EditHF-1M: A Million-Scale Rich Human Preference Feedback for Image Editing

Zitong Xu[1], Huiyu Duan[1*], Zhongpeng Ji[2], Xinyun Zhang[3], Yutao Liu[4], Xiongkuo Min[1], Ke Gu[5], Jian Zhang[2], Shusong Xu[2], Jinwei Chen[2], Bo Li[2*], and Guangtao Zhai[1*]
*Corresponding Authors

[1] Shanghai Jiao Tong University, Shanghai, China
{xuzitong, huiyuduan}@sjtu.edu.cn
[2] vivo BlueImage Lab, vivo Mobile Communication Co., Ltd, Shanghai, China
[3] University of Electronic and Science Technology of China, Sichuan, China
[4] Ocean University of China, Shandong, China
[5] Beijing University of Technology, Beijing, China

Abstract. Recent text-guided image editing (TIE) models have achieved remarkable progress, while many edited images still suffer from issues such as artifacts, unexpected editings, unaesthetic contents. Although some benchmarks and methods have been proposed for evaluating edited images, scalable evaluation models are still lacking, which limits the development of human feedback reward models for image editing. To address the challenges, we first introduce **EditHF-1M**, a million-scale image editing dataset with over 29M human preference pairs and 148K human mean opinion ratings, both evaluated from three dimensions, *i.e.*, visual quality, instruction alignment, and attribute preservation. Based on EditHF-1M, we propose **EditHF**, a multimodal large language model (MLLM) based evaluation model, to provide human-aligned feedback from image editing. Finally, we introduce **EditHF-Reward**, which utilizes EditHF as a reward signal to optimize the text-guided image editing models through reinforcement learning. Extensive experiments show that EditHF achieves superior alignment with human preferences and demonstrates strong generalization on other datasets. Furthermore, we fine-tune the Qwen-Image-Edit using EditHF-Reward, achieving significant performance improvements, which demonstrates the ability of EditHF to serve as a reward model to scale-up the image editing. Both the dataset and code will be released in our GitHub repository: https://github.com/IntMeGroup/EditHF.

1 Introduction

The rapid advancement of image editing allows fine-grained image modifications through natural language instructions [3,7,48,60,72,89]. Although some state-of-the-art editing models, such as Nano-Banana [7], SeedDream [60], Qwen-Image-Edit [72], *etc.*, have achieved impressive performance, many edited images still suffer from issues such as artifacts, unexpected editings, unaesthetic contents.

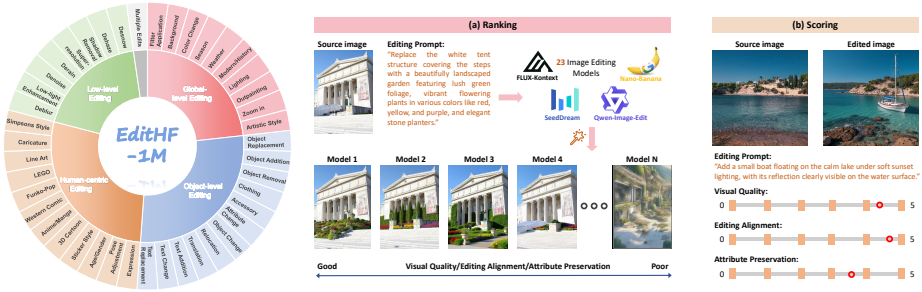

Fig. 1: Overview of EditHF-1M and human annotations: (a) Ranking: edited images from different editing models are grouped by source image and editing prompt and ranked to assess relative quality; (b) Scoring: each edited image is rated individually for evaluating absolute quality.

Table 1: Comparison of text-guided image editing model evaluation benchmarks. ✓ means publicly available, ✗ means not applicable or not available.

Databases	Source Images	Edited Images	Annotations	Editing Models	Editing Tasks	Prompt Type	Dimensions	Evaluation Model	Validated for Refinement
TedBench [24]	100	400	300 Pairs	3	4	Description	Overall	✗	✗
EditVal [2]	648	5,184	10.3K Scores	8	12	Instruction	Alignment, Preservation	✗	✗
EditEval [18]	150	1,050	4.2K Scores	7	19	Description, Instruction	Quality, Alignment, Preservation, Realism	✓	✗
AURORA-Bench [27]	400	2,000	2K Scores	5	8	Instruction	Overall	✓	✗
ImagenHub [28]	179	1,432	2.8K Scores	8	-	Description, Instruction	Quality, Preservation	✗	✗
IE-Bench [64]	301	1,505	4.5K Scores	5	11	Description, Instruction	Quality, Correspondence, Preservation	✓	✗
EBench-18K [80]	1,080	18,360	54K Scores	17	21	Description, Instruction	Quality, Alignment, Preservation	✓	✗
EditScore [45]	577	2,885	3,072 Pairs	11	13	Instruction	Quality, Alignment, Preservation	✓	✓
EditReward [75]	9,557	114.7K	200K Pairs	6	7	Instruction	Quality, Alignment	✓	✓
EditHF-1M	**44.3K**	**1.01M**	**29.1M Pairs, 148K Scores**	**23**	**43**	**Description, Instruction**	**Quality, Alignment, Preservation**	✓	✓

Thus, several datasets and benchmarks have emerged for evaluating text-guided image editing (TIE) models [2,18,45,47,64,75,80]. However, existing benchmarks exhibit several critical limitations as shown in Table 1. **(i) Limited content diversity:** Most existing benchmarks cover only a narrow set of editing tasks and rely on a small number of source images [18, 45, 64], which restricts their ability to evaluate flexible and general-purpose image editing models across diverse scenarios. **(ii) Limited coverage of editing models:** The majority of current datasets include an insufficient number of editing methods [64, 75] and often exclude recent state-of-the-art or closed-source models [2,18,47], leading to incomplete and potentially biased evaluations that fail to reflect practical model performance. **(iii) Limited effectiveness for model refinement:** Most existing datasets and benchmarks primarily serve as static evaluation tools and rarely examine or demonstrate the effectiveness for refining image editing models [18,64,80].

Besides datasets and benchmarks, some metrics have been developed or applied for evaluating image editing, including image quality assessment (IQA) metrics [8,50,52,61,62,70,71,90,91], vision-language approaches [16,34,78], and multimodal large language model (MLLM) based evaluation methods [18, 47]. Traditional IQA metrics primarily focus on natural distortions and low-level visual fidelity, but fail to capture semantic alignment with user instructions. Vision-language approaches such as CLIPScore [16] and HPSv2 [76] have shown notable progress in evaluating image generation by incorporating human visual

feedback [16,34,37,76,78], however, they mainly assess text-image alignment and are not designed to capture the semantic changes specific to image editing. More recently, several studies have explored leveraging MLLMs to evaluate image editing [18, 45, 47, 75, 80]. Nevertheless, zero-shot MLLM-based evaluations [18, 47] often exhibit limited alignment with human preferences in complex editing scenarios, while fine-tuned MLLMs [45, 75, 80] tend to suffer from restricted generalization when encountering unseen editing models and editing tasks. Moreover, previous studies have not considered training models by combining score regression and pairwise comparison to achieve both effective evaluation and reward capabilities.

In this work, we introduce **EditHF-1M**, a million-scale image editing evaluation dataset curated by trained annotators under a rigorous and standardized annotation protocol. EditHF-1M consists of **44.3K** source images collected from Pico-Banana-400K [57], EBench-18K [80], and free photography websites, each paired with an editing prompt, covering **43 editing tasks**, as shown in Figure 1. Based on these source images and prompts, we generate over **1M** edited images using **23** state-of-the-art image editing models, including both open-sourced models such as Qwen-Image-Edit [72] and OmniGen2 [73], and closed-source models such as NanoBanana [7] and Seedream4 [60]. Through an extensive subjective study, we collect over **29M** human preference pairs and **148K** human mean opinion ratings, covering perceptual quality, editing alignment, and attribute preservation dimensions. Building upon EditHF-1M, we propose **EditHF**, an MLLM-based evaluation model for image editing that assesses from multiple dimensions. Extensive experiments conducted on EditHF-1M and other benchmarks demonstrate that EditHF achieves state-of-the-art performance and exhibits strong generalization ability. Furthermore, we introduce **EditHF-Reward**, which demonstrates that EditHF can be effectively utilized as a reward model for reinforcement learning to enhance image editing models by optimizing Qwen-Image-Edit [72].

