PREFEDIT: FINE-GRAINED HUMAN-CENTRIC IMAGE EDITING WITH FAILURE-AWARE RUBRIC REWARDS

Anonymous authors
Paper under double-blind review

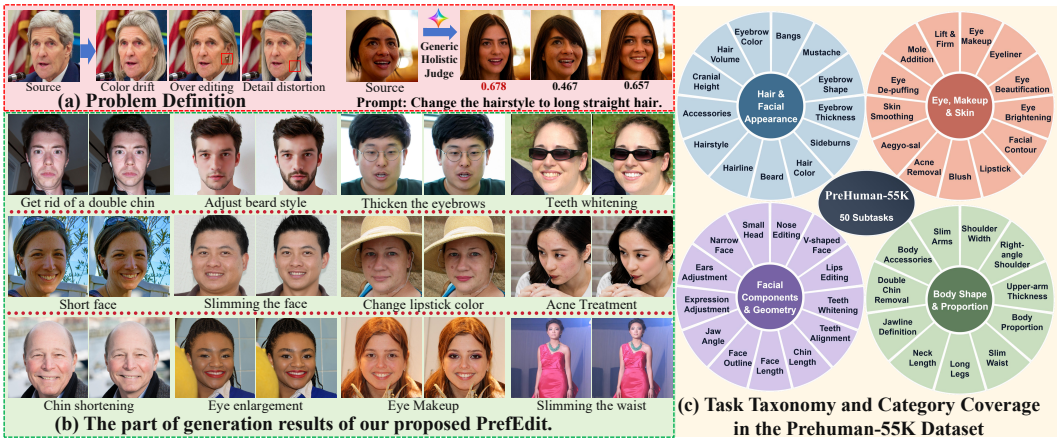


Figure 1: **Motivation and overview of our study.** (a) Two challenges of fine-grained human editing: existing editors struggle with precise local modification, causing color drift, over-editing, or detail distortion, while coarse rewards fail to reliably distinguish candidates with critical editing failures. (b) Representative results of PrefEdit across diverse facial, appearance, and body editing tasks. (c) The PreHuman-55K taxonomy, covering 50 subtasks organized into four categories.

ABSTRACT

Diffusion-based models perform well on broad semantic manipulation, but fine-grained human editing requires precise local changes while preserving non-target content. To address such limitations, we first construct **PreHuman-55K**, a high quality, large-scale dataset covering 50 anatomically localized editing subtasks (*e.g.,* eye makeup, hairline shape, double chin). We then present **PrefEdit**, which is specialized in fine-grained human image editing. Firstly, a lightweight LE-LoRA module stage learns preservation-sensitive local-editing priors. Based on such local editing ability, our core idea is to design a failure-aware rubric reward: Subtask-Specific Rubrics define task-dependent success and preservation criteria, yielding failure-aware rewards that reduce undesirable compensation from critical editing failures while preserving valid evidence from unaffected criteria. Extensive in-domain and out-of-domain experiments demonstrate that PrefEdit achieves state-of-the-art performance in instruction fulfillment, localization, preservation, and visual quality while substantially reducing critical failures.

1 INTRODUCTION

Human-centric image editing supports personalized content creation, portrait enhancement, and virtual-human applications. Despite recent progress in identity-aware and localized manipulation Li et al. (2024); Xian et al. (2025), many practical requests require subtle, anatomically specific changes, such as lowering the hairline or reducing the jaw angle. These edits impose an asymmetric objective: the target attribute must change clearly, while non-target regions remain stable. Existing models face a granularity gap between general-purpose editing supervision and the precise requirements of human-centric editing, which has two main sources, illustrated in Figure 1(a).

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.

The first challenge is the lack of fine-grained human-centric supervision. Existing editing models acquire instruction-following ability mainly through supervised fine-tuning (SFT) on paired data Brooks et al. (2023); Zhang et al. (2023); Sheynin et al. (2024), while recent datasets expand coverage to diverse open-ended operations Zhao et al. (2024); Yu et al. (2025); Zhou et al. (2025). However, they provide limited dense supervision for anatomically localized edits and their preservation constraints, causing coarse transformations, incorrect localization, and weak generalization to unseen applications. **The second challenge** is the limited granularity of existing preference feedback. Reward-based feedback Kirstain et al. (2023); Xu et al. (2023); Liang et al. (2024); Black et al. (2024) complements reference imitation with behavioral supervision, but typically emphasizes holistic preference, aesthetics, image-text alignment, or general quality. Even criterion-based evaluations Hu et al. (2023); Ma et al. (2024); Wang et al. (2025); Chen et al. (2025) provide limited information about which instruction detail failed, which anatomical region is incorrect, or which preservation constraint was violated. Consequently, visually plausible but instruction-inconsistent edits may still receive high rewards. We therefore argue that reliable local editing requires supervision, evaluation, and optimization at the same granularity as the requested edit.

To address this gap, we construct **PreHuman-55K**, comprising 50K paired samples and 5K instruction-only samples. As shown in Figure 1(c), it provides three complementary properties: (1) **Fine-grained taxonomy**, covering 50 anatomically localized appearance and geometry edits; (2) **High-quality curation**, using candidate edits from complementary systems on real portrait photographs and filtering them by subtask-specific instruction fulfillment, identity preservation, and visual quality; and (3) **Preference-optimization support**, where paired samples support fine-grained SFT and instruction-only prompts support preference post-training.

Building on PreHuman-55K, we propose **PrefEdit**, a two-stage Diffusion Transformer (DiT)-based framework for subtask-conditioned human-centric local editing (Figure 1(b)). Its central idea is to align post-training rewards with the granularity of the requested operation. LE-LoRA provides localized editing adaptation, SSR evaluates each subtask with explicit target-fulfillment, localization, preservation, and quality criteria, and FCRC applies failure-conditioned ceilings to affected criteria, limiting criterion-level compensation while preserving unrelated valid evidence.

Our contributions are summarized as follows:

- We formulate human-centric local editing as a subtask-conditioned optimization problem that jointly requires target execution and preservation of identity and non-target content.
- We present PrefEdit with a hierarchical, failure-aware reward design: SSR aggregates subtask-specific criteria under shared editing dimensions, while FCRC maps detected failure types and severities to affected criteria and caps their scores before aggregation.
- We construct PreHuman-55K, a 50-subtask benchmark for preservation-sensitive human editing, and show through experiments on both in-domain and out-of-domain splits that PrefEdit achieves stronger target execution, fewer critical failures, and better preservation of identity and non-target content.

