# EDIT-GRPO: A LOCALITY-PRESERVING POLICY OPTIMIZATION FRAMEWORK FOR IMAGE EDITING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A fundamental challenge in image editing lies in preserving spatial locality: edits should improve targeted content without inadvertently altering surrounding regions. However, most optimization-based editing approaches treat images as holistic entities, causing global policy updates that undermine locality and introduce undesired context changes. We observe that this issue stems from a mismatch between localized editing intent and globally applied optimization signals. Motivated by this insight, we propose Edit-GRPO, preserving locality while optimizing image editing, a locality-preserving policy optimization framework that explicitly decouples editing and preservation objectives. By assigning region-specific optimization signals to edit and non-edit areas, Edit-GRPO aligns policy updates with the spatial structure of editing tasks, enabling localized improvements while maintaining global visual coherence. This design effectively suppresses common artifacts such as context distortion and boundary inconsistency. Extensive experiments across diverse image editing scenarios demonstrate that Edit-GRPO significantly improves locality preservation while maintaining strong editing performance compared to existing optimization-based methods, validating the generality and effectiveness of the proposed framework.

## 1 INTRODUCTION

Recent advances in generative models, particularly diffusion models (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2022) and flow models (Lipman et al., 2022; Liu et al., 2022; Esser et al., 2024) have significantly improved visual content creation, enabling high-quality instruction-guided image editing (Labs et al., 2025; Wu et al., 2025a), which requires both precise semantic control and faithful context preservation.

Inspired by the success of reinforcement learning from human feedback (RLHF) in LLMs (Ouyang et al., 2022; Guo et al., 2025; Shao et al., 2024), recent work has explored integrating RL with visual generative models (Black et al., 2023; Fan et al., 2023). Recently, one line of work (Liu et al., 2025a; Xue et al., 2025; Li et al., 2025a) introduces GRPO-style policy optimization to diffusion models by converting the ODE sampling into an equivalent SDE. Another line of work (Zheng et al., 2025; Li et al., 2025b) performs policy optimization directly on the forward diffusion process via flow matching objectives, enabling more efficient training. Despite their differences in policy optimization, these RL-based approaches share a common design: *reward signals are computed at image level and used to guide policy updates globally*.

While this global optimization paradigm works well for general image generation, it becomes problematic when applied to image editing. Unlike generation, editing is inherently a localized task: **only the target regions should be modified, while the remaining content should remain unchanged**. However, we observe that existing optimization-based approaches (Liu et al., 2025a; Xue et al., 2025; Li et al., 2025a; Zheng et al., 2025; Li et al., 2025b) typically treat images as holistic entities and use image-level rewards to update the policy, which conflicts with the localized nature of editing. More importantly, a single global reward can **not** explicitly distinguish where the model should edit and where it should preserve, making the editing model struggle to achieve precise semantic modification and faithful context preservation simultaneously. When the model is encouraged to perform stronger semantic edits, unintended distortion beyond the editing areas arises inadvertently (such as background drift and boundary inconsistency), degrading the final



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.

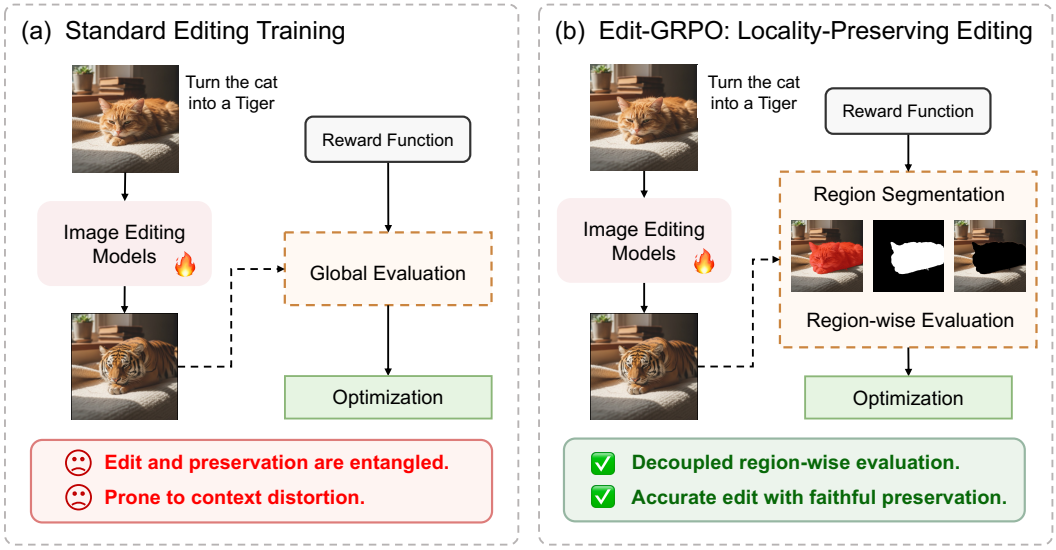


Figure 1: Overview of the proposed Edit-GRPO framework compared to standard editing post-training. Edit-GRPO performs region-wise evaluation to better preserve non-edit content.

editing quality. This observation suggests that, in image editing, **locality preservation should not be treated as a byproduct of semantic editing, but as an explicit optimization objective during post-training**. Consequently, reward signals should be region-specific rather than globally defined: semantic following reward should be evaluated in the target regions, while locality preservation reward should be evaluated individually in the non-edit regions. This insight motivates us to explore a decoupled reward mechanism that assesses semantic editing and locality preservation separately, which remains underexplored in image editing post-training.

In this work, we propose **Edit-GRPO**, a policy optimization framework that **preserves locality while optimizing image editing**. The core idea is to decouple the reward design based on the spatial regions of the edited images. Specifically, we partition the images into edit regions and non-edit regions, and design different rewards separately. For edit regions, we apply a semantic following reward to encourage precise editing that follows the instruction. Meanwhile, we introduce a locality preservation reward for the non-edit regions that explicitly encourages the model to retain the original visual content in areas unrelated to the edit. Based on the decoupled reward mechanism, we further propose a region-decoupled optimization framework to mitigate the interference between editing and preservation objectives. By decoupling optimization objectives in this way, Edit-GRPO enables joint policy optimization for improving both semantic editing and locality preservation, encouraging the model to produce more visually coherent and realistic editing results.

To validate Edit-GRPO, we apply our method to FLUX.1-Kontext [Dev] (Labs et al., 2025) and Qwen-Image-Edit [2509] (Wu et al., 2025a). Extensive experimental results demonstrate that Edit-GRPO consistently improves locality preservation, effectively suppressing unintended changes in non-edit regions. Importantly, Edit-GRPO achieves comparable or even better semantic editing performance relative to the base models, indicating that the locality-preserving improvements do not compromise edit capability. Together, these results show that Edit-GRPO improves editing precision and locality preservation simultaneously, highlighting the effectiveness of the proposed framework.

The main contributions of this work are summarized as follows:

- We identify a fundamental mismatch between localized editing intent and globally applied optimization signals in existing optimization-based image editing methods.
- We propose Edit-GRPO, a locality-preserving policy optimization framework that decouples editing and preservation objectives through region-specific optimization signals.
- Extensive experiments demonstrate that Edit-GRPO consistently improves both editing fidelity and context preservation over the base models.

# 2 Related Work

## 2.1 Instruction-Based Image Editing

Diffusion and flow models (Ho et al., 2020; Song et al., 2020a; Rombach et al., 2022; Song et al., 2020b; Lipman et al., 2022; Liu et al., 2022) have significantly advanced text-to-image generation and instruction-based image editing, allowing users to manipulate visual content through natural language instructions. Early works (Meng et al., 2021; Brooks et al., 2022; Wang et al., 2023; Zhang et al., 2023) demonstrate that diffusion models can acquire instruction-following editing capabilities through supervised fine-tuning. Recent works (Labs et al., 2025; Wu et al., 2025a; Liu et al., 2025b; Cai et al., 2025; Deng et al., 2025) further improve editing performance by scaling model capacity and training data. However, these Supervised Fine-Tuning (SFT) approaches often lead to shortcut learning and overfitting to large-scale datasets (Li et al., 2025b), limiting the generalization and controllability of editing models. To this end, post-training alignment methods based on reinforcement learning (RL) have emerged as a promising solution.

## 2.2 Policy Optimization for Generative Models

Inspired by the success of reinforcement learning from human feedback in LLMs (Ouyang et al., 2022; Guo et al., 2025; Shao et al., 2024), recent efforts have extended RL to visual generative models. Pioneer works like DDPO (Black et al., 2023) and ReFL (Xu et al., 2023) apply PPO (Schulman et al., 2017) to fine-tune diffusion models for human preference alignment. Diffusion-DPO (Wallace et al., 2023) adapts Direct Preference Optimization (DPO) (Rafailov et al., 2023) to directly optimize diffusion models from paired data, bypassing the need for explicit reward models. Subsequent studies such as Flow-GRPO (Liu et al., 2025a) and Dance-GRPO (Xue et al., 2025) enable GRPO-style policy updates by converting the ODE sampling into an equivalent SDE. MixGRPO (Li et al., 2025a) further improves training efficiency through a hybrid ODE-SDE sampling approach. More recently, DiffusionNFT (Zheng et al., 2025; Li et al., 2025b) directly optimizes diffusion models along the forward process via flow matching objectives, leading to better generation performance. Despite these advances, existing optimization methods typically treat images as holistic entities and rely on image-level rewards to update the policy, which works well for generic image generation but becomes problematic when applied to image editing, where edits are expected to occur only in specific regions without altering surrounding content. Achieving the balance between precise semantic editing and faithful locality preservation remains an open challenge in complex image editing tasks.