The main contributions of this work include:

- We introduce EditHF-1M, a million scale dataset containing over 1M edited images across diverse tasks with comprehensive human preference rankings and scores covering perceptual quality, editing alignment and attribute preservation dimensions.
- We propose EditHF, a unified model by jointly training on score prediction and pairwise comparison, which achieves human-aligned evaluation from multiple dimensions for image editing.
- We introduce EditHF-Reward, which utilizes EditHF as the reward signal for reinforcement learning to enhance the ability of image editing models.
- Extensive experiments on EditHF-1M and other image editing benchmarks demonstrate that EditHF achieves state-of-the-art performance and generalizes well. Furthermore, EditHF-Reward can effectively enhance the performance of image editing models with EditHF.

2 Related Work

2.1 Image Editing

With the development of generative models such as Stable Diffusion [58] and FLUX [11], numerous image editing methods have been proposed [18]. Based on the type of editing prompts, these methods can be divided into description-based approaches (*e.g.*, "A deer running on the grass" → "A horse running on the grass") and instruction-based approaches (*e.g.*, "Change the deer to a horse") [18, 80]. Description-based methods mainly use inversion-based modifications [12, 22, 29, 54, 79], which first invert an image into a latent representation and then generate edits according to the new prompt. Early instruction-based methods [3, 36, 89] train models on instruction-image paired datasets. More recent approaches [7, 60, 72, 73, 77] unify understanding and generation, interpreting editing instructions and enabling flexible and highly realistic image edits.

2.2 Benchmarks for Image Editing

Existing benchmarks for image editing, along with comparisons to our EditHF-1M, are summarized in Table 1. TedBench [24] is the first image editing benchmark, contains only 100 source images with prompts. EditVal [2] covers more tasks and methods, but the resulting images often suffer from low resolution and blurriness due to data source. EditEval [18] includes more editing tasks, yet scores derived directly from MLLMs and failed to align with human perception. IE-Bench [64] and EBench-18K [80] provide Mean Opinion Scores (MOSs), but are limited in image scale. EditScore [45] and EditReward [75] incorporate pairs of human preferences and demonstrate potential as reward models, but their coverage of editing tasks and source content remains limited, making them less suitable for evaluating flexible editing models. In contrast, EditHF-1M contains over one million edited images spanning 43 editing tasks and 23 editing models, enabling comprehensive evaluation.

2.3 Evaluation Metrics for Image Editing

Traditional image quality assessment (IQA) models include full-reference (FR) IQA [13, 31, 70, 82, 90, 91] and no-reference (NR) IQA [23, 50, 51, 59, 63, 85] metrics. These methods primarily evaluate low-level visual quality and natural distortions, but fail to capture the relationship between the editing instruction and the edited image. To incorporate text prompts, vision-language metrics [16, 26, 34, 37, 78] have been proposed to evaluate alignment for text-based image generation. While effective in measuring image-text correspondence, they often fail to assess the semantic alignment between the editing instruction and the edited image. With the development of MLLMs, many MLLMs demonstrate effectiveness in describing image quality [10, 40–42, 66–68, 81]. Works such as [45, 75, 80] fine-tune MLLMs on human preference data to provide more accurate assessments. [45, 75] further show that learned evaluation models can serve as reward functions to optimize image editing models via reinforcement learning, bridging the gap between evaluation and generation. However, these learned evaluators typically cover a narrow range of editing tasks and source contents,

limiting their applicability to modern, large-scale, and diverse image editing systems. To this end, we propose EditHF, which provides human-aligned feedback across various editing tasks and content, and EditHF-Reward, which leverages EditHF as a reward signal to optimize image editing models.

3 EditHF-1M

In this section, we introduce **EditHF-1M**, the first million-scale image editing benchmark featuring both human preference pairs and fine-grained scores. The benchmark consists of 44.3K high-quality source images with instruction-based and description-based prompts across 43 editing tasks, over one million edited images generated by 23 image editing models, and large-scale human annotations, including **29.1M preference pairs** and **148K scores** covering perceptual quality, editing alignment, and attribute preservation. With its diverse image content and extensive annotations, EditHF-1M serves as a comprehensive resource for evaluating image editing models.

3.1 Design of Editing Tasks

Considering both practical relevance and real-world usage, we select 43 image editing tasks, which are categorized into **global-level tasks**, **object-level tasks**, **human-centric tasks** and **low-level tasks**, as shown in Figure 1. Specifically, global-level tasks involve modifications applied to the entire image, such as overall style transfer or global color adjustment; object-level tasks focus on editing specific objects or regions, such as addition, removal, and attribute modification; human-centric tasks target edits related to human subjects, such as expression change or pose adjustment; and low-level tasks refer to pixel-level transformations aimed at improving image quality, such as denoising, deblurring, and super-resolution. Detailed descriptions of all editing tasks are provided in the Supplementary Material Section 2.

3.2 Data Collection

To construct a comprehensive benchmark for evaluating image editing models, we collect and generate a large-scale dataset of source images, editing prompts and edited images. For global-level, object-level and human-centric tasks, approximately 20K source images are collected from publicly available photography websites, each accompanied by a textual description. Based on the image content and the editing task, we leverage InternVL3.5 [69] to generate corresponding editing prompts, followed by careful manual examination and refinement. To accommodate different types of editing models, the resulting prompts include both instruction-based and description-based formulations. In addition, 26K images are sampled from Pico-Banana-400K [57], which provides instruction-based editing prompts. For these images, we further employ InternVL3.5 [69] to generate complementary description-based prompts and again conduct manual verification and revision. For low-level tasks, both the source images and the corresponding editing prompts are directly adopted from EBench-18K [80].

Next, we select 23 image editing models, covering both description-based and instruction-based approaches, including both open-source and closed-source

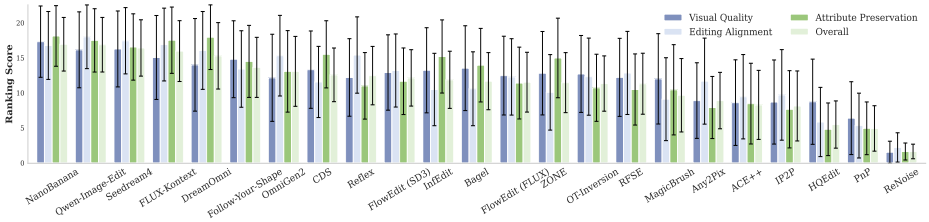

Fig. 2: Comparison of image editing models using EditHF-1M. Ranking scores are derived from win counts in pairwise group comparisons.

methods, with a variety of backbone architectures. Detailed descriptions of these models are provided in the Supplementary Material Section 3. Some models failed to produce valid edits for certain inputs, generating abnormal outputs (e.g., noisy or black images). We have filtered these images, but we ensure the images in test set that the images of each editing task and each model are same for further equal comparison. Finally, using our source images and editing prompts, we generate a total of 1.01M edited images.

3.3 Subjective Experiment

To evaluate the edited images, we conduct a subjective quality assessment experiment based on EditHF-1M, aiming to capture human feedback for edited images and ensure alignment with real-world human perception. Unlike existing benchmarks, which typically annotate scores or rankings alone, we collect both rankings and scores as shown in Figure 1.

Group rankings are effective for comparing results from different models under the same source image and editing prompt, as they reduce individual scoring bias and provide reliable relative judgments. However, rankings are limited to local comparisons and cannot be aggregated across tasks or image contents. In contrast, absolute scoring enables cross-task and cross-content analysis. By combining rankings and scores, our protocol leverages the strengths of both for a more reliable assessment of image editing models. To comprehensively evaluate image editing, we perform the subjective evaluation from three dimensions:

- **Visual Quality:** measures the overall visual quality of the edited image, such as realism, absence of artifacts, structural integrity, color consistency, and the richness of visual details.
- **Editing Alignment:** evaluates the degree to which the edited image correctly follows the given editing instruction, assessing whether the intended modifications are accurately and completely applied.
- **Attribute Preservation:** assesses how well the edited image maintains essential attributes of the source image beyond the intended edits, including preservation of subject identity, key visual characteristics, and contextual consistency.