2 Related Work

Diffusion Models (DMs) have significantly advanced since the pioneering DDPM Ho et al. (2020). Built upon latent diffusion models (LDMs) Rombach et al. (2022), diffusion models have been extended beyond image synthesis Zhang et al. (2025b); Qin et al. (2025) to diverse generative applications, including unified image generation Geng et al. (2024); Xiao et al. (2025), video generation Kara et al. (2025); Yao et al. (2025), 3D content generation Hu et al. (2025); Yang et al. (2025), and human motion synthesis Zhang et al. (2025a), which shows remarkable generated results. However, these advances mainly target general generation, macroscopic controllability. We therefore leveraged the powerful generative capabilities of diffusion models to explore fine-grained human-centric image editing remains insufficiently explored, where anatomically specific local changes must be performed while preserving identity and non-target regions.

Image Editing aims to transform a source image according to user-specified guidance, while preserving content irrelevant to the edit. Existing image editing methods mainly target broad semantic edits, such as general instruction following Brooks et al. (2023); Zhang et al. (2023), object insertion and region replacement Sheynin et al. (2024); Geng et al. (2024). Recent benchmark Pan et al. (2025); Xu et al. (2026); Qian et al. (2026) and methods Zhao et al. (2024); Yu et al. (2025);

Zhou et al. (2025) further expand the function of instruction type, but their primary objective is not systematically cover diverse anatomical operations, target-region localization, identity preservation, and non-target stability within a shared taxonomy. We therefore construct PreHuman-55K, which complements these general editing resources with 50 fine-grained human-editing subtasks, paired supervision, and instruction-only prompts for preference post-training.

Rubric-Guided Reward. Visual reward models commonly learn holistic preference Fan et al. (2023); Black et al. (2024); Wallace et al. (2024)scores for image quality or prompt alignment Kirstain et al. (2023); Wu et al. (2023); Xu et al. (2023), while recent evaluators introduce finer criteria for text-image faithfulness, compositional reasoning, and image editing quality Hu et al. (2023); Huang et al. (2023); Liang et al. (2024); Ma et al. (2024). These signals have been used for diffusion alignment and preference optimization. For example, Multi-Reward Gu et al. (2025) combines multidimensional MLLM scores with textual failure feedback, HP-Edit Li et al. (2026) uses a VLM scorer for editing post-training. However, they are still mostly generic and compensatory: a realistic image may receive a high reward even if the requested local human edit fails. PrefEdit instead constructing a hierarchical, subtask-conditioned reward rather than directly combining generic scores. SSR organizes each editing subtask from shared evaluation dimensions into subtask-specific atomic criteria and aggregates their scores bottom-up. FCRC further detects the type and severity of critical failures, maps them to affected criteria, and applies severity-dependent ceilings before aggregation. Failure signals therefore modify the reward structure directly, limiting criterion-level compensation while preserving valid evidence from unaffected criteria.

3 Dataset Construction

To narrow the granularity gap between general image-editing data and fine-grained human-centric editing, **PreHuman-55K** covers 50 fine-grained human-editing subtasks. In Figure 2,we first collect a diverse portrait pool spanning different identities, appearances, poses, clothing, lighting conditions, and compositions. Since editing systems have different strengths across subtasks, we systematically evaluate **GPT-Image-2, Nano Banana, Meitu Xiuxiu, and PixelCake** on the full task taxonomy. For each subtask, we select the system that best performs the requested local edit while preserving identity, non-target content, and visual quality. Each sample is constructed with a structured instruction that specifies the target change, the content to preserve, and the expected reconstruction quality.

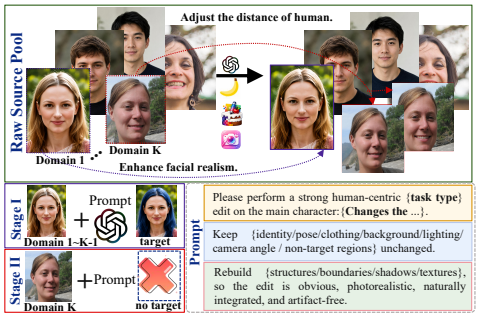


Figure 2: **Overview of the PreHuman-55K construction pipeline**. Curated multi-domain portraits are organized into 50K paired samples and 5K instruction-only samples using three-part editing prompts.

PreHuman-55K consists of 50,000 paired samples and 5,000 samples without fixed targets. Each paired sample contains a source image, an editing instruction, and a target image, and is used for supervised adaptation in Stage I. The remaining samples contain only a source image and an instruction, and are used for candidate rollout and preference post-training in Stage II. Most importantly, we construct **PreHuman-Benchmark** for evaluation. It contains 500 independently collected in-domain pairs and 1,000 out-of-domain pairs. More discussion can be seen in Appendix D.

4 Method

4.1 Overall Architecture

PrefEdit performs instruction-specified local changes while preserving identity and non-target content. In Figure 3, the framework contains two training stages and comprises: Local-Editing LoRA (LE-LoRA) module, Subtask-Specific Rubrics (SSR), and Failure-Conditioned Rubric Correction (FCRC). **In the first stage**, we encode the source image x, instruction t, and target image y from PreHuman-55K to obtain their respective latent representations c_S, c_T, c_p. Then, the concatenated tokens $Z_t = [c_S; c_T; c_p]$ are fed into the DiT block for a further step-wise denoising process. In particular, the c_s is regarded as noise-free condition tokens to preserve detailed information throughout all denoising stages. During this phase, LE-Lora module is inserted into each DiT block to realize

