MR-IQA-2: Faithful Image Quality Reflection via Fine-Grained Credit Assignment

Yuan Li, Youyuan Lin, Chenhui Chu, Shin'ya Nishida

Graduate School of Informatics, Kyoto University


MR-IQA-2

arXiv:2608.18579v1 [cs.CV] 19 Aug 2026

Abstract

Multimodal large language models (MLLMs) have shown great potential to improve image quality assessment (IQA) by making predictions more explainable through increased consistency between quality ratings and their underlying reasoning. However, most existing approaches supervise reasoning only to align it with human-provided quality ratings, paying little attention to whether the reasoning faithfully reflects the image's actual quality content. Since higher rating accuracy alone does not guarantee faithful reasoning, using a shared reward for both rating and reasoning obscures the source of supervision and may reinforce unfaithful reasoning when a correct rating is obtained by chance. To improve the faithfulness and reliability of blind IQA, we propose to (1) decouple credit assignment for reasoning and rating, and (2) provide verifiable supervision signals for faithful reasoning. To this end, we introduce MR-IQA-2, an actor-editor-judge framework for faithful IQA that operationalizes reasoning-editing-reflection. The actor first generates quality reasoning for an input image. Conditioned on this reasoning, the editor revises the image to provide a verifiable visual signal for the identified quality factors. A frozen judge then compares the original and edited images, producing reflective supervision that improves the actor's quality reasoning. MR-IQA-2 uses fine-grained credit assignment to decouple reasoning and rating supervision. Judge feedback supervises reasoning, whereas human ratings supervise the predicted rating. Masked token-specific updates distinguish these signals while preserving the causal relation from reasoning to rating. Across IQA benchmarks, MR-IQA-2 achieves competitive rating alignment with humans. Additionally, visual reflection enables the framework to acquire richer and more faithful visual understanding beyond rating. This faithful visual understanding may inform image-quality optimization and related downstream tasks. Code is available at MR-IQA-2.

1 Introduction

Image quality assessment (IQA) aims to understand how humans perceive visual quality and to model the relationship between an image and its perceived quality rating. Prior IQA work has undergone a series of framework transitions as visual content and degradation patterns have become increasingly complex. The goal of IQA is no longer limited to producing a numerical rating. A more important question is whether a model can develop a faithful understanding of image quality. Such understanding should go beyond specific degradations and cover both low-level visual attributes and semantic-level perception.

Early blind image quality assessment (BIQA) methods were largely driven by degradation-centered assumptions. In this setting, image quality was often associated with the strength of synthetic distortions, and the learning objective was commonly formulated as degradation-scale estimation. Methods such as ARNIQA (Agnolucci et al. 2024), TOPIQ (Chen et al. 2024), and LIQE (Zhang et al. 2023) represent important attempts to learn quality-aware representations under this paradigm.

With the emergence of practical IQA benchmarks such as KonIQ-10k (Hosu et al. 2020) and SPAQ (Fang et al. 2020), the focus of BIQA has gradually shifted. Real-world images contain diverse content, complex capture conditions, aesthetic factors, and semantic preferences. These factors cannot be fully described by synthetic distortion types. As a result, BIQA has moved from estimating degradation intensity toward understanding the image itself.

Recent multimodal approaches further extend this trend. Works such as Q-Instruct (Wu et al. 2024), Q-Insight (Li et al. 2025a), and VisualQuality-R1 (Wu et al. 2025) introduce language, criteria, and human-like reasoning into BIQA. These methods allow models to evaluate images through quality dimensions, ranking preferences, and explanatory outputs. This transition improves interpretability and generalization via criteria closer to human perception. But it also raises a deeper question: does the generated reasoning faithfully reflect the actual factors that determine image quality?

This question is critical for BIQA based on multimodal large language models (MLLMs). MLLMs can produce fluent and plausible quality explanations, but plausible language does not guarantee faithful visual reasoning. A model may attribute low quality to reasonable-sounding factors while failing to identify the true visual limitations. Causal reasoning collapse, unverifiable explanations, and deviations from human perception can therefore appear in quality assessment. Previous studies, including BRIQA (Li et al. 2025b) and H-IQA (Li et al. 2025c), have attempted to analyze reasoning failures, improve reasoning consistency, and align IQA with human perception. However, the reasoning process itself remains difficult to verify.

Recent methods, including Zoom-IQA (Liang et al. 2026) and Tool-IQA (Qin et al. 2026), move BIQA toward

tool-augmented reasoning through region-aware inspection. These methods show that additional visual observations help models inspect image details more reliably. Nevertheless, they focus on observation and grounding rather than verifying whether the proposed factors actually limit image quality. It therefore remains unclear whether a model has identified the factors that genuinely constrain overall image quality.