# 3 Preliminaries

In this section, we briefly review the formulation of flow matching (Lipman et al., 2022; Liu et al., 2022) and Flow-GRPO (Liu et al., 2025a).

**Flow Matching.** Let $x_0 \sim X_0$ denote a clean image sample and $x_1 \sim X_1$ denote a Gaussian noise sample. Following rectified flow, the interpolated sample at time $t \in [0, 1]$ is defined as:

$$x_t = (1 - t)x_0 + tx_1, \tag{1}$$

with target velocity field $v = x_1 - x_0$. A neural network $v_\theta(x_t, t, c)$ conditioned on instruction $c$ is trained to regress the target velocity by minimizing the flow matching objective:

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,x_0\sim X_0,x_1\sim X_1}\left[\|v - v_\theta(x_t, t, c)\|_2^2\right]. \tag{2}$$

At inference time, edited images are generated by solving the deterministic ODE conditioned on the source image $I$ and editing instruction $c$.

**Flow-GRPO.** Our approach is built upon Flow-GRPO, which formulates the denoising process as a Markov Decision Process. Specifically, given a prompt $c$, the flow model $\pi_\theta$ first samples a group of $G$ images $\{x_0^i\}_{i=1}^G$ and their corresponding reverse-time trajectories $\{(x_T^i, x_{T-1}^i, \ldots, x_0^i)\}_{i=1}^G$. Then the group-relative advantage $\hat{A}_t^i$ is computed by normalizing the rewards within the group:

$$\hat{A}_t^i = \frac{R(x_0^i, c) - \text{mean}(\{R(x_0^j, c)\}_{j=1}^G)}{\text{std}(\{R(x_0^j, c)\}_{j=1}^G)}. \tag{3}$$

The policy is optimized by maximizing the following objective:

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t=0}^{T-1}\min\Big(\rho_t^i(\theta)\hat{A}_t^i, \text{clip}(\rho_t^i(\theta), 1\pm\epsilon)\hat{A}_t^i\Big) - \beta D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Big], \quad (4)$$

where $\rho_t^i(\theta) = \frac{p_\theta(x_{t-1}^i|x_t^i,c)}{p_{\theta_{\text{old}}}(x_{t-1}^i|x_t^i,c)}$ denotes the probability ratio.

# 4 METHOD

In this section, we present Edit-GRPO, a novel policy optimization framework that preserves locality while optimizing image editing. We begin by analyzing the limitations of existing optimization-based image editing approaches in Section 4.1; then we show how to identify and decompose edit regions and design region-specific rewards in Section 4.2 and Section 4.3; finally, in Section 4.4 we propose a region-decoupled training framework. The overall pipeline of Edit-GRPO is illustrated in Figure 3.

## 4.1 MISMATCH BETWEEN LOCALIZED EDITING AND GLOBAL OPTIMIZATION SIGNAL

Image editing is fundamentally different from general image generation, as it requires not only *precise semantic control* over the target content, but also *faithful context preservation* of the background. However, as illustrated in Figure 1, existing optimization-based approaches treat images as holistic entities and conduct global evaluation over the whole image, omitting the inherently localized nature of editing. While such a *holistic* optimization strategy works well for image generation, it becomes problematic when applied to image editing tasks. As shown in Figure 2, while global optimization successfully encourages the model to follow the editing instruction, it induces undesired background distortions, leading to a suboptimal editing result. This phenomenon reveals a fundamental mismatch between localized editing intent and globally applied optimization signals, as global optimization cannot precisely localize the editing region, causing gradients to inevitably affect the background pixels.

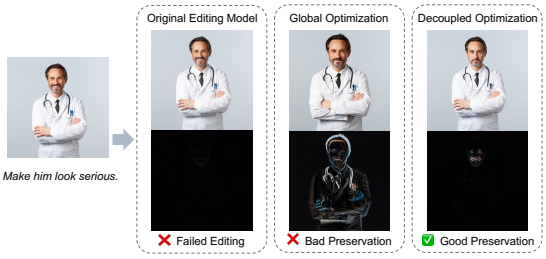


Figure 2: Visual comparison between global and decoupled optimization. Please zoom in for details.

Motivated by this observation, we argue that **reward design in image editing should be region-specific rather than globally defined**. To this end, we propose a novel locality-preserving policy optimization framework that separates the preservation objective from the editing one. By assigning region-specific reward signals to the edit and non-edit regions, the model can improve semantic following within the target area while penalizing undesired modifications elsewhere. As shown in Figure 2, the region-decoupled optimization strategy better preserves surrounding context while achieving the intended edit, resulting in more faithful and spatially consistent image editing.

## 4.2 LANGUAGE-GROUNDED REGION DECOUPLING MODULE

To achieve accurate decoupling for the edit and non-edit regions without manual annotations, we propose a Language-Grounded Region Decoupling Module that bridges free-form editing instructions and spatial segmentation of edited images. Specifically, given an input image $I$ and an editing instruction $c$, we first use an LLM to parse the instruction and extract structured descriptions of the target content. These descriptions and images are then fed into a segmentation model to generate high-resolution segmentation masks for the edit target. To acquire robust segmentation in diverse editing scenarios, we apply region decoupling to the original image $I$ and the edited image $\hat{I}$, obtaining two individual masks. The final mask $M$ is computed as the union of the two masks:

$$M = \mathcal{S}(I, c) \cup \mathcal{S}(\hat{I}, c), \quad (5)$$

where $\mathcal{S}$ denotes the language-grounded region decoupling module. The remaining areas are defined as the non-edit areas $\bar{M} = 1 - M$. By grounding editing instructions into precise spatial masks, our

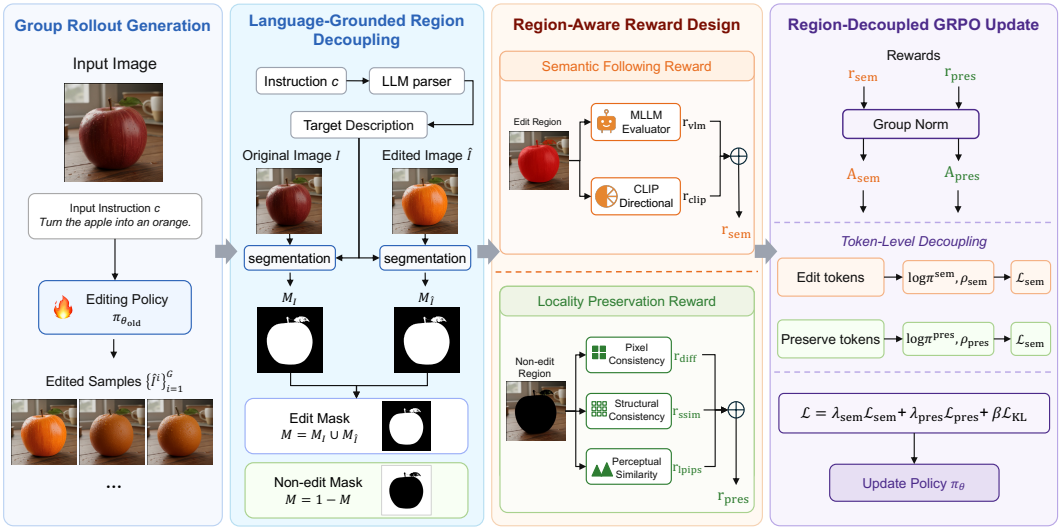


Figure 3: Overview of the proposed Edit-GRPO pipeline.

approach eliminates the need for expensive user annotations, automatically generating precise masks for the edit regions in instruction-based image editing tasks.

### 4.3 Region-Specific Reward Design

Based on the segmentation masks, we further design a region-specific reward mechanism that decomposes the optimization objective into two distinct components: a **semantic following reward**, which evaluates the precision of the intended edit, and a **locality preservation reward**, which measures the preservation of the non-edit background. We detail our reward design as follows:

**Semantic Following Reward.** This reward assesses how accurately the edited image reflects the semantic intent of the editing instruction. Recent studies (Wang et al., 2025; Xu et al., 2026; Gong et al., 2025; Wu et al., 2025c; Li et al., 2025b) have revealed that MLLMs demonstrate superior capability in evaluating image quality due to their strong visual understanding and human preference alignment. In Edit-GRPO, we leverage a pretrained MLLM as a training-free reward model to evaluate editing quality. Specifically, given the original image $I$, the edited image $\hat{I}$ and the editing instruction $c$, instead of using discrete output ratings, we utilize **logit-based** ratings calculated from the rating token probability to obtain fine-grained continuous reward signal (Wu et al., 2025b; Zhang et al., 2025; Li et al., 2025b). Formally, we define the MLLM reward as:

$$r_{\text{vlm}} = \sum_{s \in \mathcal{S}} s \cdot p(s|I, \hat{I}, c), \tag{6}$$

where $\mathcal{S} = \{0, 1, 2, 3, 4, 5\}$ is the set of predefined numerical rating tokens and $p(s|\cdot)$ is the softmax probability of that token. In addition, to ensure the edit aligns well with the instruction, we further incorporate a CLIP-based directional reward:

$$r_{\text{clip}} = \cos\Big(f_{\text{CLIP}}(\hat{I}) - f_{\text{CLIP}}(I), f_{\text{CLIP}}(c)\Big). \tag{7}$$

The final semantic following reward is defined as the weighted sum of the two rewards above:

$$r_{\text{sem}} = \lambda_{\text{vlm}}\, r_{\text{vlm}} + \lambda_{\text{clip}}\, r_{\text{clip}}. \tag{8}$$

where $\lambda_{\text{vlm}}$ and $\lambda_{\text{clip}}$ are coefficients balancing the contributions of the MLLM-based semantic reward and the CLIP-based directional reward, respectively.