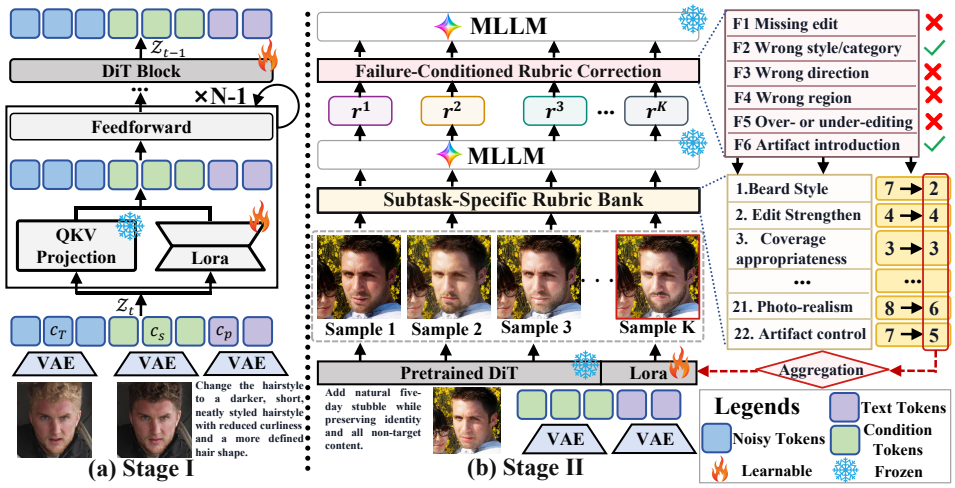


Figure 3: **Framework of PrefEdit.** PrefEdit follows a two-stage training scheme with the following components: The first stage trains LE-LoRA module to acquire fine-grained local-editing capability. In the second stage, the pretrained model in Stage I generates multiple candidates, SSR evaluates them using subtask-specific human judgments, and FCRC constrains failure-dependent criteria before the corrected scores are aggregated into rewards for preference post-training.

precise, localized edits while minimizing unintended modifications. This stage equips the model with preservation-sensitive capabilities for subtle human edits involving hair, beard, makeup, skin, and body attributes. **In the second stage**, the source image x and instruction t are regarded as the conditions of the above pretrained model to generate K candidate editing results $\{y^{(k)}\}_{k=1}^{K}$. The reward estimation is performed outside the editing model by a frozen multi-modal evaluator (*e.g.*,Gemini and Qwen). For each candidate $y^{(k)}$, the evaluator first uses the designed SSR corresponding to instruction t to produce raw $\mathcal{L}_3$ criterion scores. Then, the FCRC re-examines critical task-level failures and constrains the affected criteria to avoid default high-score judgments. After correction, the scores are aggregated in a bottom-up manner, from $\mathcal{L}_3$ criteria to four shared dimensions and finally to a scalar reward. For each source-instruction pair, the resulting reward vector $\mathbf{r} = [r^{(1)}, \ldots, r^{(K)}]$ provides candidate-level preference signals for post-training, encouraging the editor to generate results that correctly execute the requested local edit while preserving identity, non-target regions, and visual realism.

4.2 Local-Editing LoRA Module

Although FLUX-Kontext Black Forest Labs et al. (2025) provides a general editing prior, it lacks adaptation to fine-grained human-centric local edits. We therefore freeze its DiT backbone and insert LE-LoRA modules into each block, allowing the model to learn task-specific local transformation patterns, such as facial geometry, makeup, and hairline, within a dedicated adaptation space. In each DiT block, LE-LoRA module is implemented as a set of trainable low-rank residual adapters attached to the attention projections. Given the concatenated multi-modal token sequence c, the original attention projections are computed as $Q = cW_Q^\top, K = cW_K^\top, V = cW_V^\top$, where W_Q, W_K, and W_V are frozen pretrained projection matrices. To provide a dedicated adaptation space for local human editing, LE-LoRA augments each frozen projection with a trainable low-rank residual update: $\Delta W_Q = B_Q A_Q, \Delta W_K = B_K A_K, \Delta W_V = B_V A_V$, where only the adapter parameters $(A., B.)$ are optimized and each update satisfies $(\mathrm{rank}(\Delta W.) \leq r)$, with r=64. The adapted query, key, and value features are computed as $Q' = C(W_Q + \Delta W_Q)^\top, K' = C(W_K + \Delta W_K)^\top, V' = C(W_V + \Delta W_V)^\top$. The LE-LoRA reformulated Multi-modality Attention Esser et al. (2024) as $\mathrm{softmax}\left(\frac{Q'(K')^\top}{\sqrt{d}}\right)V'$. These low-rank residual adapters enable the attention layers to learn human-centric local transformation patterns, improving their sensitivity to subtle target-region edits.

4.3 Subtask-Specific Rubrics

Fine-grained human edits share a common evaluation structure comprising target fulfillment, localization, preservation, and visual quality. However, the visual evidence for these dimensions must be conditioned on the requested attribute and target region. For example, beard editing, hairline adjust-

ment, lip reshaping, and shoulder-width modification involve different local cues and preservation constraints under the same evaluation structure. A holistic reward or generic evaluation prompt cannot capture such edit-conditioned criteria and may therefore overlook whether the requested local transformation is correctly performed. To address this issue, we construct a subtask-specific rubric bank $\mathcal{B} = \{\mathcal{R}_s\}_{s=1}^{S}$, where each rubric $\mathcal{R}_s$ provides dedicated evaluation criteria for editing instructions associated with subtask s. Each rubric follows a shared hierarchical structure but adopts subtask-specific criterion semantics. Specifically, every $\mathcal{R}_s$ contains 22 $\mathcal{L}_3$ criteria organized under four shared dimensions: **Prompt Compliance**, **Localization and Anatomical Integration**, **Subject Preservation**, and **Task Realism and Image Quality**. The shared hierarchy provides a unified evaluation interface, while the $\mathcal{L}_3$ descriptions are customized to the visual requirements of each subtask. Given a source image x, an editing instruction t, and the k-th edited candidate $y^{(k)}$, the frozen multi-modal evaluator J_ϕ produces structured criterion-level judgments according to the corresponding subtask rubric $\mathcal{R}_s$: $\tilde{\mathbf{s}}^{(k)} = J_\phi\big(x, t, y^{(k)}; \mathcal{R}_s\big), \tilde{\mathbf{s}}^{(k)} \in \{0, \ldots, 9\}^{22}$. Here, $\tilde{\mathbf{s}}^{(k)}$ denotes the raw $\mathcal{L}_3$ score vector over the 22 rubric criterion. These scores do not directly serve as the final reward. Instead, they provide interpretable fine-grained evidence for subsequent FCRC and then bottom-up reward aggregation.

