, these qualitative results demonstrate that MoRE-OPD provides more reliable reward signals across diverse editing tasks, encouraging both accurate edit execution and appropriate preservation while maintaining natural visual quality.

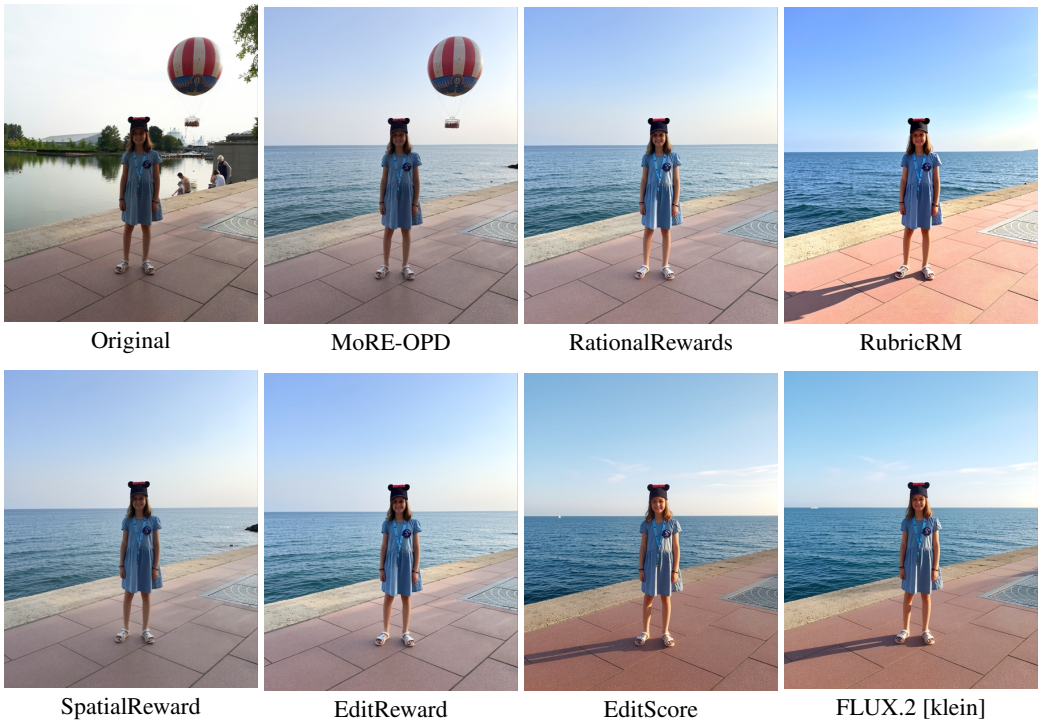


Figure 6: Additional qualitative comparison. **Instruction:** "Change the background to the sea."



Figure 7: Additional qualitative comparison. **Instruction:** "Change this bag to red."