**Locality Preservation Reward.** To alleviate unintended changes outside the target regions, we design a locality preservation reward based on the non-edit regions to measure visual consistency. Specifically, we propose a reconstruction-based reward to capture pixel-level consistency between

original image $I$ and edited image $\hat{I}$:

$$r_{\text{diff}} = 1 - \text{L1}\left(I \odot \bar{M}, \hat{I} \odot \bar{M}\right), \tag{9}$$

This term enforces strict pixel-level consistency in the non-edit regions. Additionally, to preserve structural consistency and perceptual similarity of the edited images, we further introduce a SSIM-based structural reward and a LPIPS-based perceptual reward:

$$r_{\text{ssim}} = \text{SSIM}\Big(I \odot \bar{M},\ \hat{I} \odot \bar{M}\Big),\ r_{\text{lpips}} = -\,\text{LPIPS}\Big(I \odot \bar{M},\ \hat{I} \odot \bar{M}\Big) \tag{10}$$

The overall locality preservation reward in Edit-GRPO is formulated as the weighted sum of pixel-level reward $r_{\text{diff}}$, structural reward $r_{\text{ssim}}$ and perceptual reward $r_{\text{lpips}}$:

$$r_{\text{pres}} = \lambda_{\text{diff}}\, r_{\text{diff}} + \lambda_{\text{ssim}}\, r_{\text{ssim}} + \lambda_{\text{lpips}}\, r_{\text{lpips}}, \tag{11}$$

The scalar $\lambda_{\text{diff}}$, $\lambda_{\text{ssim}}$ and $\lambda_{\text{lpips}}$ govern the relative importance of the three preservation rewards.

### 4.4 Region-Decoupled Policy Optimization

To mitigate interference between editing and preservation objectives, we further propose a region-decoupled policy gradient optimization framework. The core idea is to apply each reward signal only to the corresponding spatial region, rather than mixing them as one reward and conducting global policy updates. Specifically, during training iterations, we first sample a mini-batch of image-instruction pairs $(I, c)$ and generate a group of $G$ edited samples $\{\hat{I}^i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid I, c)$ from the current policy. Then, for each sample $\hat{I}^i$, we infer the editing mask $M$ and compute editing reward $r_{\text{sem}}$ and preservation reward $r_{\text{pres}}$, respectively. Instead of combining them into a single scalar, we feed them into group normalization independently to obtain their corresponding advantages:

$$A_{\text{sem}} = \frac{r_{\text{sem}} - \text{mean}\big(\{r^i_{\text{sem}}\}_{i=1}^{G}\big)}{\text{std}\big(\{r^i_{\text{sem}}\}_{i=1}^{G}\big)},\ A_{\text{pres}} = \frac{r_{\text{pres}} - \text{mean}\big(\{r^i_{\text{pres}}\}_{i=1}^{G}\big)}{\text{std}\big(\{r^i_{\text{pres}}\}_{i=1}^{G}\big)}. \tag{12}$$

To ensure the editing reward supervises the edit regions while the preservation reward supervises the non-edit regions, we further decouple the policy gradient at token level. Let $\pi_\theta^{(k)}$ denote the generation probability of the $k$-th image token, we partition the token set with editing mask $M$ and compute region-aggregated log probabilities:

$$\log \pi_\theta^{\text{sem}} = \frac{1}{|M|} \sum_{k \in M} \log \pi_\theta^{(k)},\ \log \pi_\theta^{\text{pres}} = \frac{1}{|\bar{M}|} \sum_{k \in \bar{M}} \log \pi_\theta^{(k)}. \tag{13}$$

Similarly, we define the corresponding log probabilities under the old policy, denoted by $\log \pi_{\theta_{\text{old}}}^{\text{sem}}$ and $\log \pi_{\theta_{\text{old}}}^{\text{pres}}$. Based on these quantities, we construct two independent optimization objectives:

$$\mathcal{L}_{\text{sem}} = \text{ClipPG}(A_{\text{sem}}, \rho_{\text{sem}}),\ \mathcal{L}_{\text{pres}} = \text{ClipPG}(A_{\text{pres}}, \rho_{\text{pres}}), \tag{14}$$

where $\text{ClipPG}(A, \rho) \triangleq \min\big(\rho\, A,\ \text{clip}(\rho, 1 \pm \epsilon)A\big)$ denotes the GRPO-style clipped policy gradient objective, following the same structure as Eq. 4; $\rho_{\text{sem}} = \exp(\log \pi_\theta^{\text{sem}} - \log \pi_{\theta_{\text{old}}}^{\text{sem}})$ and $\rho_{\text{pres}} = \exp(\log \pi_\theta^{\text{pres}} - \log \pi_{\theta_{\text{old}}}^{\text{pres}})$ are the region-specific policy ratios. Finally, the overall training objective is defined as the weighted combination of the two region-level losses together with a KL regularization term:

$$\mathcal{L} = \lambda_{\text{sem}}\, \mathcal{L}_{\text{sem}} + \lambda_{\text{pres}}\, \mathcal{L}_{\text{pres}} + \beta \mathcal{L}_{\text{KL}}, \tag{15}$$

where $\lambda_{\text{sem}}$ and $\lambda_{\text{pres}}$ balance the relative importance of editing and preservation during optimization. The overall optimization framework is summarized in the Appendix B.

## 5 Experiment

### 5.1 Setup

**Dataset.** We adopt HumanEdit (Bai et al., 2024) as our training dataset where each sample consists of an input image and a natural language editing instruction describing the expected modification.

Table 1: Quantitative comparison on GEdit-Bench (Liu et al., 2025b). Best scores are in **bold**, second-best in underlined.

| Model | Preservation | | | | GEdit-Bench-EN | | |
|---|---|---|---|---|---|---|---|
| | UR ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | G_SC ↑ | G_PQ ↑ | G_O ↑ |
| Instruct-Pix2Pix (Brooks et al., 2022) | - | - | - | - | 3.58 | 5.49 | 3.68 |
| UniWorld-V1 (Lin et al., 2025) | - | - | - | - | 4.93 | 7.43 | 4.85 |
| OmniGen (Xiao et al., 2025) | - | - | - | - | 5.96 | 5.89 | 5.06 |
| FLUX.1-Kontext [Dev] | 57.56 | 22.36 | 0.639 | 0.201 | 6.58 | 7.43 | 6.04 |
| + Edit-R1 (Li et al., 2025b) | 52.01 | 19.53 | 0.589 | 0.236 | **7.32** | 7.54 | **6.84** |
| **+ Edit-GRPO(Ours)** | **60.30** | **22.91** | **0.663** | **0.185** | 6.81 | **7.56** | 6.34 |
| Qwen-Image-Edit [2509] | 56.69 | 21.61 | 0.616 | 0.208 | 8.23 | **7.98** | 7.70 |
| + Edit-R1 (Li et al., 2025b) | 50.97 | 20.53 | 0.597 | 0.229 | 8.34 | 7.88 | 7.77 |
| **+ Edit-GRPO(Ours)** | **63.04** | **23.07** | **0.657** | **0.182** | **8.40** | 7.94 | **7.83** |

Table 2: Quantitative comparison on ImgEdit (Ye et al., 2025). Best scores are in **bold**, second-best in underlined.