The experiment is conducted using a Python-based graphical user interface displayed on a calibrated LED monitor with a resolution of 3840 × 2160, where images are presented at a resolution of 1024 × 1024. Annotators are seated approximately 2 feet from the monitor in a controlled environment. Each image

ranking group is annotated by three professional annotators. Each score in the training set is derived from ratings by five professional annotators, while scores in the test set are obtained from fifteen professional annotators. All annotators undergo rigorous training following a standardized protocol. A pre-test is conducted to assess participants' comprehension of the criteria and their alignment with the standard examples. Participants who did not meet the required accuracy threshold were excluded from further participation.

For ranking annotation, EditHF-1M is divided into 44.3K image groups, where each group contains edited images produced by different editing models given the same source image and editing instruction. Annotators are asked to rank the edited images along the three evaluation dimensions independently. Rankings are collected from three annotators per group, and inter-annotator consistency is assessed. Groups exhibiting low agreement are re-annotated by additional three annotators to ensure reliable and consistent rankings. The final ranking for each image group is computed as the average rank from the three annotators. We further generate human preference pairs based on the rankings, for the group including M images, C_M^2 preference pairs generated.

For scoring annotation, 49.4K edited images from EditHF-1M across all the editing tasks and models are selected, and these samples are ensured with diverse source image content and editing prompts. Participants are randomly shown a source image, an edited image, and the corresponding editing prompt, and asked to rate the edited image on a 5-point continuous scale across three aspects. Following [20], those beyond 2 standard deviations from the mean are regarded as outliers and excluded, and participants with over 5% outlier ratings are removed. Remaining ratings are converted to Z-scores and linearly scaled to [0,100]. The final score for each image is computed as follows:

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

where s_{ij} is the raw score given by the i-th subject to the j-th image, μ_i is the mean rating and σ_i is the standard deviation provided by the i-th subject and N_j is the number of ratings for the j-th image. Scores and pairs are further compared and any annotations where the scores and preference pairs are inconsistent are rejected. Detailed descriptions of the annotation protocol and procedures are provided in Supplementary Material, Section 4. in Finally, we have obtained 29.1M pairs and 148K scores from three evaluation dimensions.

3.4 Data Analysis

To benchmark the editing models, we use group rankings from EditHF-1M. Each image is assigned a ranking score based on its win count across these pairwise comparisons. Following [80], an overall score is calculated by combining the visual quality score Score_v, the editing alignment score Score_e, and the attribute preservation score Score_p according to the following equation:

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

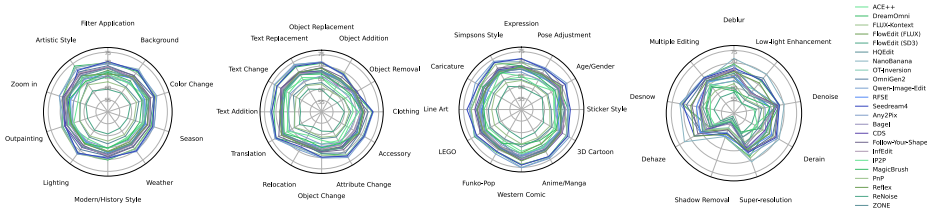

Fig. 3: Comparison of image editing models across different task types: (a) Global-level, (b) Object-level, (c) Human-centric, and (d) Low-level tasks.

The editing alignment score is assigned a higher weight to emphasize the importance of satisfying editing instructions. Figure 2 shows the results of the editing models in EditHF-1M. Qwen-Image-Edit [83] performs best in editing alignment while NanoBanana [7] perform best in other evaluation dimensions.

We further benchmark each model's performance across different editing tasks. For each task, a task-level score is computed by averaging the scores over all editing samples within that task. As shown in Figure 3, models generally perform better on high-level editing tasks, indicating strong capabilities in handling semantic-level modifications. Among the high-level tasks, pose adjustment and object relocation exhibit lower performance, suggesting that precise geometric control and spatial consistency remain challenging. Low-level editing tasks show overall weaker results, reflecting limitations in fine-grained pixel-level editing. Shadow removal performs the worst among all tasks, highlighting the difficulty of modeling complex illumination interactions and preserving visual realism.

4 EditHF

In this section, we present EditHF, a unified image editing evaluation model that supports both fine-grained quality score prediction and pairwise preference comparison, jointly assessing perceptual quality, editing alignment, and attribute preservation in alignment with human perception.

4.1 Model Architecture

Inspired by the success of MLLMs in multimodal reasoning, we adopt an MLLM as the backbone for EditHF. The model jointly reasons over an edited image I_e, the corresponding source image I_s, and the editing prompt T, producing fine-grained scores on three dimensions, as shown in Figure 4.

Visual features are first extracted from I_e and I_s using a pre-trained vision encoder E_v based on InternViT [6]. These features are then projected into the language embedding space via a trainable projector P_φ with parameters φ, yielding the visual tokens

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

Meanwhile, the editing prompt is tokenized into text tokens T_p. The visual and textual tokens are concatenated and fed into an advanced MLLM, InternVL3.5 [69], to perform multimodal feature fusion and joint reasoning. Defining the LLM backbone as H_ψ with trainable parameters ψ, the last hidden states are obtained as $\mathbf{h} = H_\psi(T_e, T_s, T_p)$. To support both textual response generation and fine-grained score prediction, EditHF employs an adaptive decoding strategy. The

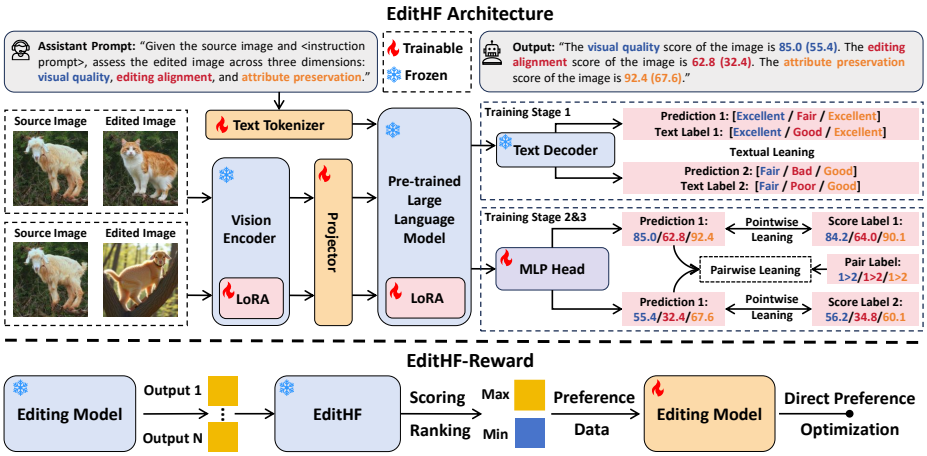

Fig. 4: Overview of the EditHF architecture and EditHF-Reward strategy.

last hidden states $\mathbf{h}$ are first decoded by a text decoder, enabling the model to acquire an initial understanding and generate human-interpretable responses. Once the text output stabilizes, the same hidden states are fed into an MLP head, S_ω with parameters ω, to produce quality scores s_v, s_e, s_p corresponding to the three dimensions. The whole process can be formally written as:

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

where the parameters φ, ψ, ω are optimized during training. The output scores are used to assess perceptual quality, editing alignment, and attribute preservation, and can be further employed for pairwise preference comparison.

4.2 Training Strategy

Our training consists of three stages: **textual learning, pointwise learning, and pairwise learning**, each with a specific loss function targeting distinct objectives.

Textual learning serves as the initial phase, providing the model with a coarse-grained understanding of image editing quality through textual categories. Continuous quality scores are converted into discrete text-based levels, leveraging the fact that MLLMs often exhibit stronger comprehension of textual information than raw numerical values. Specifically, the score range between the maximum and minimum values is uniformly divided into five intervals, corresponding to five quality levels: *bad*, *poor*, *fair*, *good*, and *excellent*. Each score is assigned to its corresponding level based on its interval. This stage is trained using a **cross-entropy loss**.