4.4 Failure-Conditioned Rubric Correction

Although SSR provides fine-grained scores, independently evaluated $\mathcal{L}_3$ criteria may remain compensation effect. For example, a realistic full beard may score highly in texture and preservation despite violating a five-day-stubble instruction. Such critical failures should therefore constrain all logically dependent criteria. To prevent critical editing failures from being obscured by high scores on unrelated criteria, we introduce Failure-Conditioned Rubric Correction (FCRC).

For each candidate $y^{(k)}$, the evaluator J_ϕ detects failures according to the corresponding rubric $\mathcal{R}_s$: $\mathcal{F}_s^{(k)} = J_\phi^F(x, t, y^{(k)}; \mathcal{R}_s)$, where ϕ denotes the frozen evaluator parameters, and the superscript F indicates failure-detection prompting. The output $\mathcal{F}_s^{(k)}$ is a set of detected failures for candidate k. Each element $(f, \nu) \in \mathcal{F}_s^{(k)}$ contains a failure type f defined by rubric $\mathcal{R}_s$ and its severity $\nu \in \{1, \ldots, 5\}$. The failure types include missing, reversed, or substituted edits, misplaced target regions, excessive modifications, and preservation violations. FCRC associates each failure type f with an affected $\mathcal{L}_3$ criterion set $\mathcal{A}_s(f)$. Let $\tilde{s}_j^{(k)}$ denote the raw SSR score for criterion j, and let $c_{s,j}(f, \nu)$ denote its maximum permitted score under failure f with severity ν. The corrected $\mathcal{L}_3$ score can be rewritten as: $s_j^{(k)} = \min\left(\tilde{s}_j^{(k)}, \min_{\substack{(f,\nu)\in\mathcal{F}_s^{(k)} \\ j\in\mathcal{A}_s(f)}} c_{s,j}(f, \nu)\right)$. The inner minimum is omitted when no detected failure affects criterion j, in which case $s_j^{(k)} = \tilde{s}_j^{(k)}$. If multiple failures affect the same criterion, the most restrictive ceiling is applied. FCRC thus corrects only failure-dependent criteria while preserving unrelated evidence.

After correction, the $\mathcal{L}_3$ scores are normalized and aggregated bottom-up. Let $\mathcal{C}_{s,j}$ denote the $\mathcal{L}_3$ criteria belonging to the m-th shared dimension and $w_{s,mj}$ their normalized weights, which can be expressed as: $q_m^{(k)} = \sum_{j\in\mathcal{C}_{s,m}} w_{s,mj} \frac{s_j^{(k)}}{9}$, where $\sum_{j\in\mathcal{C}_{s,m}} w_{s,mj} = 1$. The four dimensions are Prompt Compliance, Localization and Anatomical Integration, Subject Preservation, and Task Realism and Image Quality. Their scores are aggregated into the final reward, which can be written as: $r^{(k)} = \sum_{m=1}^{4} \alpha_{s,m} q_m^{(k)}$, where $\sum_{m=1}^{4} \alpha_{s,m} = 1$, and $\alpha_{s,m} \geq 0$ controls the importance of dimension m for subtask s. Prompt Compliance receives the largest weight, so a realistic or well-preserved result should not receive a high reward when the requested edit is missing, incorrect, reversed, or substituted. FCRC therefore produces a non-compensatory reward that penalizes critical editing failures without suppressing unrelated valid evidence.

5 Experiments

Implementation Details. We conduct all experiments using 8 NVIDIA H200 GPUs. The original input images have varying resolutions, including 1024×1024 and 512×1024, and are uniformly resampled to 512×512 before training. In the first stage, we initialize the model from the pretrained FLUX.1 Kontext Black Forest Labs et al. (2025) and perform supervised adaptation using LoRA

with a rank of 64, a batch size of 1 per device, and a learning rate of 1×10^{-5}. In the second stage, we perform rubric-guided diffusion post-training based on Diffusion-NFT Zheng et al. (2025) with 16 trajectories per prompt, a batch size of 4, and 800 optimization steps, while retaining the same learning rate. More details can be seen in Appendix. A.

Benchmark. We evaluate on PreHuman-Bench, comprising 500 held-out in-domain (ID) pairs from PreHuman-55K and 1,000 out-of-domain (OOD) pairs collected from external photographic sources with different styles and resolutions. Both splits cover 50 subtasks, are disjoint from training, and are evaluated separately using macro-averaged metrics. More details can be seen in Appendix. D.

Metrics. We report rubric-based metrics include mean Edit Success Rate (mESR), mean Rubric Score (mRS), Target Fulfillment (TF), Subject Preservation (SP), Visual Quality (VQ), and Critical Failure Rate (CFR). Reward-independent metrics include Face Identity Similarity (Face-ID) Deng et al. (2019), DINOv2-based Preservation Similarity (DINO-Pres) Oquab et al. (2024), and CLIP Directional Similarity (CLIP-Dir) Radford et al. (2021). More Details can be seen in Appendix. C.

Baseline Methods. Publicly Research: FLUX.1 Kontext Black Forest Labs et al. (2025), FireRed-Image-Edit Super Intelligence Team et al. (2026), Qwen-Image-Edit-2509 Qwen Team (2025), FLUX.2 [klein] 9B Black Forest Labs (2026), and LongCat-Image-Edit Meituan LongCat Team et al. (2025). Commercial API: GPT Image 2 OpenAI (2026) and Nano Banana Google (2026).

5.1 Main Results

Figure 4 demonstrates the generated results of PrefEdit, which effectively conducts prompt while preserving the semantics and fine details. More generated results can be seen in Appendix E.1.