| Model | Preservation | | | | ImgEdit | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | UR ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | Add | Adjust | Extract | Replace | Remove | Background | Style | Hybrid | Action | **Overall** |
| Instruct-Pix2Pix (Brooks et al., 2022) | - | - | - | - | 2.45 | 1.83 | 1.44 | 2.01 | 1.50 | 1.44 | 3.55 | 1.20 | 1.46 | 1.88 |
| OmniGen (Xiao et al., 2025) | - | - | - | - | 3.47 | 3.04 | 1.71 | 2.94 | 2.43 | 3.21 | 4.19 | 2.24 | 3.38 | 2.96 |
| UniWorld-V1 (Lin et al., 2025) | - | - | - | - | 3.82 | 3.64 | 2.27 | 3.47 | 3.24 | 2.99 | 4.21 | 2.96 | 2.74 | 3.26 |
| FLUX.1 Kontext [Dev] | 56.67 | 22.92 | 0.690 | 0.163 | 3.67 | 3.81 | 2.10 | 4.18 | 3.01 | 3.75 | 4.42 | 2.61 | 3.92 | 3.52 |
| + Edit-R1 (Li et al., 2025b) | 45.84 | 18.82 | 0.581 | 0.221 | 3.86 | 3.90 | 2.67 | 4.40 | 3.39 | 4.07 | 4.59 | 3.04 | 4.12 | **3.82** |
| **+ Edit-GRPO(Ours)** | **59.65** | **23.25** | **0.706** | **0.161** | 3.68 | 3.88 | 2.47 | 4.31 | 3.33 | 3.85 | 4.35 | 2.78 | 3.92 | 3.65 |
| Qwen-Image-Edit [2509] | 47.70 | 19.68 | 0.567 | 0.220 | 4.31 | 4.37 | 3.33 | 4.60 | 4.42 | 4.26 | 4.81 | 3.61 | 4.27 | 4.26 |
| + Edit-R1 (Li et al., 2025b) | 38.18 | 17.91 | 0.528 | 0.256 | 4.34 | 4.21 | 3.86 | 4.66 | 4.48 | 4.20 | 4.88 | 3.42 | 4.47 | **4.34** |
| **+ Edit-GRPO(Ours)** | **51.81** | **20.65** | **0.608** | **0.204** | 4.27 | 4.40 | 3.42 | 4.66 | 4.43 | 4.26 | 4.85 | 3.55 | 4.56 | 4.30 |

The dataset covers diverse editing types, including object addition, removal, replacement, action modification, and counting-based edits, providing a broad range of editing scenarios.

**Implementation Details.** We use FLUX.1-Kontext [Dev] (Labs et al., 2025) and Qwen-Image-Edit [2509] (Wu et al., 2025a) as our base models. For editing mask generation, we employ Qwen2.5-14B-Instruct (Yang et al., 2024) as the object prompt extractor and SAM3 (Carion et al., 2025) as the segmentation model. We employ Qwen3-VL-32B-Instruct (Bai et al., 2025) as the MLLM-based score evaluator. More implementation details are provided in the Appendix A.

**Semantic Editing Evaluation.** For quantitative semantic editing evaluation, we employ two representative benchmarks: GEdit-Bench (Liu et al., 2025b), which evaluates image editing through diverse language instructions, and ImgEdit (Ye et al., 2025), which unifies multiple tasks into a common comparison framework.

**Locality Preservation Evaluation.** We introduce several locality-aware metrics to quantify locality preservation. First, we introduce **Unchanged Ratio (UR)**, defined as the proportion of pixels in non-edit regions whose absolute difference falls below a small threshold before and after editing. In addition, we further report PSNR, SSIM and LPIPS to measure low-level fidelity, structural consistency, and perceptual similarity, respectively. More evaluation details are provided in Appendix A.2.

### 5.2 Main Results

We evaluate Edit-GRPO on both semantic editing and locality preservation. Quantitative results are reported in Table 1 and Table 2. Qualitative comparisons are shown in Figure 4. We further visualize the smoothed reward trajectory during training in Figure 5.

**Global reward optimization improves editing scores at the expense of locality preservation.** As shown in Table 1 and Table 2, the global optimization method Edit-R1 (Li et al., 2025b) improves semantic editing scores while substantially harming locality preservation. On GEdit-Bench, applying Edit-R1 to FLUX.1-Kontext [Dev] increases G_O to 6.84, but causes clear degradation in all preservation metrics (UR: 57.56% → 52.01%, PSNR: 22.36 → 19.53, SSIM: 0.639 → 0.589, LPIPS: 0.201 → 0.236). Similar experimental results are observed for Qwen-Image-Edit [2509] and on the ImgEdit benchmark. These results indicate that global optimization tends to encourage aggressive edits, introducing substantial content modifications outside the target regions.

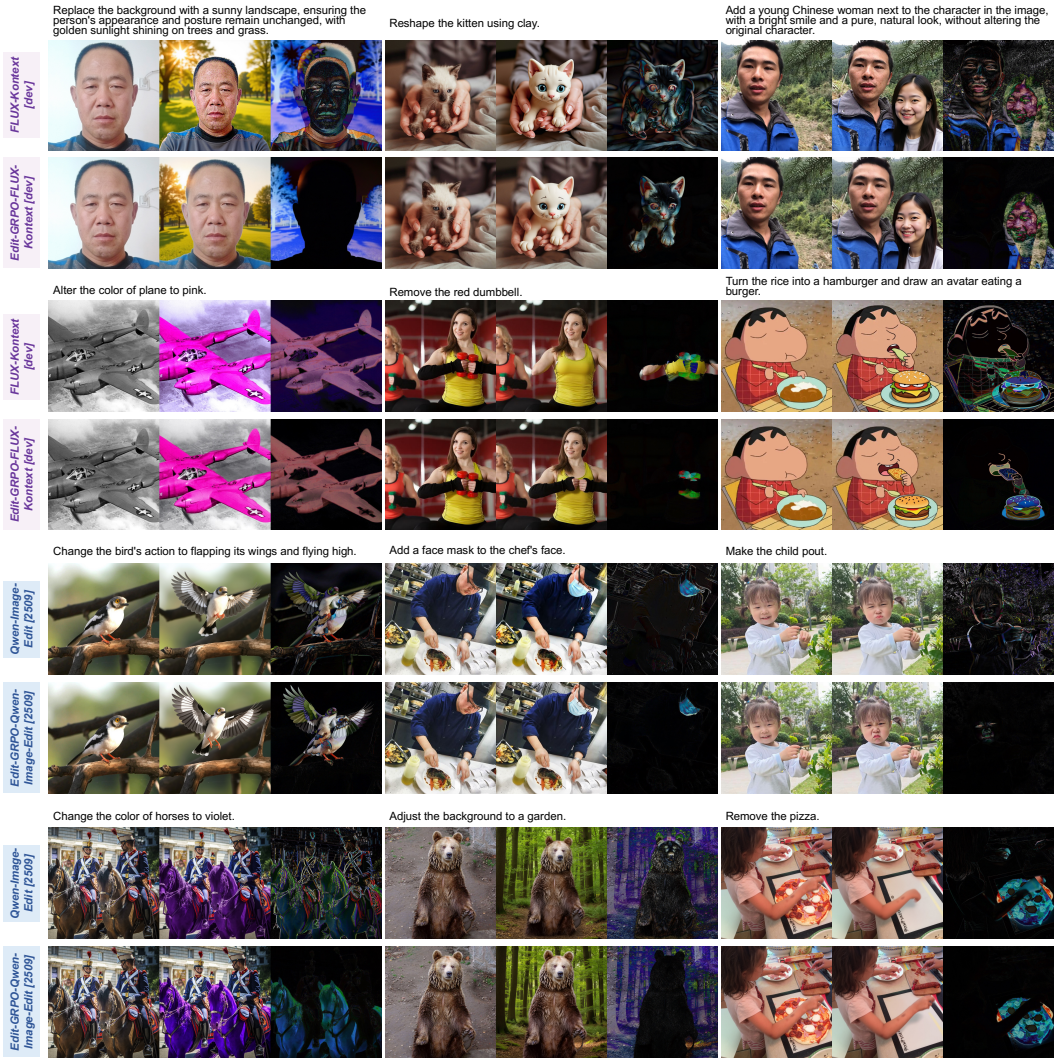


Figure 4: Qualitative comparison of editing capabilities before and after policy optimization. Our approach demonstrates superior semantic editing and locality preservation over the base models.

**Edit-GRPO consistently improves locality preservation while retaining competitive editing performance.** On GEdit-Bench (Table 1), Edit-GRPO increases UR from 57.56% to 60.30% for FLUX.1-Kontext [Dev] and from 56.69% to 63.04% for Qwen-Image-Edit [2509]. Meanwhile, semantic editing performance is preserved or improved relative to the base models, with G_O increasing from 6.04 to 6.34 for FLUX.1-Kontext [Dev] and from 7.70 to 7.83 for Qwen-Image-Edit [2509]. Similar trends are observed on ImgEdit (Table 2), where Edit-GRPO improves all preservation metrics while maintaining competitive overall editing scores. These results demonstrate that Edit-GRPO effectively improves locality preservation while retaining competitive editing performance, mitigating the preservation-editing trade-off in prior approaches.

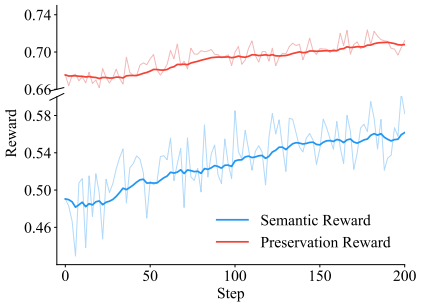


Figure 5: Training reward curves on FLUX.1-Kontext [Dev].

**Qualitative results.** As illustrated in Figure 4, Edit-GRPO produces edits that are more localized to the target regions and better preserve surrounding content compared to the base models, further demonstrating that our method can improve locality preservation while maintaining editing precision.