Pointwise learning aims to provide the model the ability to predict fine-grained quality scores. Leveraging samples with human-annotated scores from EditHF-1M, the model learns to estimate absolute quality values, reflecting direct perceptual judgments. This stage is supervised using the **mean squared error loss**, which encourages the model to approximate human-provided scores as closely as possible.

Pairwise learning focuses on relative quality assessment among images sharing the same source and editing instructions. Preference pairs derived from group rankings in EditHF-1M are used to train the model. Pairs with larger rank differences are prioritized during training, followed by pairs with smaller differences, facilitating a coarse-to-fine understanding of relative quality. The model is optimized with the following **pairwise loss**:

$$\mathcal{L}_{\text{pairwise}} = \frac{1}{N} \sum_{i=1}^{N} \log \Big(1 + \exp(s_{\text{neg},i} - s_{\text{pos},i})\Big), \tag{5}$$

where $s_{\text{pos},i}$ and $s_{\text{neg},i}$ denote the predicted scores for the preferred and less-preferred images in the i-th pair. This loss encourages the model to assign higher scores to perceptually superior images, enhancing its sensitivity to relative image editing quality.

While pointwise training enables the model to predict absolute quality scores that reflect direct human perception, it is inherently insensitive to subtle relative differences among images derived from the same source and editing instruction. Conversely, pairwise training focuses on modeling relative preferences within groups, capturing fine-grained distinctions between similar edits, but it lacks the ability to produce absolute scores that are comparable across different sources or editing instructions. By combining both training paradigms, our model acquires both the capability of absolute score prediction and relative quality comparison. This synergistic training strategy ensures that the model can serve as a reliable evaluation benchmark, while also functioning as a reward model for preference-guided optimization in image editing.

5 EditHF-Reward

In this section, we present EditHF-Reward, a strategy to utilize the EditHF as a reward signal to optimize the image editing models.

5.1 Preference Data Construction

To refine the editing model, preference data are first constructed, as shown in Figure 4. For each source image and editing prompt, a group of edited images is generated. These images are scored using EditHF across three evaluation dimensions, and an overall score is computed using Equation 2. Within each group, images are ranked by their overall scores, with the highest ranked image regarded as the chosen sample and the lowest as the rejected sample. To avoid including ambiguous or failure-prone cases, pairs where the chosen sample has a low overall score are discarded.

5.2 Reinforcement Learning

Direct Preference Optimization (DPO) is employed to align the outputs of image editing models with human preferences. Specifically, we freeze the VAE module and optimize the Transformer by comparing the flow prediction errors between

Table 2: Performance comparisons of quality evaluation methods on EditHF-1M from perspectives of Visual Quality, Editing Alignment, and Attribute Preservation. $SRCC_{global}$ and $PLCC_{global}$ are reported to measure correlation in terms of **score prediction**. $SRCC_{group}$ denotes the average SRCC computed in terms of the **ranking within each editing group**. Acc represents the accuracy in terms of **pairwise comparison prediction**. ♠ Traditional FR IQA metrics, ♡ traditional NR IQA metrics, ♣ deep learning-based FR IQA methods, ♢ deep learning-based NR IQA methods, ★ vision-language methods, ☆ MLLM-based models. The fine-tuned results are marked with *. The best results are highlighted in **red**, and the second-best results are highlighted in **blue**.

Dimension	Visual Quality				Editing Alignment				Attribute Preservation			
Methods/Metrics	$SRCC_{global}$	$PLCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$PLCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$PLCC_{global}$	$SRCC_{group}$	Acc
♡DIIVINE [53]	0.2062	0.3162	0.1795	0.5633	0.0618	0.1947	0.0962	0.5332	0.1641	0.2686	0.1506	0.5522
♡BRISQUE [50]	0.2446	0.2996	0.2020	0.4286	0.0494	0.0975	0.0495	0.4827	0.2019	0.2457	0.1654	0.4426
♡NIQE [51]	0.2816	0.3145	0.2386	0.4142	0.0635	0.1432	0.0811	0.4723	0.2623	0.2778	0.2405	0.4157
♠MSE	0.1421	0.1418	0.2900	0.3968	0.0600	0.0601	0.1730	0.4412	0.3108	0.1893	0.4736	0.3268
♠PSNR	0.1270	0.0232	0.2844	0.6011	0.0572	0.0134	0.1694	0.5575	0.2978	0.1851	0.4690	0.6708
♠GMSD [82]	0.0176	0.0222	0.2715	0.4019	0.0527	0.0486	0.1903	0.4339	0.1926	0.1398	0.4750	0.3265
♠MSSIM [71]	0.1023	0.1041	0.2360	0.5821	0.0894	0.0915	0.1768	0.5597	0.2914	0.2490	0.4563	0.6628
♠SSIM [70]	0.0807	0.0762	0.2454	0.5868	0.0688	0.0646	0.1641	0.5557	0.2448	0.2107	0.4480	0.6605
♠SCSSIM [15]	0.0398	0.0226	0.0867	0.5283	0.0337	0.0329	0.0567	0.5187	0.1140	0.1054	0.3029	0.6053
♠VIF [61]	0.1769	0.0134	0.3703	0.6343	0.2009	0.0922	0.3217	0.6133	0.3367	0.1167	0.5625	0.7087
♠FSIM [90]	0.1777	0.1840	0.3357	0.6219	0.1350	0.1461	0.2461	0.5846	0.3683	0.3271	0.5387	0.6993
♢CNNIQA* [23]	0.6438	0.5824	0.6394	0.7609	0.1927	0.2066	0.1719	0.5605	0.3229	0.2153	0.3086	0.6149
♢MANIQA* [85]	0.8150	0.8104	0.7893	0.8152	0.3068	0.3381	0.2569	0.5907	0.5090	0.5192	0.4968	0.6853
♢TOPIQ* [4]	0.7926	0.8007	0.7359	0.7823	0.2977	0.3175	0.2630	0.5926	0.6006	0.6403	0.5314	0.6936
♢Q-align* [74]	**0.8285**	**0.8403**	0.7659	0.7948	0.3322	0.3520	0.3340	0.6189	0.6401	0.7023	0.5639	0.7049
♣LPIPS [91]	0.2760	0.2629	0.3664	0.3666	0.1783	0.1690	0.2920	0.3988	0.4783	0.4192	0.6007	0.2758
♣ST-LPIPS [13]	0.2896	0.2953	0.3763	0.3611	0.1825	0.1986	0.2903	0.3992	0.4919	0.4376	0.6082	0.2710
♣CVRKD* [88]	0.7649	0.7777	0.7430	0.7917	0.3282	0.3344	0.2943	0.6039	0.7831	0.7989	0.7931	0.8155
♣AHIQ* [31]	0.8172	0.8249	0.8219	0.8326	0.3538	0.3680	0.3254	0.6157	0.8110	0.8210	0.8223	0.8308
★CLIPScore [16]	0.0348	0.0487	0.0879	0.4659	0.1430	0.1585	0.1085	0.5377	0.0422	0.0569	0.1035	0.4597
★BLIPScore [34]	0.1158	0.1235	0.0171	0.4921	0.2562	0.2580	0.2185	0.5805	0.1330	0.1376	0.0345	0.4861
★ImageReward [78]	0.0973	0.1019	0.0555	0.5188	0.2017	0.1894	0.2340	0.5853	0.1092	0.1195	0.0269	0.5062
★HPSv2 [76]	0.1128	0.1291	0.0280	0.5098	0.2912	0.3216	0.2429	0.5851	0.2278	0.2484	0.1643	0.5571
★VQAScore [32]	0.0904	0.0930	0.0029	0.4999	0.2763	0.2850	0.2536	0.5900	0.3573	0.3754	0.2910	0.6032
☆LLaVA-1.5 (7B) [39]	0.0700	0.0942	0.0729	0.1677	0.0253	0.0271	0.0163	0.0396	0.1463	0.1548	0.1122	0.1100
☆mPLUG-Owl3 (7B) [87]	0.1435	0.0801	0.1242	0.5338	0.1173	0.0645	0.1601	0.5606	0.1962	0.0331	0.2368	0.5932
☆MiniCPM-V2.6 (8B) [86]	0.1482	0.0913	0.1224	0.3648	0.2283	0.0220	0.2473	0.4098	0.2765	0.0671	0.2479	0.4029
☆Ovis2.5 (8B) [44]	0.2011	0.0984	0.1858	0.4255	0.2274	0.0377	0.3004	0.4647	0.2467	0.0716	0.2385	0.4376
☆LLama3.2-V (11B) [49]	0.4004	0.2283	0.3115	0.2813	0.3743	0.3334	0.3337	0.2071	0.2181	0.1944	0.0945	0.1526
☆DeepSeekVL2 (small) [38]	0.1634	0.1081	0.0824	0.3440	0.1751	0.0663	0.1491	0.3335	0.2562	0.1232	0.1781	0.3408
☆Qwen2.5-VL (8B) [1]	0.2818	0.3065	0.3362	0.4033	0.4528	0.4572	0.6193	0.5195	0.4664	0.4278	0.5871	0.4869
☆Qwen3-VL (8B) [83]	0.3256	0.2263	0.2982	0.5096	0.4578	0.4606	0.6237	0.4990	0.4791	0.4415	0.6296	0.4884
☆InternVL3 (8B) [92]	0.2462	0.2995	0.2947	0.4710	0.4649	0.4354	0.5853	0.5851	0.3604	0.3095	0.4335	0.5512
☆InternVL3.5 (8B) [69]	0.3607	0.3055	0.3976	0.3148	0.4671	0.3700	0.6051	0.4103	0.5065	0.3324	0.5887	0.4077
☆EditScore (Qwen3) [45]	0.3678	0.3297	0.3730	0.6440	0.3700	0.2657	0.4149	0.6753	0.4110	0.3054	0.4639	0.6863
☆EditReward (MiMo) [75]	0.3940	0.3320	0.4321	0.3412	0.3410	0.3290	0.3705	0.3657	0.3941	0.3544	0.4266	0.3449
☆DeepSeekVL2 (small)* [38]	0.7931	0.7771	0.8111	0.8330	0.8176	0.8077	0.8273	0.8341	0.7790	0.7984	0.8870	0.8693
☆Qwen3-VL (8B)* [83]	0.7751	0.7673	**0.8674**	0.8588	0.8246	0.8137	0.8498	0.8472	**0.8173**	**0.8342**	**0.9167**	**0.8904**
☆InternVL3 (8B)* [92]	0.8044	0.7852	0.8284	**0.8817**	**0.8248**	**0.8156**	**0.8540**	**0.8498**	0.8120	0.8297	0.9137	0.8883
EditHF (Ours)*	**0.8548**	**0.8422**	**0.9223**	**0.9152**	**0.8488**	**0.8348**	**0.9319**	**0.9298**	**0.8475**	**0.8442**	**0.9251**	**0.9197**