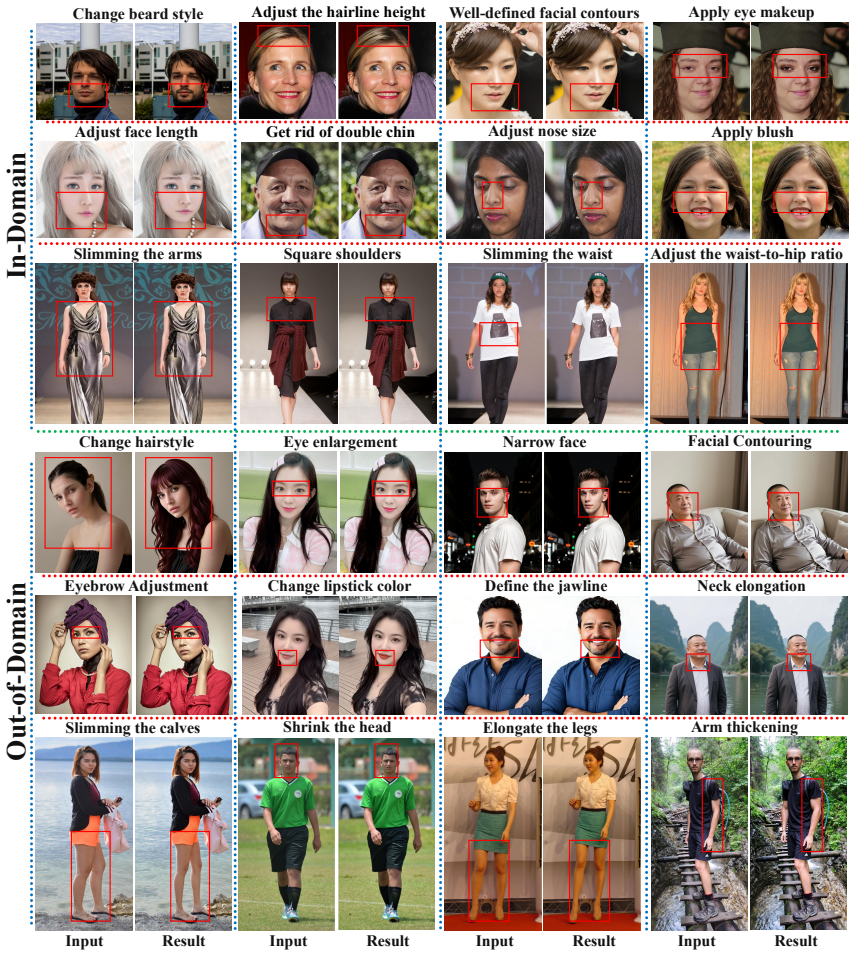


Figure 4: **The generated results of PrefEdit**.

Quantitative Evaluation. As shown in Table 1, on the in-domain split, PrefEdit outperforms the per-metric runner-up by 3.4, 2.5, and 5.4 points in mESR, mRS, and TF, respectively, while reducing CFR from 35.8 to 9.0 (a 74.9% relative reduction) and achieving the best SP. Relative to the best

commercial result for each metric, it improves mESR, mRS, TF, and Face-ID by 17.4%, 3.7%, 14.0%, and 6.2%, respectively, while reducing CFR by 76.7%. On the OOD split, PrefEdit improves over FLUX.2 [klein] by 0.5, 2.1, 2.6, 3.5, and 9.1 points in mESR, mRS, SP, VQ, and Face-ID, respectively, although FLUX.2 achieves a higher CLIP-Dir score (5.0 vs. 4.3) and Nano Banana obtains a higher TF score (58.4 vs. 53.9). Since TF and CLIP-Dir mainly measure visible semantic alignment, these gains may partly reflect stronger image changes that do not fully account for non-target modifications. PrefEdit instead prioritizes localized editing and portrait preservation. This trend is also observed against commercial APIs: compared with GPT-Image-2, it improves mESR, mRS, SP, VQ, and Face-ID by 14.3, 6.0, 4.0, 5.6, and 5.7 points, respectively, while reducing CFR by 5.1 points. compared with Nano Banana, it improves SP and Face-ID by 17.9 and 6.3 points, respectively, and VQ by 11.4 points, indicating gains in both preservation and visual quality. Nano Banana nevertheless has higher TF and lower CFR, suggesting a more aggressive target-execution strategy. PrefEdit therefore adopts a preservation-oriented operating point, accepting lower nominal target-execution when necessary to keep identity and non-target regions stable.

Table 1: **Comparison on PreHuman-Bench.** Bold and underline denote the best and second results.

In-domain		**Rubric-aligned evaluation**						**Reward-free evaluation**		
Model type	**Method**	mESR↑	mRS↑	TF↑	SP↑	VQ↑	CFR↓	Face-ID↑	DINO-Pres↑	CLIP-Dir↑
Research models	FLUX.1-Kontext-dev	15.4	62.5	57.2	67.6	65.1	<u>35.8</u>	**90.8**	96.9	3.77
	FireRed-Image-Edit	<u>22.2</u>	65.1	<u>59.1</u>	68.5	68.2	38.4	81.4	96.3	3.05
	Qwen-Image-Edit	11.6	60.9	56.6	61.6	65.8	42.8	81.9	92.6	**7.16**
	FLUX.2 [klein] 9B	11.4	59.5	54.9	61.3	63.9	47.0	86.2	97.2	4.49
	LongCat-Image-Edit	8.2	58.9	54.4	59.2	64.6	47.4	72.3	92.1	<u>4.88</u>
Commercial APIs	GPT-Image-2	21.8	65.3	56.6	68.5	70.6	38.6	81.2	97.5	1.91
	Nano Banana	19.0	<u>66.7</u>	53.0	<u>73.9</u>	**72.3**	38.8	84.2	**98.5**	3.58
Ours	**PrefEdit**	**25.6**	**69.2**	**64.5**	**74.0**	<u>70.9</u>	**9.0**	<u>89.4</u>	<u>98.0</u>	3.54
Out-of-domain		**Rubric-aligned evaluation**						**Reward-free evaluation**		
Model type	**Method**	mESR↑	mRS↑	TF↑	SP↑	VQ↑	CFR↓	Face-ID↑	DINO-Pres↑	CLIP-Dir↑
Research models	FLUX.1-Kontext-dev	30.2	59.0	47.1	71.5	67.4	52.2	<u>93.0</u>	97.2	3.4
	FireRed-Image-Edit	31.1	59.7	48.3	73.9	66.6	49.9	82.0	90.7	4.5
	Qwen-Image-Edit	38.3	52.6	49.2	59.1	54.2	56.4	88.5	<u>97.5</u>	4.5
	FLUX.2 [klein] 9B	<u>46.5</u>	<u>67.9</u>	52.9	<u>84.0</u>	<u>77.6</u>	<u>45.0</u>	85.7	96.2	**5.0**
	LongCat-Image-Edit	35.7	64.3	50.1	78.6	73.9	46.2	75.2	89.7	<u>4.7</u>
Commercial APIs	GPT-Image-2	32.7	64.0	45.6	82.6	75.5	52.5	89.1	94.8	2.3
	Nano Banana	41.3	64.8	**58.4**	68.7	69.7	**44.3**	88.5	<u>97.5</u>	4.5
Ours	**PrefEdit**	**47.0**	**70.0**	<u>53.9</u>	**86.6**	**81.1**	47.4	**94.8**	**97.6**	4.3