Table 3: Ablation on optimization objectives. Best scores are in **bold**, second-best in underlined.

| Loss Design | UR ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | G_SC ↑ | G_PQ ↑ | G_O ↑ |
|---|---|---|---|---|---|---|---|
| FLUX.1-Kontext [Dev] | 57.56 | 22.36 | 0.639 | 0.201 | 6.58 | 7.43 | 6.04 |
| Semantic-Only | 52.54 | 21.23 | 0.624 | 0.214 | **7.02** | 7.29 | **6.44** |
| Preservation-Only | **64.10** | **24.49** | **0.700** | **0.176** | 5.42 | **7.76** | 5.09 |
| Combined Optimization | 57.90 | 22.80 | 0.658 | 0.191 | 6.63 | 7.54 | 6.15 |
| Decoupled Optimization | 60.30 | 22.91 | 0.663 | 0.185 | 6.81 | 7.56 | 6.34 |

Table 4: Ablation on semantic reward designs.

| Reward Design | G_SC ↑ | G_PQ ↑ | G_O ↑ |
|---|---|---|---|
| FLUX.1-Kontext [Dev] | 6.58 | 7.43 | 6.04 |
| Preservation-Only | 5.42 | **7.76** | 5.09 |
| + CLIP reward | 6.35 | 7.50 | 5.94 |
| + MLLM (discrete) reward | 6.67 | 7.39 | 6.14 |
| + MLLM (logit) reward | **6.81** | 7.56 | **6.34** |

Table 5: Ablation on preservation reward designs.

| Reward Design | UR ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| FLUX.1-Kontext [Dev] | 57.56 | 22.36 | 0.639 | 0.201 |
| Semantic-Only | 52.54 | 21.23 | 0.624 | 0.214 |
| +LPIPS reward | 53.09 | 21.66 | 0.638 | 0.202 |
| +SSIM reward | 53.47 | 22.08 | 0.653 | 0.192 |
| + L1 reward | **60.30** | **22.91** | **0.663** | **0.185** |

### 5.3 Ablation Study

In this section, we conduct a systematic analysis to evaluate the contribution of each component in Edit-GRPO. All experiments are trained on FLUX.1-Kontext [Dev] and evaluated on GEdit-Bench.

**Decoupled Optimization vs. Global Optimization.** We first study different objective designs. As shown in Table 3, Semantic-Only optimization achieves the highest editing score (G_O 6.44) but substantially degrades locality preservation (UR: 57.56% → 52.54%), while Preservation-Only optimization yields the best preservation metrics (UR 64.10%) but significantly hurts editing performance (G_O: 6.04 → 5.09). This phenomenon indicates a clear conflict between the two objectives when optimized alone. In contrast, our decoupled optimization improves preservation metrics (UR 60.30%) while maintaining a strong editing score (G_O 6.34), enabling simultaneous optimization of semantic editing and locality preservation. Notably, while mixing the two rewards into a scalar for global optimization yields slight improvements (G_O: 6.04 → 6.15, UR: 57.56% → 57.90%), decoupled optimization achieves stronger improvements (G_O: 6.34, UR: 60.30%), demonstrating the effectiveness of the proposed decoupled optimization framework.

**Semantic Reward Design.** We ablate different semantic reward designs in Table 4. Without semantic reward, the model suffers from a clear performance degradation in semantic following (G_O: 6.04 → 5.09). Adding the CLIP-based directional reward improves G_O to 5.94 and incorporating the MLLM-based discrete reward further improves G_O to 6.14. Replacing the MLLM-based discrete reward with logit reward further improves G_O from 6.14 to 6.34, suggesting that the logit reward can provide more fine-grained signals for instruction-compliant editing.

**Preservation Reward Design.** We further ablate preservation reward design in Table 5. Without preservation reward, locality preservation degrades substantially compared to the base model (UR: 57.56% → 52.54%). Adding the LPIPS reward improves perceptual preservation, reducing LPIPS from 0.214 to 0.202. Further incorporating the SSIM reward strengthens structural consistency, achieving SSIM of 0.653. Finally, adding the pixel-level L1 reward achieves the largest improvement in locality preservation, increasing UR from 53.47% to 60.30%. These results suggest that perceptual, structural, and pixel-level rewards provide complementary preservation signals, with the L1 reward playing a crucial role in suppressing unintended modifications in non-edit regions.

## 6 Conclusion

In this paper, we present Edit-GRPO, a general locality-preserving policy optimization framework for image editing. By explicitly aligning optimization signals with the spatial structure of editing tasks, Edit-GRPO enables targeted improvements in editing quality while effectively preserving surrounding context. Extensive experiments demonstrate that region-decoupled optimization is a principled and effective solution to the long-standing locality challenge in image editing. We hope this work inspires future research on spatially aware optimization strategies for generative models.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with constructing an evaluation subset, as detailed in Appendix D.4. We did not use generative AI tools for any other tasks requiring disclosure under the ICLR policy that are applicable to this work. Additionally, we used generative AI tools to assist with drafting parts of the manuscript and editing the text to improve readability, fluency, grammar, spelling, and word choice.

We reviewed and verified all AI-assisted text for accuracy and consistency with the intended scientific meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Jinbin Bai, Wei Chow, Ling Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Shuicheng Yan. Humanedit: A high-quality human-rewarded dataset for instruction-based image editing. *arXiv preprint arXiv:2412.04280*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Feng Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zi-Qiang Gong, Jia-Wern Pan, Yingzhi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *ArXiv*, abs/2505.22705, 2025. URL https://api.semanticscholar.org/CorpusID:278996234.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiao-Ping Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *ArXiv*, abs/2505.14683, 2025. URL https://api.semanticscholar.org/CorpusID:278768720.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, P. Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning

for fine-tuning text-to-image diffusion models. *ArXiv*, abs/2305.16381, 2023. URL https://api.semanticscholar.org/CorpusID:258947323.

Yuan Gong, Xionghui Wang, Jie Wu, Shiyin Wang, Yitong Wang, and Xinglong Wu. Onereward: Unified mask-guided image generation via multi-task human preference learning. *arXiv preprint arXiv:2508.21066*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025a.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Feize Wu, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, et al. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025b.

Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yunyang Ge, Yatian Pang, and Li Yuan. Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation. *ArXiv*, abs/2506.03147, 2025. URL https://api.semanticscholar.org/CorpusID:279119654.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025a.

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*, 2025b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://api.semanticscholar.org/CorpusID:245704504.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2022. URL https://api.semanticscholar.org/CorpusID:254854389.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023. URL https://api.semanticscholar.org/CorpusID:258959321.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017. URL https://api.semanticscholar.org/CorpusID:28695052.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020a. URL https://api.semanticscholar.org/CorpusID:222140788.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2020b. URL https://api.semanticscholar.org/CorpusID:227209335.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq R. Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238, 2023. URL https://api.semanticscholar.org/CorpusID:265352136.

Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Instructedit: Improving automatic masks for diffusion-based image editing with user instructions. *ArXiv*, abs/2305.18047, 2023. URL https://api.semanticscholar.org/CorpusID:258959425.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 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.

Jie Wu, Yu Gao, Zi-Nuo Ye, Ming Li, Liang Li, Hanzhong Guo, Jie Liu, Zeyue Xue, Xiaoxia Hou, Wei Liu, Yangyang Zeng, and Weilin Huang. Rewarddance: Reward scaling in visual generation. *ArXiv*, abs/2509.08826, 2025b. URL https://api.semanticscholar.org/CorpusID:281247213.

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*, 2025c.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

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. *ArXiv*, abs/2304.05977, 2023. URL https://api.semanticscholar.org/CorpusID:258079316.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *ArXiv*, abs/2505.20275, 2025. URL https://api.semanticscholar.org/CorpusID:278911803.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *ArXiv*, abs/2306.10012, 2023. URL https://api.semanticscholar.org/CorpusID:259187796.

Qihui Zhang, Munan Ning, Zheyuan Liu, Yanbo Wang, Jiayi Ye, Yue Huang, Shuo Yang, Xiao Chen, Yibing Song, and Li Yuan. Upme: An unsupervised peer review framework for multimodal large language model evaluation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9165–9174, 2025. URL https://api.semanticscholar.org/CorpusID:277113471.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

# A Implementation Details

## A.1 Training Details

In our implementation, we use FLUX.1-Kontext [Dev] (Labs et al., 2025) and Qwen-Image-Edit [2509] (Wu et al., 2025a) as our base models. For editing mask generation, we employ Qwen2.5-14B-Instruct (Yang et al., 2024) as the object prompt extractor and SAM3 (Carion et al., 2025) as the segmentation model. We employ Qwen3-VL-32B-Instruct (Bai et al., 2025) as the MLLM-based score evaluator for the edited images. All images are resized to a resolution of $512 \times 512$ during training. To improve memory efficiency, we adopt Fully Sharded Data Parallelism (FSDP) for the text encoder and enable gradient checkpointing when training Qwen-Image-Edit [2509]. All training experiments are conducted in `bfloat16` mixed precision.

We use the AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of $3 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $10^{-4}$. We apply a KL regularization term with coefficient $\beta = 1 \times 10^{-4}$ to constrain the policy from deviating excessively from the reference model. We maintain an exponential moving average (EMA) of model parameters with decay 0.9. We adopt per-prompt normalization with global standard deviation across the group to reduce variance. For the reward function, we set the semantic reward weights to $\lambda_{\text{vlm}} = 4.0$ and $\lambda_{\text{clip}} = 1.0$, and the preservation reward weights to $\lambda_{\text{diff}} = 4.0$, $\lambda_{\text{ssim}} = 2.0$, and $\lambda_{\text{lpips}} = 1.0$. The region-decoupled loss weights are set to $\lambda_{\text{sem}} = 0.5$ for the semantic branch and $\lambda_{\text{pres}} = 1.0$ for the preservation branch.