the original pre-trained editing model, which serves as the reference, as follows:

$$\begin{aligned} \mathcal{L}(\theta) &= -\mathbb{E}_{(x_0^c, x_0^r)\sim\mathcal{D}_{\text{Gen}},\, t\sim\mathcal{U}(0,T)}\Big[\log\sigma\big(\Delta_\theta(x_t^c,t) - \Delta_\theta(x_t^r,t)\big)\Big], \\ \Delta_\theta(x_t,t) &= \beta_g\Big(\|(\epsilon - x_0) - v_\theta(x_t,t)\|_2^2 - \|(\epsilon - x_0) - v_{\text{ref}}(x_t,t)\|_2^2\Big). \end{aligned} \tag{6}$$

where ϵ denotes the Gaussian noise sampled from a standard normal distribution, which is used to perturb the clean image x_0. x_t^c and x_t^r denote the noisy latents of the chosen and rejected samples at timestep t, respectively. $v_\theta(\cdot)$ and $v_{\text{ref}}(\cdot)$ denote the flow predictions of the fine-tuned and reference editing models. β_g is a temperature parameter controlling the optimization strength, and $\sigma(\cdot)$ denotes the logistic function. This loss encourages the fine-tuned editing model to reduce the denoising error for chosen samples while increasing it for rejected ones, while referencing the editing model before fine-tuning to enhance training stability.

Table 3: Comparisons of the alignment between different evaluation methods and human perception in evaluating image editing models. The best results are highlighted in **red**, and the second-best results are highlighted in **blue**. * denotes fine-tuned models.

Dimensions	Perceptual Quality				Editing Alignment				Attribute Preservation				Overall Score				Overall Rank	
Models/Metrics	Human↑	Ours*	Q-align*	EditReward↑	Human↑	Ours*	Qwen3-VL*	EditReward↑	Human↑	Ours*	AHIQ*	EditScore↑	Human↑	Ours*	EditScore	EditReward↑	Human↓	Ours
NanoBanana [7]	17.54	17.61	16.20	5.863	17.01	17.97	17.24	6.746	18.56	18.43	18.65	18.85	17.68	17.70	5.608	1.810	1	1
Qwen-Image-Edit [72]	15.26	16.53	14.51	6.453	18.17	18.26	17.38	6.952	17.51	17.86	19.95	18.19	17.15	17.33	5.463	1.799	2	2
Seedream4 [60]	15.94	16.77	14.85	6.377	17.31	18.06	17.31	7.013	16.57	17.38	19.34	17.85	16.74	17.19	5.352	1.799	3	4
FLUX-Kontext [30]	15.12	16.37	14.87	6.602	17.12	18.04	17.13	7.126	17.40	18.06	20.13	17.05	16.65	17.25	5.259	1.794	4	3
DreamOmni [77]	14.33	15.91	13.57	6.658	16.14	17.45	16.69	7.109	18.08	18.22	20.06	16.59	16.19	16.93	5.234	1.794	5	5
Follow-Your-Shape [43]	14.85	16.41	14.81	10.29	12.94	16.53	14.81	11.01	14.24	15.80	16.20	13.78	13.92	16.01	4.664	1.695	6	6
OmniGen2 [84]	11.89	15.16	12.60	10.21	15.25	16.86	16.33	10.32	13.26	15.67	18.39	13.41	13.66	15.73	4.581	1.704	7	7
Reflex [25]	12.03	15.47	13.03	10.87	14.83	16.59	15.89	11.08	11.11	14.48	18.07	11.99	12.89	15.36	4.413	1.686	8	8
CDS [54]	13.11	15.59	13.29	11.57	11.07	14.71	14.15	11.95	14.65	16.50	12.47	10.59	12.71	15.23	4.370	1.665	9	9
InfEdit [79]	13.10	15.49	13.52	10.98	10.73	14.30	13.79	10.86	14.80	16.25	14.40	12.78	12.60	14.96	4.552	1.687	10	10
FlowEdit (SD3) [29]	12.07	15.38	12.56	11.22	13.04	14.82	12.86	11.45	11.43	15.20	14.46	13.37	12.30	14.85	4.542	1.676	11	11
Bagel [9]	13.33	15.67	13.23	11.30	10.46	14.26	13.61	11.83	13.41	15.76	15.46	8.310	12.16	14.86	4.040	1.670	12	12
RFSE [65]	12.44	15.62	12.91	11.11	12.86	15.11	14.63	11.40	10.81	14.10	17.27	11.58	12.15	14.71	4.100	1.678	13	14
FlowEdit (FLUX) [29]	12.21	15.46	12.94	10.70	12.26	15.01	14.33	10.99	11.63	14.50	13.97	12.17	12.09	14.74	4.461	1.689	14	13
ZONE [35]	12.17	15.19	12.54	10.95	9.934	14.17	13.28	11.59	14.30	16.21	12.42	12.62	11.84	14.81	4.472	1.678	15	15
OT-Inversion [46]	12.48	15.62	12.44	12.26	12.14	14.91	14.32	12.58	10.39	14.06	16.54	10.47	11.74	14.61	3.948	1.648	16	16
Any2Pix [36]	9.270	14.12	10.06	14.93	11.64	14.41	13.89	14.79	8.384	12.49	17.32	8.400	9.951	13.51	3.459	1.583	17	17
MagicBrush [89]	11.69	14.02	12.36	11.41	8.504	10.77	10.31	10.75	9.693	13.10	12.22	12.34	9.772	12.16	4.388	1.682	18	18
ACE++ [48]	8.925	13.52	9.703	13.86	9.781	11.86	12.67	13.76	8.905	9.961	11.79	8.760	9.278	11.54	3.832	1.610	19	19
IP2P [3]	8.557	12.93	9.904	11.33	8.637	10.10	9.751	11.15	6.968	9.844	11.38	13.35	8.116	10.63	4.499	1.678	20	20
HQEdit [19]	8.834	14.32	10.05	20.20	6.175	8.588	8.420	20.20	5.275	9.011	7.398	4.040	6.641	10.08	2.655	1.441	21	21
PnP [22]	6.608	12.23	8.570	11.90	5.917	9.502	9.171	11.21	5.416	9.431	9.591	12.97	5.959	10.07	4.523	1.669	22	22
ReNoise [12]	1.544	1.402	3.446	12.52	2.367	2.193	2.400	11.19	1.719	1.572	1.377	12.68	1.892	1.736	4.361	1.661	23	23
SRCC to human ↑		**0.976**	**0.951**	0.766		**0.987**	**0.952**	0.632		**0.996**	**0.734**	0.612		**0.998**	0.755	**0.833**		**0.998**

6 Experiments

In this section, we evaluate the performance of EditHF as both an evaluation and reward model through extensive experiments.