Qualitative Comparison. Figure 5 explains the differences in Table 1: TF and CLIP-Dir mainly measure visible semantic alignment, whereas SP, Face-ID, and DINO-Pres. evaluate identity and non-target preservation, and VQ assesses overall visual quality. For In-domain, PrefEdit localizes edits such as eye makeup, blush, and waist shaping while preserving facial structure, hairstyle, clothing, pose, and background. For example, its eye-makeup and blush changes remain concentrated around the face, whereas several baselines introduce broader color or appearance changes. The same trend is clearer OOD: for the hairstyle prompt, Nano Banana produces a stronger and broader change in hair silhouette and facial framing, while PrefEdit changes the hairstyle but retains the subject's facial structure and expression. In the double-chin example, PrefEdit restricts the change to the chin region, whereas some baselines also reshape the surrounding face; in the slim-arm example, it modifies the arms while better preserving the torso and clothing. These stronger baseline transformations explain why Nano Banana's higher TF and CLIP-Dir scores, but may also reduce preservation or visual quality, as reflected by SP, VQ, and Face-ID. PrefEdit instead reduces edit magnitude when broader changes would destabilize the portrait, achieving a better overall balance as reflected by its higher mESR and mRS. More results are provided in Appendix E.

Human Validation of the Reward Evaluators. We evaluate 100 held-out candidate pairs (50 ID and 50 OOD, with one pair per split for each of 50 subtasks), yielding 200 outputs from training-disjoint sources. These cases were not used to define or tune the evaluators. Ten experts independently score each output on TF, LI, SP, and VQ and select the preferred result within each pair. For Generic, SSR, and SSR+FCRC, Pearson's r is computed per dimension against mean expert ratings, while PA is the

Table 2: Alignment of automatic rubric evaluators with human experts.

Evaluator	TF	LI	SP	VQ	PA (%)
Generic	0.68	0.62	0.77	0.76	74
SSR	0.76	0.69	0.83	**0.78**	80
SSR+FCRC	**0.78**	**0.72**	**0.84**	**0.78**	**84**
Expert ICC	0.94	0.91	0.96	0.93	–

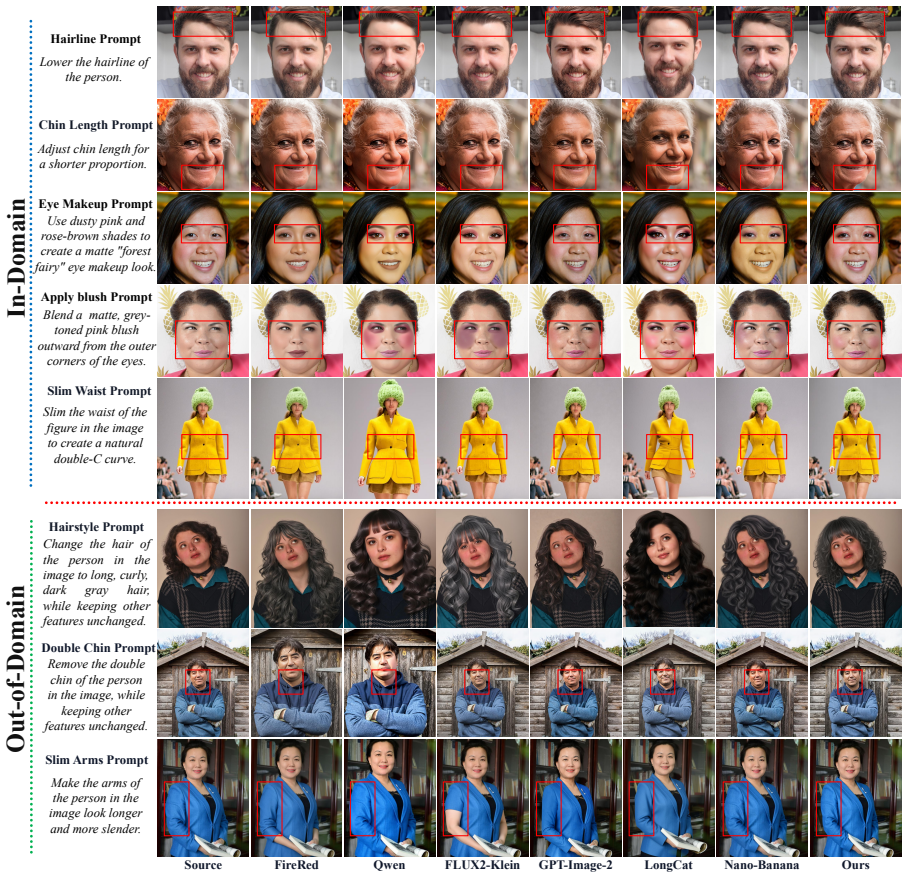


Figure 5: **Qualitative comparisons** between PrefEdit and other SOTA methods.

percentage of the 100 pairs matching the experts' majority preference. Automatic ties use the evaluator's four-dimensional mean, and 5:5 expert ties use the corresponding mean expert rating. Table 2 reports these results and inter-expert ICC. SSR+FCRC achieves the highest observed PA and highest or tied-highest correlations, suggesting closer alignment with expert preferences on this validation.

6 Ablation Study

Table 3 demonstrates the effectiveness of each component. The Baseline denotes the frozen FLUX.1-Kontext-dev editor. The w/ LE-LoRA configuration adds the Stage-I localized editing adaptation, while w/ SSR further introduces Stage-II SSR optimization. The w/(SSR+FCRC) configuration additionally applies FCRC and corresponds to the full PrefEdit model.