## A.2 Evaluation Details

The evaluation experiments consists of two distinct components: semantic editing evaluation and locality preservation evaluation. For semantic editing evaluation, we strictly follow the GEdit-Bench (Liu et al., 2025b) and the ImgEdit (Ye et al., 2025) benchmarks, using GPT4.1 for the semantic score evaluation. For locality preservation evaluation, we design an auto mask-based evaluation pipeline. Specifically, we first use Qwen2.5-14B-Instruct (Yang et al., 2024) to parse the editing instruction and extract structured descriptions of the target content. These descriptions and images are then fed into the SAM3 (Carion et al., 2025) model to generate high-resolution masks for the edit target. To acquire robust segmentation in diverse editing scenarios, we apply region decoupling to both the original image and the edited image, obtaining two editing masks accordingly. The final segmentation mask is computed as the union of the two masks. All the remaining areas are treated as the non-edit regions.

Based on the region segmentation masks, we compute the following locality preservation metrics exclusively in non-edit regions:

**Unchanged Ratio (UR).** This metric measures the proportion of pixels in non-edit regions that remain nearly unchanged after editing. For each pixel $p$ in non-edit regions $\bar{M}$, we compute the L1 distance across RGB channels between the original image $I$ and the edited image $\hat{I}$. A pixel is considered unchanged if the distance falls below a threshold $\tau$. In our implementation, we set $\tau = 20$ in RGB pixel space. Formally, the UR metric is defined as:

$$\text{UR} = \frac{\left|\left\{p \in \bar{M} : \|I(p) - \hat{I}(p)\|_1 < \tau\right\}\right|}{|\bar{M}|}. \tag{16}$$

**PSNR.** Peak Signal-to-Noise Ratio measures pixel-level reconstruction fidelity in non-edit regions. We compute the mean squared error (MSE) between the original image $I$ and the edited image $\hat{I}$ over pixels in the non-edit regions $\bar{M}$ across all RGB channels:

$$\text{PSNR} = 10 \log_{10}\left(\frac{255^2}{\text{MSE}}\right), \tag{17}$$

where $\text{MSE} = \frac{1}{3|\bar{M}|} \sum_{p \in \bar{M}} \|I(p) - \hat{I}(p)\|_2^2$.

**SSIM.** Structural Similarity Index (SSIM) evaluates the structural consistency between the original and edited images. Following the standard protocol, we convert both images to grayscale and compute SSIM within the non-edit regions $\bar{M}$.

Table 6: Hyperparameter settings for Edit-GRPO. We report the configurations for both FLUX.1-Kontext [Dev] and Qwen-Image-Edit [2509] backbones.

| **Hyperparameter** | **FLUX.1-Kontext [Dev]** | **Qwen-Image-Edit [2509]** |
|---|---|---|
| | *Training Settings* | |
| Batch Size (per device) | 1 | 1 |
| Gradient Accumulation Steps | 24 | 24 |
| Learning Rate | $3 \times 10^{-4}$ | $3 \times 10^{-4}$ |
| Adam ($\beta_1$, $\beta_2$) | $0.9, 0.999$ | $0.9, 0.999$ |
| Weight Decay | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ |
| EMA Decay | 0.9 | - |
| | *Sampling Settings* | |
| Image Resolution | $512 \times 512$ | $512 \times 512$ |
| Sampling Inference Steps | 6 | 10 |
| Images Per Prompt | 8 | 8 |
| Noise Level | 0.9 | 0.7 |
| | *Reward Weights* | |
| MLLM Semantic Score | 4.0 | 4.0 |
| CLIP Directional Score | 1.0 | 1.0 |
| L1 Preservation Score | 4.0 | 4.0 |
| SSIM Preservation Score | 2.0 | 2.0 |
| LPIPS Preservation Score | 1.0 | 1.0 |
| | *Loss Coefficients* | |
| Semantic Loss | 0.5 | 0.5 |
| Preservation Loss | 1.0 | 1.0 |
| KL Divergence ($\beta$) | $1 \times 10^{-4}$ | $1 \times 10^{-4}$ |

**LPIPS.** Learned Perceptual Image Patch Similarity (LPIPS) measures perceptual similarity using deep visual features. Following the standard LPIPS implementation, images are first normalized to the range $[-1, 1]$ and then passed through a pretrained VGG-based network (Zhang et al., 2018). The perceptual distance is computed in non-edit regions $\bar{M}$, where lower LPIPS values indicate better perceptual preservation.

In practice, when conducting locality preservation evaluation, we explicitly **exclude** samples whose task type belongs to *style_change*, *text_change* and *tone_transfer* as these editing categories involve global or diffuse transformations where SAM3 (Carion et al., 2025) cannot accurately segment a well-defined local edit region, making the mask-based locality evaluation unreliable. We further apply two filtering criteria to handle cases where the generated segmentation masks are inaccurate: (i) samples where the non-edit regions cover less than 5% of the image are discarded, as they likely represent near-global edits for which mask-based evaluation is not meaningful; (ii) samples where the edit regions cover less than 1% of the image are also discarded, as this typically indicates that SAM3 fails to localize the edited object and generate a meaningful segmentation mask.

### A.3 Hyperparameters and Computation Resources

We report the key hyperparameters of Edit-GRPO in Table 6. All training and evaluation experiments are conducted on a single machine with 8 NVIDIA H20 96 GB GPUs.

## B Algorithm of Edit-GRPO

We provide the overall algorithm of the proposed Edit-GRPO in Algorithm 1.

## C Prompt Template

We provide the prompt template used for MLLM-based semantic score evaluation in Figure 6.

**Algorithm 1** Region-Decoupled GRPO Training (**Edit-GRPO**).

**Require:** Editing policy $\pi_\theta$, dataset $\{(I, c)\}$, region inference module $\mathcal{S}$, semantic evaluator $\mathcal{E}_{\text{sem}}$, preservation evaluator $\mathcal{E}_{\text{pres}}$, group size $G$.
**Initialize:** Reference policy $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$.

**for** each training iteration **do**
  **for** each sampled pair $(I, c)$ **do** ▷ Rollout Step
    Sample edited images $\{\hat{I}^i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|I, c)$.
    $M^i \leftarrow \mathcal{S}(I, c) \cup \mathcal{S}(\hat{I}^i, c)$.
    $r^i_{\text{sem}} \leftarrow \mathcal{E}_{\text{sem}}(\hat{I}^i, c, M^i), \quad r^i_{\text{pres}} \leftarrow \mathcal{E}_{\text{pres}}(\hat{I}^i, I, \bar{M}^i)$.
    $\log \pi^{\text{sem},i}_{\theta_{\text{old}}} \leftarrow \text{Mean}_{k\in M^i}\big[\log \pi^{(k)}_{\theta_{\text{old}}}\big], \quad \log \pi^{\text{pres},i}_{\theta_{\text{old}}} \leftarrow \text{Mean}_{k\in \bar{M}^i}\big[\log \pi^{(k)}_{\theta_{\text{old}}}\big]$.
  **end for**
  **for** each training mini-batch **do** ▷ Policy Update
    $A_{\text{sem}}, A_{\text{pres}} \leftarrow \text{GroupNorm}(r_{\text{sem}}, r_{\text{pres}})$.
    $\log \pi^{\text{sem}}_\theta \leftarrow \text{Mean}_{k\in M}\big[\log \pi^{(k)}_\theta\big], \quad \log \pi^{\text{pres}}_\theta \leftarrow \text{Mean}_{k\in \bar{M}}\big[\log \pi^{(k)}_\theta\big]$.
    $\rho_{\text{sem}} = \exp(\log \pi^{\text{sem}}_\theta - \log \pi^{\text{sem}}_{\theta_{\text{old}}}), \quad \rho_{\text{pres}} = \exp(\log \pi^{\text{pres}}_\theta - \log \pi^{\text{pres}}_{\theta_{\text{old}}})$.
    $\mathcal{L}_{\text{sem}} = \text{ClipPG}(A_{\text{sem}}, \rho_{\text{sem}}), \quad \mathcal{L}_{\text{pres}} = \text{ClipPG}(A_{\text{pres}}, \rho_{\text{pres}})$.
    $\theta \leftarrow \theta - \lambda\nabla_\theta\big(\lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \beta\mathcal{L}_{\text{KL}}\big)$.
  **end for**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$.
**end for**

**Output:** Optimized policy $\pi_\theta$.

**Template for MLLM-based Semantic Score Evaluation:**
You are an image editing evaluator. Evaluate how well the edited image follows the instruction.
**Original Image**: [First image]
**Edited Image**: [Second image]
**Instruction**: {instruction}
You need to rate the editing result from 0 to 5 based on the accuracy and quality of the edit.
**Scoring Rules:**
0: The wrong object was edited, or the edit completely fails to meet the requirements.
5: The correct object was edited, the requirements were met, and the visual result is high quality.
**Response Format:**
Directly output the score number (0–5).