6.1 Experiment Setup

We train the EditHF on our EditHF-1M, which is split to training, validating and testing set with a 5:1:1 ratio. Each image in the testing set includes both ranking and score annotation. We utilize the LoRA [17] for fine-tuning, with LoRA modules are applied to both vision model and language model. The models are implemented with PyTorch and trained on 2 NVIDIA RTX A6000 GPU.

6.2 Evaluation on EditHF-1M

Table 2 presents the performance of EditHF compared with traditional and deep learning-based FR IQA and NR IQA methods, vision-language models, and MLLM-based approaches. We report the Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) over the test set to evaluate overall score prediction, denoted as SRCC_{global} and PLCC_{global}. To evaluate pairwise comparisons, we report the average SRCC and comparison accuracy among image groups sharing the same source image and editing prompt, denoted as SRCC_{group} and Acc.

Traditional IQA methods perform poorly, as their features are primarily designed to capture low-level distortions and are ineffective for high-level semantic changes. Deep learning-based IQA methods achieve better results in visual quality assessment but still struggle to evaluate editing alignment. Vision-language models show limited effectiveness because they focus on text-image alignment and fail to capture editing semantics. Zero-shot performance of MLLM-based methods is also poor due to insufficient understanding of the editing evaluation task. EditScore [45] and EditReward [75], which are trained on other image editing datasets, exhibit limited performance due to the limited size of their dataset. Our EditHF achieves the best performance in aligning with human perception across all evaluation dimensions. Table 2 further shows that MLLMs fine-tuned

Table 4: Ablation study on the different backbones and training strategy.

Backbone&Strategy				Perceptual Quality			Editing Alignment			Attribute Preservation		
Backbone	MLP Head	Pointwise Training	Pairwise Training	$SRCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$SRCC_{group}$	Acc	$SRCC_{global}$	$SRCC_{group}$	Acc
InternVL3.5 [69]		✓		0.7184	0.7215	0.7904	0.7025	0.7312	0.7750	0.7241	0.7506	0.7862
InternVL3.5 [69]	✓	✓		0.7762	0.7829	0.7826	0.7806	0.7964	0.7825	0.7752	0.7864	0.7821
InternVL3.5 [69]	✓		✓	0.8018	0.8792	0.8663	0.8213	0.8894	0.8762	0.8173	0.8726	0.8765
InternVL3.5 [69]	✓	✓	✓	**0.8548**	**0.9223**	**0.9152**	**0.8488**	**0.9319**	**0.9298**	**0.8475**	**0.9251**	**0.9197**
InternVL3 [5]	✓	✓	✓	0.8316	0.8964	0.8826	0.8155	0.9109	0.9035	0.8141	0.9022	0.8943
Qwen3-VL [83]	✓	✓	✓	0.8417	0.9032	0.8985	0.8231	0.9127	0.9046	0.8240	0.9068	0.9074
DeepSeekVL2 [38]	✓	✓	✓	0.8274	0.8823	0.8970	0.8285	0.8963	0.8856	0.8106	0.8769	0.8811
LLaVA-NeXT [33]	✓	✓	✓	0.8164	0.8723	0.8657	0.8108	0.8814	0.8741	0.8037	0.8806	0.8794

Table 5: Comparison results on other public image editing benchmarks. The best results are highlighted in **red**, and the second-best results are highlighted in **blue**. "-" indicates that the model was trained on this dataset, and thus its result is not reported to ensure a fair cross-dataset evaluation.

Methods/Benchmarks	GenAI-Bench [21]	AROURA-Bench [27]	ImagenHub [28]	EBench-18K [80]	EditScore-Bench [45]	EditReward-Bench [75]
GPT-4o [55]	0.5354	0.5081	0.3821	0.4108	0.6730	0.2831
GPT-5 [56]	0.5961	0.4727	0.4085	0.4260	**0.7524**	0.3781
Gemini-2.0-Flash [14]	0.5332	0.4431	0.2369	0.3892	0.6958	0.3347
Gemini-2.5-Flash [7]	0.5701	0.4763	**0.4162**	**0.4337**	0.7026	0.3802
Qwen3-VL (7B) [83]	0.4458	0.3428	0.2426	0.3726	0.4251	0.2731
InternVL3.5 (7B) [69]	0.4517	0.3692	0.2071	0.3511	0.4518	0.2406
EditScore (Qwen3) [45]	0.5024	0.4218	0.3062	0.3206	-	0.2590
EditReward (MiMo) [75]	**0.6572**	0.6362	0.3520	0.3718	0.7125	-
LMM4Edit [80]	0.6327	**0.6521**	0.3726	-	0.7064	**0.3958**
EditHF (Ours)	**0.6792**	**0.6608**	**0.4526**	**0.4627**	**0.7750**	**0.4325**

on our EditHF-1M achieve significant improvements, demonstrating the value of our dataset for advancing image editing quality assessment.

Table 3 demonstrates the effectiveness of our model to benchmark image editing models. Our model achieves the highest SRCC with human scores, demonstrating its strong capability in benchmarking image editing models. Moreover, the results reveal a clear performance gap between closed-source image editing methods and open-source models, motivating us to further refine open-source models using data derived from closed-source methods.

6.3 Ablation Study

We conduct comprehensive ablation studies, with the results shown in Table 4. Rows 1-2 illustrate the importance of the MLP head for precise score regression. Rows 2-4 demonstrate that combining pointwise and pairwise training yields the best performance. Rows 4-8 compare different MLLM backbones with similar parameter scales, among which InternVL3.5 achieves the best results.

6.4 Zero-shot Performance on Other Benchmarks

Table 5 presents the performance of EditHF compared with other advanced MLLMs on several image editing benchmarks. For benchmarks with pairwise annotations, including GenAI-Bench [21], AROURA-Bench [27], and EditScore-Bench [45], we report prediction accuracy. For benchmarks with pointwise annotations, including ImagenHub [28], EBench-18K [80], and EditReward-Bench [75], we report SRCC with human ratings. For benchmarks that provide multiple evaluation dimensions, including ImagenHub [28], EBench-18K [80], EditScore-Bench [45], and EditReward-Bench [75], we report the average result across different dimensions. The results show that EditHF achieves strong performance across diverse benchmarks, showing its robust generalization ability.

Table 6: Results of Qwen-Image-Edit [72] refined with EditHF-Reward on different benchmarks. Overall scores from human study and our EditHF are reported.