Table 3: Ablation study on the in-domain and out-of-domain splits of PreHuman-Bench.

	In-domain									Out-of-domain								
	Rubric-aligned evaluation						Reward-free evaluation			Rubric-aligned evaluation						Reward-free evaluation		
Method	mESR↑	mRS↑	TF↑	SP↑	VQ↑	CFR↓	Face-ID↑	DINO-Pres.↑	CLIP-Dir↑	mESR↑	mRS↑	TF↑	SP↑	VQ↑	CFR↓	Face-ID↑	DINO-Pres.↑	CLIP-Dir↑
Baseline	15.4	62.5	57.2	67.6	65.1	35.8	90.8	96.9	3.77	30.2	59.0	47.1	71.5	67.4	52.2	93.0	97.2	3.4
w/ LE-LoRA	25.1	68.4	63.8	73.2	65.6	16.8	87.9	96.4	3.89	35.1	59.9	41.7	79.3	69.3	61.0	91.1	97.2	3.6
w/ SSR	25.3	68.9	64.2	73.6	68.3	13.2	88.6	97.3	3.78	43.8	67.4	52.1	84.9	74.7	52.7	92.0	97.3	4.5
w/(SSR+FCRC)	25.6	69.2	64.5	74.0	70.9	9.0	89.4	98.0	3.54	47.0	70.0	53.9	86.6	81.1	47.4	94.8	97.6	4.3

Effect of LE-LoRA module. As shown in Table 3, on the ID split, LE-LoRA improves ID performance: mESR/mRS/TF/SP/VQ increase by 9.7/5.9/6.6/5.6/0.5 points and CFR decreases from 35.8 to 16.8, but Face-ID and DINO-Pres. decline from 90.8/96.9 to 87.9/96.4 despite a CLIP-Dir. increase from 3.77 to 3.89. OOD gains remain for mESR/mRS/SP/VQ/CLIP-Dir. (4.9/0.9/7.8/1.9/0.2 points), while TF drops by 5.4 points, CFR rises from 52.2 to 61.0, and Face-ID falls from 93.0 to 91.1, indicating limited preservation under distribution shift. Figure 6 provides a visual explanation of this behavior. In the face-contour example, LE-LoRA removes the baseline's large gray artifact and preserves the surrounding face, although the contour edit remains weak. In the eye-makeup example, it confines changes to the eyes and introduces the requested orange-brown

tones and highlights, but does not fully reproduce the color combination and inner eyeliner; for the hairstyle prompt, it approaches the requested shoulder-length straight style while preserving the subject, clothing, and background, although the hair length and edge structure remain imperfect. Overall, LE-LoRA effectively injects fine-grained local-editing priors and improves editing locality.

Effect of SSR. In Table 3, SSR improves the ID mESR/mRS/TF/SP/VQ scores by 0.2/0.5/0.4/0.4/2.7 points and reduces CFR by 3.6 points, while partially recovering preservation: Face-ID and DINO-Pres. increase from 87.9/96.4 to 88.6/97.3, with only a slight CLIP-Dir. decrease from 3.89 to 3.78. The OOD gains are larger, with mESR/mRS/TF/SP/VQ improving by 8.7/7.5/10.4/5.6/5.4 points and CFR decreasing by 8.3 points. Face-ID, DINO-Pres., and CLIP-Dir. further rise from 91.1/97.2/3.6 to 92.0/97.3/4.5. These results indicate that SSR provides more robust task-specific guidance for balancing target fulfillment and non-target preservation. Figure 6 provides visual support: relative to the weak or nearly unchanged LE-LoRA outputs, $w/$ SSR produces clearer local edits without the baseline's gray artifact, strengthens the face-contour modification, renders the requested eye-makeup cues more consistently, and brings the hairstyle closer to the specified shoulder-length straight style while preserving identity, clothing, and background. Thus, SSR guides edits according to the task rather than uniformly increasing their magnitude. Figure 7 further shows that generic Gemini scores occupy a narrow range of $[0.593, 0.670]$ and rank Sample 2 highest despite its weak satisfaction of the hairstyle and eye-covering requirements, indicating a bias toward overall plausibility. SSR instead expands the range to $[0.553, 0.879]$ and assigns higher scores to Samples 1 and 5, providing more discriminative supervision for instruction fulfillment and portrait preservation.

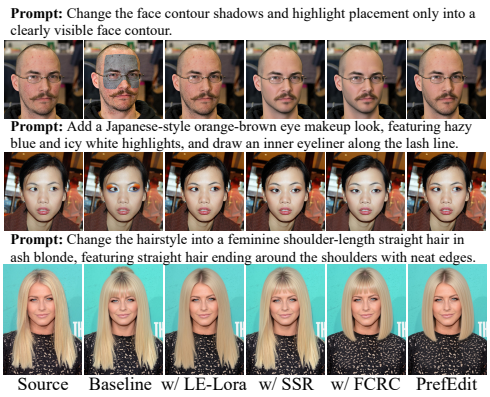


Figure 6: Ablation shows that full settings ensure strong editing consistency.

Effect of FCRC. FCRC applies failure-conditioned ceilings to criteria affected by detected critical errors, limiting compensation from unrelated high scores. Relative to SSR in Table 3, it improves ID/OOD mESR, mRS, TF, SP, and VQ by $0.3/3.2$, $0.3/2.6$, $0.3/1.8$, $0.4/1.7$, and $2.6/6.4$ points, respectively; reduces CFR by $4.2/5.3$ points; increases Face-ID by $0.8/2.8$ and DINO-Pres. by $0.7/0.3$ points; and decreases CLIP-Dir. by $0.24/0.2$ points. Figure 6 shows that the full model reduces the facial-structure distortion, excessive eye-makeup changes, and unintended hairstyle over-editing observed in some ablated variants. Figure 7 shows that SSR can assign a high score to a candidate that satisfies salient visual cues but still contains a task-critical mismatch: Sample 5 is ranked above Sample 1 by SSR (0.879 vs. 0.863), although its hairstyle is less consistent with the requested long and straight structure. FCRC selectively corrects this ranking by lowering Sample 5 to 0.684 while changing Sample 1 by only 0.007, whereas Sample 7 remains unchanged at 0.564. This example illustrates that FCRC limits criterion-level compensation and aligns the reward more closely with localized instruction fulfillment, without imposing a uniform penalty on all candidates.