Another persistent issue in reinforcement learning (RL)-based BIQA methods, including Q-Insight (Li et al. 2025a), VisualQuality-R1 (Wu et al. 2025), MR-IQA (Li et al. 2026), Zoom-IQA (Liang et al. 2026), and Tool-IQA (Qin et al. 2026), is that reasoning supervision is often determined by rating performance. Consequently, incorrect reasoning may still receive high rewards when the rating is accurate, potentially undermining reasoning faithfulness.

In this work, we decompose the learning of faithful quality reasoning into two tasks: (1) decoupling reasoning and rating supervision and (2) constructing an independent supervision signal for reasoning. For the first task, we propose fine-grained credit assignment. We use masked token-specific updates so that better reasoning or a more accurate rating is rewarded accordingly within its corresponding output while preserving causality. The second task is therefore to construct a reliable supervision signal for reasoning. We use visual reflection to estimate this signal from observed quality changes. A reasoning proposal receives higher credit when its corresponding edit alleviates the identified quality-limiting factors. Finally, we integrate these two mechanisms into an actor-editor-judge framework termed MR-IQA-2.

Prior work primarily pursues stronger rating performance, as exemplified by MR-IQA (Li et al. 2026), or richer reasoning processes, as in Zoom-IQA (Liang et al. 2026) and Tool-IQA (Qin et al. 2026). In contrast, MR-IQA-2 investigates whether the generated reasoning faithfully identifies quality-limiting factors and yields visual understanding that can inform broader downstream tasks, including image editing and visual recognition.

The main contributions are summarized as follows:

- We introduce fine-grained credit assignment for the reasoning–rating structure in BIQA. Masked token-specific updates decouple reasoning and rating supervision, assigning signal to its corresponding output while preserving the causal relation from reasoning to rating.
- We construct a verifiable supervision signal for faithful quality reasoning through visual reflection. Reasoning-conditioned editing serves as a visual intervention, while a frozen judge evaluates the observed quality change and assigns credit to the corresponding reasoning.
- We integrate these two mechanisms into MR-IQA-2, a modular actor-editor-judge framework that operationalizes reasoning-editing-reflection. The actor, editor, and judge can be independently instantiated or replaced for different models and downstream settings.

2 Related Work

2.1 Faithful Quality Reasoning with MLLMs.

The adoption of MLLMs has improved the interpretability of BIQA by enabling language-based image understanding and quality explanations. However, faithful quality reasoning remains a key challenge. Early methods such as DepictQA (You et al. 2024) and Q-Instruct (Wu et al. 2024) mainly rely on supervised fine-tuning (SFT), which encourages plausible explanation patterns. Constrained by their training data, these models may overfit templated responses and exhibit limited logical coherence. Later RL-based frameworks, including Q-Insight (Li et al. 2025a), VisualQuality-R1 (Wu et al. 2025), H-IQA (Li et al. 2025c), and MR-IQA (Li et al. 2026), optimize generated reasoning using rating supervision. By strengthening the model's underlying perceptual capabilities, these methods mitigate template overfitting, produce richer and more consistent reasoning, and improve rating alignment. Nevertheless, the reasoning itself is generated by the MLLM backbone without a reliable reasoning-specific supervision signal, making its faithfulness difficult to verify. Our framework instead evaluates reasoning faithfulness through visual intervention by verifying whether edits derived from the identified reasoning factors lead to observable improvements in image quality.

2.2 Tool-Augmented IQA Methods

Before MLLM-based BIQA, degradation modeling and task-specific image processing were already used to derive quality cues. ARNIQA (Agnolucci et al. 2024) applies predefined degradation operators to construct contrastive views and learn distortion-aware representations. More recently, Zoom-IQA (Liang et al. 2026) introduces region-aware zooming for detail inspection. Tool-IQA (Qin et al. 2026) further employs a magnifier and a gamma corrector to provide additional visual evidence for quality rating. Although these tools enable richer and more localized inspection, the resulting reasoning is still evaluated mainly through rating performance. Rather than enriching the visual evidence during reasoning for rating optimization, our design provides direct supervision for the faithfulness of quality reasoning.

3 Methods

3.1 MR-IQA-2 Framework Overview

In this work, we propose MR-IQA-2 (Figure 1), an actor-editor-judge framework for faithful IQA. MR-IQA-2 makes quality reasoning verifiable rather than assessing it solely through rating performance. The actor is a general MLLM-based BIQA model that produces interpretable quality reasoning and a rating. A frozen editor, initialized from an image-editing diffusion model, applies edits conditioned on the reasoning. A frozen MLLM judge then evaluates the quality change between the original and edited images. The observed change provides a reasoning-specific supervision signal for the actor. The following sections introduce (1) the interaction among the actor, editor, and judge, (2) fine-grained credit assignment, and (3) the training optimization procedure.