Benchmarks	EditHF-1M		AROURA-Bench [27]		ImagenHub [28]		EBench-18K [80]		EditScore-Bench [45]		EditReward-Bench [75]	
Methods/Metrics	Human	Ours	Human	Ours	Human	Ours	Human	Ours	Human	Ours	Human	Ours
Qwen-Image-Edit	75.39	78.06	83.21	68.47	75.64	73.26	73.25	78.29	70.17	75.22	80.15	77.03
Qwen-Image-Edit (Self-DPO)	78.36	81.55	84.28	69.26	76.55	75.14	73.80	79.05	72.18	76.55	81.17	78.46
Improvement	+3.94%	+4.47%	+1.29%	+1.15%	+1.20%	+2.57%	+0.75%	+0.97%	+2.86%	+1.77%	+1.27%	+1.86%
Qwen-Image-Edit (Global-DPO)	84.26	86.15	85.46	69.12	79.12	78.06	74.18	79.60	75.23	78.41	82.60	78.12
Improvement	+11.77%	+10.36%	+2.70%	+0.95%	+4.60%	+6.55%	+1.27%	+1.67%	+7.21%	+4.24%	+3.06%	+1.42%

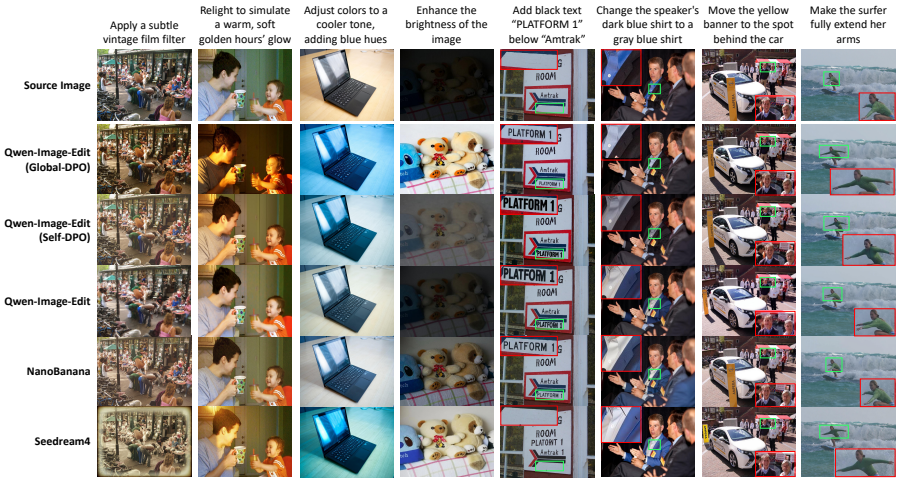

Fig. 5: Examples from Qwen-Image-Edit refined with EditHF-Reward and other competitive editing models.

6.5 Refinement Results with EditHF-Reward

To validate the effectiveness of EditHF as a reward model, we apply EditHF-Reward to refine Qwen-Image-Edit [72] that achieves the best performance on EditHF-1M among all open-source editing models. We use two complementary methods, Self-DPO and Global-DPO, to construct preference pairs for refinement. In Self-DPO, edited image groups are generated using Qwen-Image-Edit with 20 different seeds, allowing us to assess the ability of EditHF to guide a single model. However, relying on a model to generate outputs may reinforce its inherent biases and overlook alternative high-quality outputs. In Global-DPO, edited image groups are generated from multiple editing models in EditHF-1M, capturing diverse editing styles across models. Preference pairs are then obtained via EditHF, and within each editing task, only pairs from the top half of scored samples are retained to keep high-quality edits. In total, this process yields 18K high-quality preference pairs for Self-DPO and 22K for Global-DPO.

We compare the performance of the original Qwen-Image-Edit with the version refined using EditHF-Reward across several benchmarks. A subjective study is conducted following the scoring protocol described in Section 3.3. Results in Table 6 show that Qwen-Image-Edit refined with EditHF-Reward consistently achieves performance gains across these benchmarks in both human evaluations and our EditHF metric, with greater gains observed when using Global-DPO.

Figure 5 presents editing examples produced by refined Qwen-Image-Edit, alongside those from other competitive editing models, demonstrating that our refined model exhibits a strong capability to handle fine-grained editing tasks with higher fidelity and better alignment with the given instructions.

7 Conclusion

In this paper, we introduce EditHF-1M, the first million-scale dataset for multi-dimensional evaluation of image editing including both preference pairs and fine-grained scores from rigorous human annotations. Based on this benchmark, we propose EditHF, which effectively supports both score regression and pairwise comparison. We further propose the EditHF-Reward strategy, which utilizes EditHF as a reward signal for model refinement. Extensive experiments demonstrate that EditHF achieves state-of-the-art performance and exhibits strong generalization ability and EditHF-Reward can effectively improve image editing models.

8 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62401365, 62225112, 62522116, 62271312, 62132006, U24A20220, and in part by the China Postdoctoral Science Foundation under Grants BX20250411, 2025M773473, and in part by STCSM under Grant 22DZ2229005.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 11
2. Basu, S., Saberi, M., Bhardwaj, S., Chegini, A.M., Massiceti, D., Sanjabi, M., et al.: Editval: Benchmarking diffusion based text-guided image editing methods. arXiv preprint arXiv:2310.02426 (2023) 2, 4
3. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18392–18402 (2023) 1, 4, 12
4. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing (TIP) **33**, 2404–2418 (2024) 11
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2025) 13
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198 (2024) 8
7. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025) 1, 3, 4, 8, 12, 13
8. Damera-Venkata, N., Kite, T.D., Geisler, W.S., Evans, B.L., Bovik, A.C.: Image quality assessment based on a degradation model. IEEE Transactions on Image Processing (TIP) **9**(4), 636–650 (2000) 2
9. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 12
10. Duan, H., Hu, Q., Wang, J., Yang, L., Xu, Z., Liu, L., Min, X., Cai, C., Ye, T., Zhang, X., Zhai, G.: Finevq: Fine-grained user generated content video quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025) 4
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the International Conference on Machine Learning (ICML) (2024) 4
12. Garibi, D., Patashnik, O., Voynov, A., Averbuch-Elor, H., Cohen-Or, D.: Renoise: Real image inversion through iterative noising. In: European Conference on Computer Vision (ECCV). pp. 395–413 (2024) 4, 12
13. Ghildyal, A., Liu, F.: Shift-tolerant perceptual similarity metric. In: Proceedings of the European Conference on Computer Vision (ECCV). p. 91–107 (2022) 4, 11
14. Google: Gemini 2.0. https://deepmind.google/technologies/gemini/ (2024) 13
15. Gu, K., Zhai, G., Yang, X., Zhang, W.: An improved full-reference image quality metric based on structure compensation. In: Proceedings of the conference on Asia-Pacific Signal and Information Processing Association (APSIPA). pp. 1–6 (2012) 11
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 2, 3, 4, 11
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al.: Lora: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022) 12

18. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., et al.: Diffusion model-based image editing: A survey. IEEE transactions on pattern analysis and machine intelligence (TPAMI) **PP** (2024) 2, 3, 4
19. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., et al.: Hq-edit: A high-quality dataset for instruction-based image editing. arXiv preprint arXiv:2404.09990 (2024) 12
20. (ITU), I.T.U.: Methodology for the subjective assessment of the quality of television pictures. Tech. Rep. Rec. ITU-R BT.500-13, International Telecommunication Union (ITU) (Jan 2012) 7
21. Jiang, D., Ku, M., Li, T., Ni, Y., Sun, S., Fan, R., et al.: Genai arena: An open evaluation platform for generative models. arXiv preprint arXiv:2406.04485 (2024) 13
22. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024) 4, 12
23. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 4, 11
24. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., et al.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 4
25. Kim, J., Park, J., Song, Y., Kwak, N., Rhee, W.: Reflex: Text-guided editing of real images in rectified flow via mid-step feature extraction and attention adaptation. arXiv preprint arXiv:2507.01496 12
26. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). pp. 36652–36663 (2023) 4
27. Krojer, B., Vattikonda, D., Lara, L., Jampani, V., Portelance, E., Pal, C., et al.: Learning Action and Reasoning-Centric Image Editing from Videos and Simulations. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2024) 2, 13, 14
28. Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., et al.: Imagenhub: Standardizing the evaluation of conditional image generation models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024) 2, 13, 14
29. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19721–19730 (2025) 4, 12
30. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., et al.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint (2025) 12
31. Lao, S., Gong, Y., Shi, S., Yang, S., Wu, T., Wang, J., et al.: Attentions help cnns see better: Attention-based hybrid image quality assessment network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 1140–1149 (2022) 4, 11
32. Li, B., Lin, Z., Pathak, D., Li, J., Fei, Y., Wu, K., et al.: Evaluating and improving compositional text-to-visual generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 11

33. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., et al.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024) 13
34. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 12888–12900 (2022) 2, 3, 4, 11
35. Li, S., Zeng, B., Feng, Y., Gao, S., Liu, X., Liu, J., et al.: Zone: Zero-shot instruction-guided local editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6254–6263 (2024) 12
36. Li, S., Singh, H., Grover, A.: Instructany2pix: Flexible visual editing via multimodal instruction following. arXiv preprint arXiv:2312.06738 (2023) 4, 12
37. Li, Y., Liu, H., Cai, M., Li, Y., Shechtman, E., Lin, Z., et al.: Removing distributional discrepancies in captions improves image-text alignment. arXiv preprint arXiv:2410.00905 (2024) 3, 4
38. Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., et al.: Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434 (2024) 11, 13
39. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (2024) 11
40. Liu, L., Cai, C., Shen, S., Liang, J., Ouyang, W., Ye, T., Mao, J., Duan, H., Yao, J., Zhang, X., et al.: Moa-vr: A mixture-of-agents system towards all-in-one video restoration. IEEE Journal of Selected Topics in Signal Processing (2025) 4
41. Liu, L., Duan, H., Hu, Q., Yang, L., Cai, C., Ye, T., Liu, H., Zhang, X., Zhai, G.: F-bench: Rethinking human preference evaluation metrics for benchmarking face generation, customization, and restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10982–10994 (2025) 4
42. Liu, L., Duan, H., Zhu, C., Lu, J., Jiang, H., Yang, L., Hu, Q., Zhai, G., Zhang, X.: Ll-bench: Rethinking low-level vision evaluation in the era of large-scale generative models. arXiv preprint arXiv:2606.02535 (2026) 4
43. Long, Z., Zheng, M., Feng, K., Zhang, X., Liu, H., Yang, H., et al.: Follow-your-shape: Shape-aware image editing via trajectory-guided region control. arXiv preprint arXiv:2508.08134 (2025) 12
44. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., et al.: Ovis2.5 technical report. arXiv:2508.11737 (2025) 11
45. Luo, X., Wang, J., Wu, C., Xiao, S., Jiang, X., Lian, D., et al.: Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. arXiv preprint arXiv:2509.23909 (2025) 2, 3, 4, 11, 12, 13, 14
46. Lupascu, M., Stupariu, M.S.: Optimal transport for rectified flow image editing: Unifying inversion-based and direct methods. arXiv preprint arXiv:2508.02363 (2025) 12
47. Ma, Y., Ji, J., Ye, K., Lin, W., Zheng, Y., Zhou, Q., et al.: I2ebench: A comprehensive benchmark for instruction-based image editing. In: Proceedings of the Advances in Neural Information Processing Systems (NeurPIS) (2024) 2, 3
48. Mao, C., Zhang, J., Pan, Y., Jiang, Z., Han, Z., Liu, Y., Zhou, J.: Ace++: Instruction-based image creation and editing via context-aware content filling. arXiv preprint arXiv:2501.02487 (2025) 1, 12
49. Meta, A.: Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Meta AI Blog. Retrieved December (2024) 11

50. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: Proceedings of the Asilomar Conference on Signals, Systems and Computers (ACSSC). pp. 723–727 (2011) 2, 4, 11
51. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters (SPL) **20**(3), 209–212 (2013) 4, 11
52. Moorthy, A.K., Bovik, A.C.: A modular framework for constructing blind universal quality indices. IEEE Signal Processing Letters (SPL) (2009) 2
53. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing (TIP) **20**(12), 3350–3364 (2011) 11
54. Nam, H., Kwon, G., Park, G.Y., Ye, J.C.: Contrastive denoising score for text-guided latent diffusion image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9192–9201 (June 2024) 4, 12
55. OpenAI: Chatgpt-4o: Advanced multimodal chat model (2025), https://openai.com/chatgpt 13
56. OpenAI: Gpt-5. https://www.openai.com (2025) 13
57. Qian, Y., Bocek-Rivele, E., Song, L., Tong, J., Yang, Y., Lu, J., et al.: Pico-banana-400k: A large-scale dataset for text-guided image editing. arXiv preprint arXiv:2510.19808 (2025) 3, 5
58. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) 4
59. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. IEEE Transactions on Image Processing (TIP) **21**(8), 3339–3352 (2012) 4
60. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint (2025) 1, 3, 4, 12
61. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing (TIP) **15**(2), 430–444 (2006) 2, 11
62. Sheikh, H.R., Bovik, A.C., Veciana, G.D.: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Transactions on Image Processing (TIP) **14**(12), 2117–2128 (2005) 2
63. Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., et al.: Blindly assess image quality in the wild guided by a self-adaptive hyper network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4
64. Sun, S., Qu, B., Liang, X., Fan, S., Gao, W.: Ie-bench: Advancing the measurement of text-driven image editing for human perception alignment. arXiv preprint arXiv:2501.09927 (2025) 2, 4
65. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., et al.: Taming rectified flow for inversion and editing. arXiv preprint arXiv:2411.04746 (2024) 12
66. Wang, J., Duan, H., Zhao, Y., Wang, J., Zhai, G., Min, X.: Lmm4lmm: Benchmarking and evaluating large-multimodal image generation with lmms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR). pp. 17312–17323 (2025) 4

67. Wang, J., Wang, J., Duan, H., Kang, J., Zhai, G., Min, X.: I2i-bench: A comprehensive benchmark suite for image-to-image editing models. arXiv preprint arXiv:2512.04660 (2025) 4
68. Wang, P., Sun, W., Zhang, Z., Jia, J., Jiang, Y., Zhang, Z., et al.: Large multimodality model assisted ai-generated image quality assessment. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). p. 7803–7812 (2024) 4
69. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 5, 8, 11, 13
70. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP) **13**(4), 600–612 (2004) 2, 4, 11
71. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Proceedings of the Asilomar Conference on Signals, Systems & Computers (ACSSC). vol. 2, pp. 1398–1402 (2003) 2, 11
72. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025) 1, 3, 4, 12, 14
73. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025) 3, 4
74. Wu, H., Zhang, Z., Zhang, W., Chen, C., Li, C., Liao, L., et al.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090 (2023) 11
75. Wu, K., Jiang, S., Ku, M., Nie, P., Liu, M., Chen, W.: Editreward: A human-aligned reward model for instruction-guided image editing. arXiv preprint arXiv:2509.26346 (2025) 2, 3, 4, 11, 12, 13, 14
76. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 2, 3, 11
77. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., et al.: Dreamomni2: Multimodal instruction-based editing and generation (2025) 4, 12
78. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., et al.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS). pp. 15903–15935 (2023) 2, 3, 4, 11
79. Xu, S., Huang, Y., Pan, J., Ma, Z., Chai, J.: Inversion-free image editing with natural language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9192–9201 (2024) 4, 12
80. Xu, Z., Duan, H., Liu, B., Ma, G., Wang, J., Yang, L., et al.: Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. In: Proceedings of the ACM International Conference on Multimedia (ACM MM). pp. 6908–6917 (October 2025) 2, 3, 4, 5, 7, 13, 14
81. Xu, Z., Duan, H., Ma, G., Yang, L., Wang, J., Wu, Q., et al.: Harmonyiqa: Pioneering benchmark and model for image harmonization quality assessment. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2025) 4
82. Xue, W., Zhang, L., Mou, X., Bovik, A.C.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Transactions on Image Processing (TIP) **23**(2), 684–695 (2013) 4, 11

83. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 8, 11, 13
84. Yang, L., Duan, H., Zhu, Y., Liu, X., Liu, L., Xu, Z., Ma, G., Min, X., Zhai, G., Callet, P.L.: Omni^2: Unifying omnidirectional image generation and editing in an omni model (2025) 12
85. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., et al.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1191–1200 (2022) 4, 11
86. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024) 11
87. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., et al.: mplug-owl3: Towards long image-sequence understanding in multimodal large language models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024) 11
88. Yin, G., Wang, W., Yuan, Z., Han, C., Ji, W., Sun, S., et al.: Content-variant reference image quality assessment via knowledge distillation. In: Proceedings of the Conference on Association for the Advancement of Artificial Intelligence (AAAI). vol. 36, pp. 3134–3142 (2022) 11
89. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) (2023) 1, 4, 12
90. Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing (TIP) **20**(8), 2378–2386 (2011) 2, 4, 11
91. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2, 4, 11
92. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025) 11