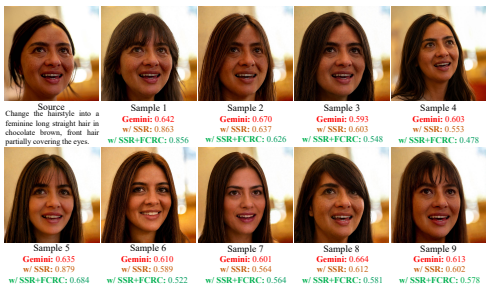


Figure 7: Reward comparison shows that full settings or with SSR more reliably rank edits than generic rewards.

7 Conclusion

In this paper, we first construct **PreHuman-55K**, spanning 50 localized human-editing subtasks. Building on this dataset, we propose **PrefEdit**, a two-stage framework in which lightweight LE-LoRA adapts the DiT backbone to preservation-sensitive local edits. More importantly, SSR provides edit-conditioned evaluation criteria, while FCRC prevents critical failures from being masked by compensatory scores, yielding reliable failure-aware rubric rewards. Experiments on both ID and OOD splits demonstrate SOTA overall performance, with stronger human-preference alignment, target fulfillment, localization, visual quality, and stable identity and non-target preservation.

AI USE STATEMENT

Generative AI tools assisted manuscript writing, translation, document formatting, literature search and organization, and feedback on the clarity and consistency of the method and experimental protocol. They were not used to implement the reported model, preprocess datasets, train models, analyze raw experimental data, or generate reported results. The authors manually checked AI-assisted text against the cited literature, the documented method, and the reported tables, and take responsibility for the final manuscript, technical claims, and results.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

Black Forest Labs. FLUX.2 [klein] 9B. Hugging Face model card: `https://huggingface.co/black-forest-labs/FLUX.2-klein-9B`, 2026. Accessed: 2026-07-29.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Sherry X. Chen, Yi Wei, Luowei Zhou, and Suren Kumar. Adiee: Automatic dataset creation and scorer for instruction-guided image editing evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18345–18356, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

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, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12709–12720. IEEE, 2024.

Google. Gemini 2.5 Flash Image (Nano Banana). `https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-image`, 2026. Accessed: 2026-07-29.

Xin Gu, Ming Li, Libo Zhang, Fan Chen, Longyin Wen, Tiejian Luo, and Sijie Zhu. Multi-reward as condition for instruction-based image editing. In *International Conference on Learning Representations*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Hanzhe Hu, Tianwei Yin, Fujun Luan, Yiwei Hu, Hao Tan, Zexiang Xu, Sai Bi, Shubham Tulsiani, and Kai Zhang. Turbo3d: Ultra-fast text-to-3d generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23668–23678. IEEE, 2025.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78723–78747, 2023.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28405–28415. IEEE, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Fan Li, Chonghuinan Wang, Lina Lei, Yuping Qiu, Jiaqi Xu, Jiaxiu Jiang, Xinran Qin, Zhikai Chen, Fenglong Song, Zhixin Wang, Renjing Pei, and Wangmeng Zuo. Hp-edit: A human-preference post-training framework for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Nannan Li, Qing Liu, Krishna Kumar Singh, Yilin Wang, Jianming Zhang, Bryan A. Plummer, and Zhe Lin. Unihuman: A unified model for editing human images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2039–2048, 2024.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Alexander Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, and Feng Yang. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2EBench: A comprehensive benchmark for instruction-based image editing. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, et al. LongCat-Image technical report, 2025.

OpenAI. GPT Image 2 Model. `https://developers.openai.com/api/docs/models/gpt-image-2`, 2026. Accessed: 2026-07-29.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yulin Pan, Xiangteng He, Chaojie Mao, Zhen Han, Zeyinzi Jiang, Jingfeng Zhang, and Yu Liu. Ice-bench: A unified and comprehensive benchmark for image creating and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16586–16596. IEEE, 2025.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-banana-400k: A large-scale dataset for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, et al. Lumina-image 2.0: A unified and efficient image generative framework. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20031–20042. IEEE, 2025.

Qwen Team. Qwen-Image-Edit-2509. Hugging Face model card: `https://huggingface.co/Qwen/Qwen-Image-Edit-2509`, 2025. Accessed: 2026-07-29.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

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.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Super Intelligence Team, Changhao Qiao, Chao Hui, et al. FireRed-Image-Edit-1.0 technical report, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Qian Wang, Aleksandar Cvejić, Abdelrahman Eldesokey, and Peter Wonka. Editclip: Representation learning for image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15960–15970, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, 2023.

Xiaole Xian, Xilin He, Zenghao Niu, Junliang Zhang, Weicheng Xie, Siyang Song, Zitong Yu, and Linlin Shen. Ca-edit: Causality-aware condition adapter for high-fidelity local facial attribute editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8593–8601, 2025. doi: 10.1609/aaai.v39i8.32928.

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. In *Advances in Neural Information Processing Systems*, volume 36, 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.

Yuanbo Yang, Jiahao Shao, Xinyang Li, Yujun Shen, Andreas Geiger, and Yiyi Liao. Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2857–2869. IEEE, 2025.

Shuaizhen Yao, Xiaoya Zhang, Xin Liu, Mengyi Liu, and Zhen Cui. Stdd: Spatio-temporal dual diffusion for video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12575–12584. IEEE, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jianrong Zhang, Hehe Fan, and Yi Yang. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17592–17602. IEEE, 2025a.

Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23464–23473. IEEE, 2025b.

Kai Zhang, Lingzhi Li, Jingjing Wang, Jing Zhang, Yang Liu, Jian Sun, et al. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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, 2025.

Jun Zhou, Jiahao Li, Zunnan Xu, Hanhui Li, Yiji Cheng, Fa-Ting Hong, Qin Lin, Qinglin Lu, and Xiaodan Liang. Fireedit: Fine-grained instruction-based image editing via region-aware vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.