Figure 6: Prompt template used for MLLM-based semantic score evaluation.

# D More Experiments

## D.1 Human preference evaluation

To better align our evaluation with human perception, we conduct an additional user study. Each participant evaluates cases spanning diverse editing scenarios, including object addition, removal, replacement, pose changes, and local attribute modifications. Each case consists of an original image, an editing instruction, and two anonymized edited images generated by the base model and the optimized model. Participants are asked to independently compare the two edited images in terms of edit faithfulness, locality preservation, and overall quality and to select the better output or indicate no clear preference. As shown in Table 7, participants consistently prefer Edit-GRPO over the base model across all three evaluation criteria, with preference rates of 45.0% versus 35.0% for edit faithfulness, 39.2% versus 35.0% for locality preservation, and 48.8% versus 33.8% for overall quality. These results demonstrate that Edit-GRPO follows editing instructions more accurately while better preserving non-edit regions.

## D.2 Evaluation with trained reward model

We further evaluate the overall editing performance of Edit-GRPO using EditReward (Wu et al., 2025c), a human-aligned reward model developed specifically for instruction-guided image editing.

Table 7: Human preference evaluation in terms of across edit faithfulness, locality preservation, and overall quality. Values indicate the percentage of participant preferences under each criterion (higher is better).

| Preference | Edit Faithfulness | Locality Preservation | Overall Quality |
|---|---|---|---|
| No Preference | 20.0% | 25.8% | 17.5% |
| Base Model | 35.0% | 35.0% | 33.8% |
| Edit-GRPO | **45.0%** | **39.2%** | **48.8%** |

Table 8: Evaluation on EditReward.

| Method | EditReward ↑ |
|---|---|
| FLUX.1-Kontext [Dev] | 0.88 |
| + Edit-GRPO | **0.92** |
| Qwen-Image-Edit [2509] | 1.40 |
| + Edit-GRPO | **1.45** |

Table 9: Evaluation with SAM2-based segmentation pipeline.

| Method | UR ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|
| FLUX.1-Kontext [Dev] | 55.70 | 21.72 | 0.694 | 0.197 |
| + Edit-GRPO | **58.34** | **22.29** | **0.719** | **0.185** |
| Qwen-Image-Edit [2509] | 54.38 | 20.881 | 0.677 | 0.205 |
| + Edit-GRPO | **61.01** | **22.42** | **0.734** | **0.177** |

For each edited result, we provide EditReward with the source image, editing instruction, and edited image, and report the mean reward score over GEdit-Bench. As shown in Table 8, Edit-GRPO increases the EditReward score from 0.88 to 0.92 for FLUX.1-Kontext [Dev] and from 1.40 to 1.45 for Qwen-Image-Edit [2509]. These consistent improvements provide complementary evidence that Edit-GRPO improves overall editing quality under a learned human-preference evaluator.

### D.3 Evaluation with independent segmentation pipeline

To evaluate whether the locality preservation improvements depend on the segmentation model used during training, we perform an additional evaluation using an independent SAM2-based segmentation pipeline. Specifically, we use the SAM2 (Ravi et al., 2024) model to generate edit masks and recompute the locality preservation metrics on GEdit-Bench. This pipeline is used only for evaluation and is independent of the SAM3-based masks employed during training. As shown in Table 9, Edit-GRPO consistently improves UR, PSNR, and SSIM while reducing LPIPS for both models. These results show that Edit-GRPO improves locality preservation when evaluated with an independent segmentation pipeline, suggesting that the preservation gains are not limited to the SAM3-based masks used during training.

### D.4 Evaluation on geometry-changing edits

To further investigate the performance on *geometry-changing edits*, where a geometrically complex edit requires contextual changes outside the edit mask, such as updated shadows, reflections, occlusion boundaries, or newly exposed background, we provide an additional quantitative evaluation focused on this challenging setting. We use GPT-5.6-Sol to automatically identify geometry-sensitive editing cases in GEdit-Bench and retain 33 cases spanning four categories:

- **Shadow/illumination** (6 cases): edits expected to alter lighting or cast shadows.
- **Reflection** (4 cases): edits involving reflective surfaces or scene elements whose reflections should change.
- **Occlusion boundaries** (9 cases): edits involving contact, overlap, pose, or spatial-relation changes.
- **Background completion** (14 cases): object-removal edits requiring reconstruction of newly exposed content.

As reported in Table 10, Edit-GRPO consistently improves overall editing performance (G_O: $6.23 \rightarrow 7.16$ for FLUX.1-Kontext and $8.56 \rightarrow 8.61$ for Qwen-Image-Edit) while largely maintaining or improving locality preservation. In particular, across the four geometry-sensitive categories, Edit-GRPO improves PSNR, SSIM, and LPIPS on FLUX.1-Kontext, and improves all four preservation metrics on Qwen-Image-Edit. For FLUX.1-Kontext, UR slightly decreases on shadow/illumination

Table 10: Evaluation on geometry-changing edits.

| Backbone | Challenge | Method | UR ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ | G_SC ↑ | G_PQ ↑ | G_O ↑ |
|---|---|---|---|---|---|---|---|---|---|
| FLUX.1-Kontext | Shadow/illumination | Base | 44.60 | 19.16 | 0.528 | 0.312 | 4.50 | 8.17 | 4.76 |
| | | Ours | 42.84 | 19.72 | 0.547 | 0.288 | 7.50 | 8.17 | 7.33 |
| | Reflection | Base | 59.19 | 17.59 | 0.638 | 0.169 | 5.00 | 8.25 | 5.39 |
| | | Ours | 58.05 | 17.90 | 0.646 | 0.154 | 7.00 | 8.00 | 7.12 |
| | Occlusion boundaries | Base | 67.16 | 23.07 | 0.691 | 0.201 | 9.00 | 6.78 | 7.69 |
| | | Ours | 67.96 | 24.03 | 0.695 | 0.185 | 9.00 | 7.11 | 7.87 |
| | Background completion | Base | 61.51 | 23.45 | 0.681 | 0.156 | 6.50 | 8.64 | 6.16 |
| | | Ours | 61.86 | 23.51 | 0.683 | 0.167 | 7.00 | 8.43 | 6.65 |
| | Overall | Base | **59.69** | 21.85 | 0.651 | 0.198 | 6.64 | **8.00** | 6.23 |
| | | Ours | 59.60 | **22.28** | **0.657** | **0.192** | **7.64** | 7.97 | **7.16** |
| Qwen-Image-Edit | Shadow/illumination | Base | 44.27 | 19.13 | 0.479 | 0.328 | 8.67 | 8.50 | 8.56 |
| | | Ours | 45.63 | 18.89 | 0.485 | 0.305 | 9.00 | 8.83 | 8.91 |
| | Reflection | Base | 61.93 | 19.49 | 0.587 | 0.192 | 9.00 | 8.75 | 8.84 |
| | | Ours | 52.14 | 18.40 | 0.552 | 0.205 | 9.00 | 8.50 | 8.68 |
| | Occlusion boundaries | Base | 55.80 | 21.52 | 0.634 | 0.262 | 9.56 | 7.44 | 8.32 |
| | | Ours | 59.36 | 21.32 | 0.653 | 0.240 | 9.56 | 7.44 | 8.34 |
| | Background completion | Base | 51.52 | 19.18 | 0.528 | 0.242 | 9.07 | 8.79 | 8.63 |
| | | Ours | 61.29 | 20.93 | 0.604 | 0.206 | 8.93 | 9.00 | 8.63 |
| | Overall | Base | 52.63 | 19.84 | 0.555 | 0.257 | 9.12 | 8.36 | 8.56 |
| | | Ours | **56.81** | **20.36** | **0.590** | **0.233** | **9.12** | **8.48** | **8.61** |

($44.60 \rightarrow 42.84$) and reflection ($59.19 \rightarrow 58.05$) edits. This likely occurs because Edit-GRPO performs the intended semantic edits more faithfully, as indicated by the substantial G_O improvements ($4.76 \rightarrow 7.33$ and $5.39 \rightarrow 7.12$), thereby introducing changes to shadows and reflections that are counted as pixel differences by UR. These results further demonstrate that our region-decoupled optimization framework remains effective for complex geometry-changing edits. Although the inherent trade-off between precise editing and locality preservation cannot be completely eliminated, Edit-GRPO improves locality preservation while maintaining strong editing performance and visual consistency as much as possible.

# E Limitations

Despite the improvements achieved by our method, several limitations remain to be addressed. First, the effectiveness of our Edit-GRPO relies on the quality of the segmentation masks used to separate edit and non-edit areas. However, in practice, the quality of automatically generated masks can vary depending on the complexity of the editing scenario. Imperfect masks may lead to inaccurate region assignments, which can affect the stability and effectiveness of the optimization process.

Second, although our region-decoupled optimization framework can alleviate the conflict between semantic editing and locality preservation objectives, this conflict is not eliminated completely. Future work could explore more advanced collaborative optimization strategies to better balance editing strength and background preservation, further improving the robustness of instruction-based image editing systems.