3.2 Actor-Editor-Judge Trajectory

This section describes the interaction flow among the Actor, Editor, and Judge. For a local batch of N images,

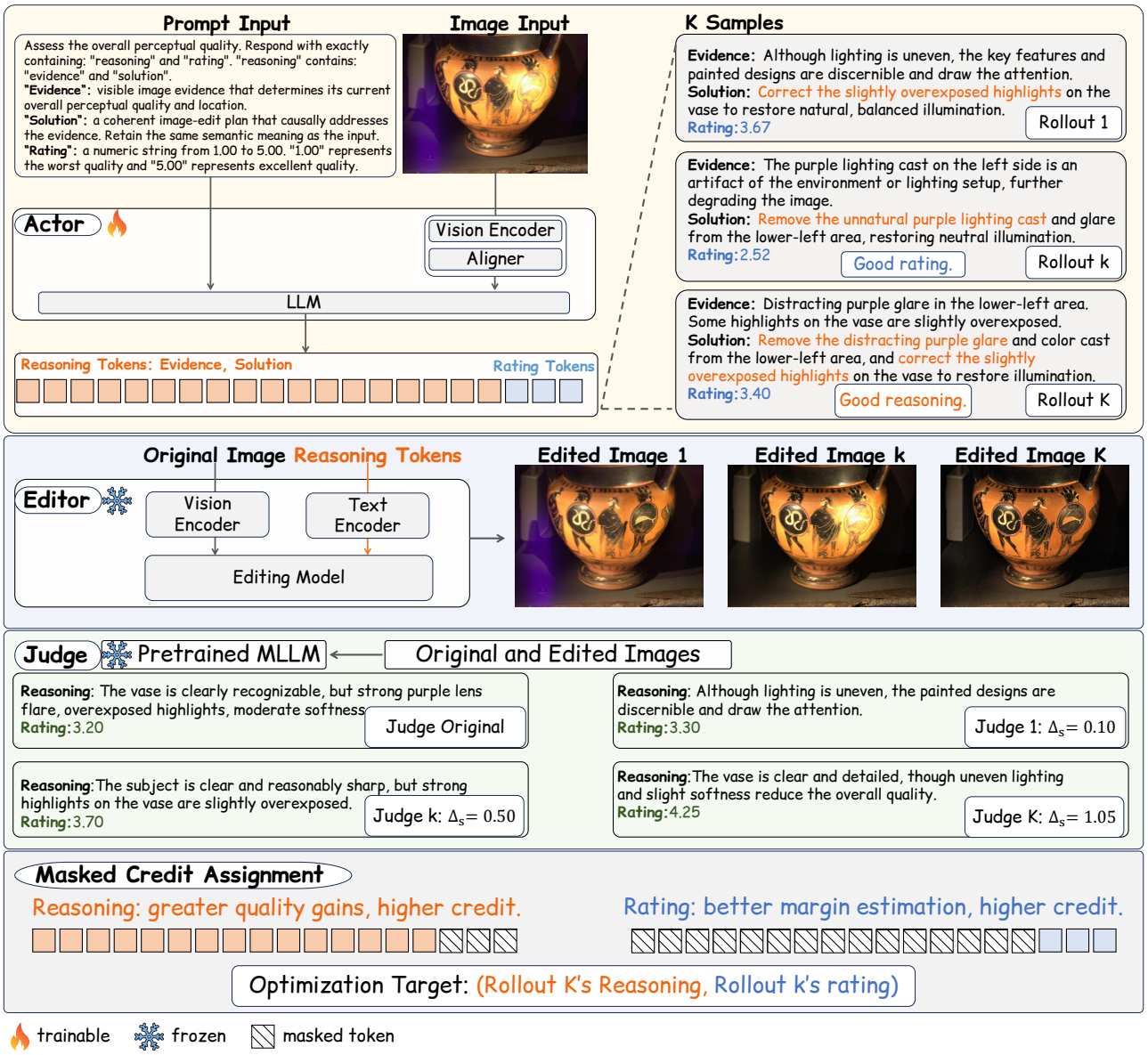

Figure 1: Overview of MR-IQA-2. Prior MLLM-based IQA produces an explanation and a rating without visually verifying the explanation. Our actor converts a quality hypothesis into multiple targeted low-level edits. The independent judge compares the resulting quality changes and returns evidence-based feedback. Only visually supported interventions reinforce the actor, yielding faithful reasoning and a deeper understanding of causal quality-limiting factors. Here, Δs denotes the Judge-score change between an edited image and the original image, as defined in Eq. (5).