# F Societal Impacts

## F.1 Positive Impacts

The proposed Edit-GRPO framework improves the controllability and reliability of instruction-based image editing by explicitly optimizing both semantic editing and locality preservation without requiring additional manual annotations. By decoupling editing and preservation objectives during policy optimization, our method encourages models to perform targeted modifications while maintaining the surrounding visual context. This improvement can benefit a range of applications such as digital content creation, photo retouching, graphic design, and e-commerce visualization, where localized

edits are often required while the rest of the image should remain unchanged. Beyond professional applications, improved locality-aware editing can also make visual editing tools more accessible to non-expert users. With natural language instructions, users can perform targeted edits without extensive experience in professional editing software, potentially lowering the barrier to creative workflows and improving productivity. In addition, explicitly optimizing preservation objectives promotes more predictable editing behavior, which helps reduce unintended visual artifacts and improve user trust in AI-assisted editing systems.

### F.2 Negative Impacts

Despite these benefits, the proposed method may also introduce potential societal risks. More precise and realistic image editing capabilities could be misused to manipulate visual content in subtle ways while preserving most of the original scene. Such edits may become harder to detect and could facilitate the creation of misleading or deceptive visual media, including manipulated photographs or visual disinformation. Furthermore, more precise and realistic image editing may also be used to modify personal or sensitive images, raising privacy and ethical concerns.

## G More Visualization Results

We present more visualization results of Edit-GRPO-FLUX-Kontext [dev] and Edit-GRPO-Qwen-Image-Edit [2509] in Figure 7 and Figure 8. We also provide global editing results in Figure 9, showing that the Edit-GRPO framework retains the ability to perform global contextual editing.

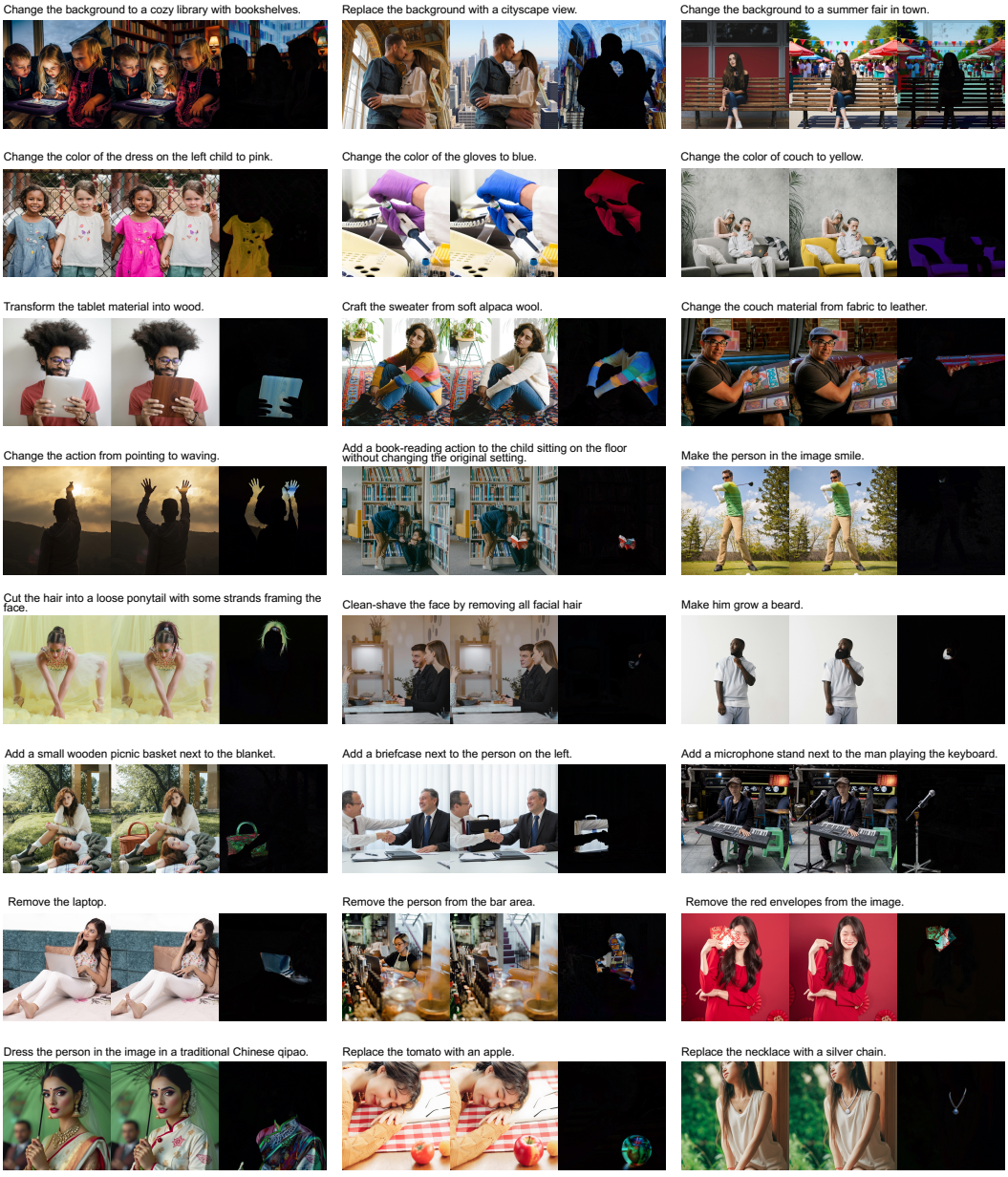


Figure 7: The visualization results of Edit-GRPO-FLUX-Kontext [dev].




Figure 8: The visualization results of Edit-GRPO-Qwen-Image-Edit [2509].

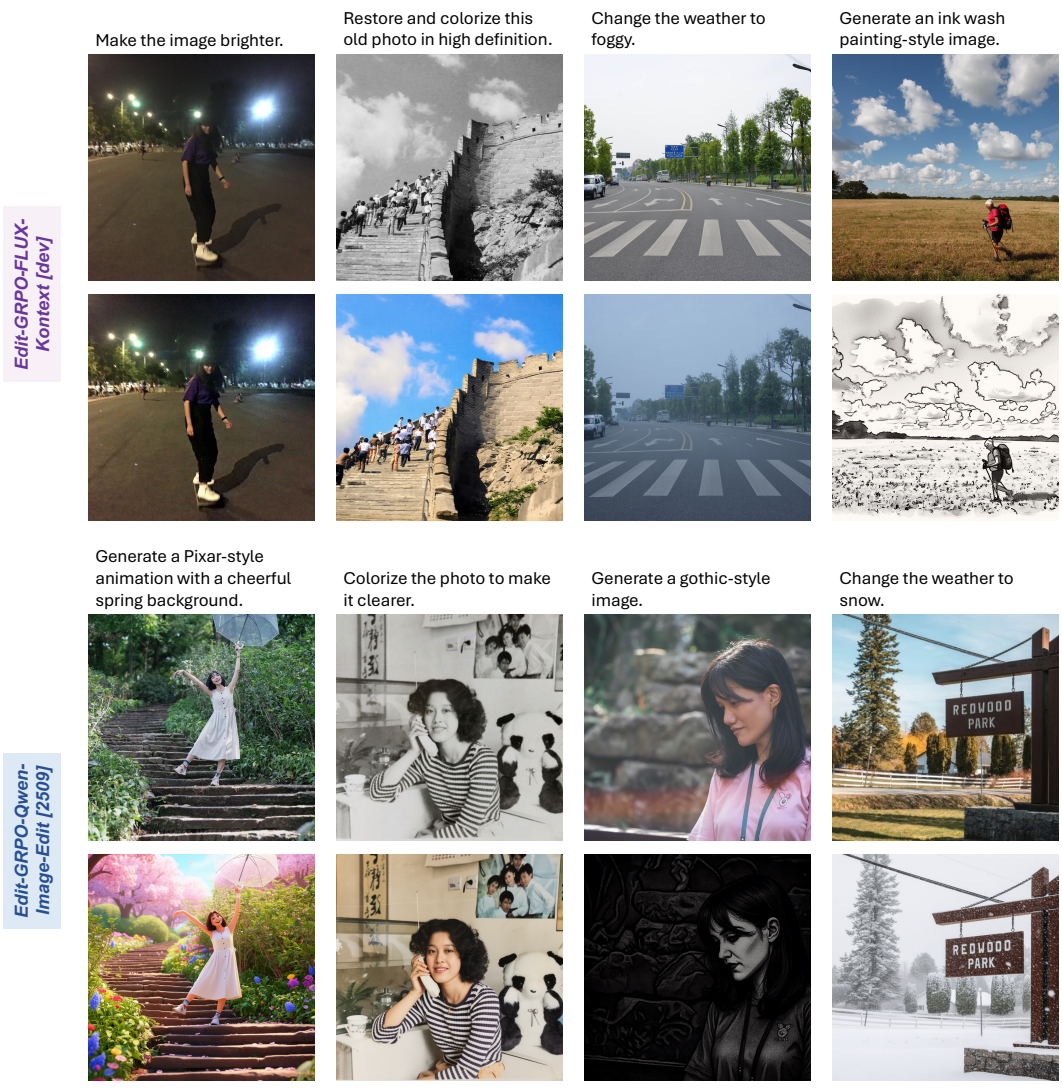


Figure 9: The visualization results of global contextual editing.