$i \in \{1, \ldots, N\}$ indexes an image, K denotes the number of Actor samples per image, and $k \in \{1, \ldots, K\}$ indexes one sample. Within each training trajectory, superscript $b \in \{0, 1\}$ denotes the original-image and post-editing stages, respectively.

Actor receives an original image I_i^0 and uses a prompt P_A to generate quality reasoning and a rating:

$$a_{i,k}^0 = \left(r_{i,k}^0, q_{i,k}^0\right) \sim \pi_{\text{old}}\left(\cdot \mid I_i^0, P_A\right), \tag{1}$$

where π_{old} denotes the original Actor policy used to generate the rollouts, and $r_{i,k}^0$ and $q_{i,k}^0$ denote reasoning and rating.

Editor corrects the quality factors identified by the Actor:

$$I_{i,k}^1 = E_{\phi_E}\left(I_i^0, r_{i,k}^0; P_E\right), \qquad \phi_E \text{ fixed.} \tag{2}$$

where P_E is a fixed editing prompt template and $I_{i,k}^1$ is the edited output.

Judge independently rates the original and edited images:

$$s_i^0 = J_{\phi_J}\left(I_i^0\right), \qquad s_{i,k}^1 = J_{\phi_J}\left(I_{i,k}^1\right). \tag{3}$$

The resulting trajectory is

$$\mathcal{T}_{i,k} = \left(I_i^0, a_{i,k}^0, I_{i,k}^1, s_i^0, s_{i,k}^1\right). \tag{4}$$

The Judge-observed quality change is

$$\Delta s_{i,k} = s^1_{i,k} - s^0_i. \tag{5}$$

3.3 Fine-Grained Credit Assignment

Motivation. As discussed in the Introduction, higher rating accuracy alone does not guarantee faithful reasoning. Moreover, assigning a shared reward to reasoning and rating obscures the source of supervision. We therefore define separate rewards for reasoning, rating, and format, and use each reward to update its corresponding output tokens.

Reasoning reward. For each image, the Actor samples K reasoning–rating outputs indexed by k. Each reasoning proposal is evaluated with the frozen Editor and Judge. The editing prompt restricts interventions to low-level quality attributes while preserving image content. We measure reasoning faithfulness by the resulting Judge-score improvement:

$$R^{\text{reasoning}}_{i,k} = \Delta s_{i,k}. \tag{6}$$

Under these controlled conditions, reasoning that produces a larger quality improvement receives a higher reward.

Rating reward. Following MR-IQA (Li et al. 2026), we supervise rating through relative quality margins. Let y_i be the human mean opinion score (MOS). For another image $j \neq i$ in the same batch, the scale-controlled margin error is

$$z^{\text{rating}}_{i,k,j} = \frac{\left(q^0_{i,k} - \frac{1}{K}\sum_{k'=1}^{K} q^0_{j,k'}\right) - (y_i - y_j)}{\tau_{ij}}, \tag{7}$$

where $\tau_{ij} > 0$ controls the margin-error scale. Using the L2 estimator, we average the $N-1$ pairwise rewards:

$$R^{\text{rating}}_{i,k} = \frac{1}{N-1}\sum_{\substack{j=1 \\ j\neq i}}^{N} \mathrm{e}^{-\frac{1}{2}\left(z^{\text{rating}}_{i,k,j}\right)^2}. \tag{8}$$

Format reward. The Actor output must be a valid JSON object containing the ordered fields **reasoning** and **rating**. Let $\mathcal{F}$ denote this output contract. The format reward is

$$R^{\text{format}}_{i,k} = \mathbf{1}\left[a^0_{i,k} \in \mathcal{F}\right]. \tag{9}$$

Masked credit assignment. Let c index a supervision type. For token t, $\chi^c_{i,k,t}$ is its binary mask, and $\Omega^c_{i,k}$ is the selected token set:

$$\begin{aligned} \mathcal{C} &= \{\text{reasoning}, \text{rating}, \text{format}\}, \\ \Omega^c_{i,k} &= \left\{t \,\middle|\, \chi^c_{i,k,t} = 1\right\}, \qquad c \in \mathcal{C}. \end{aligned} \tag{10}$$

Each reward is normalized independently within the K samples of image i:

$$A^c_{i,k} = \frac{R^c_{i,k} - \mu^c_i}{\sigma^c_i + \epsilon_A}. \tag{11}$$

Here, μ^c_i and σ^c_i denote the corresponding group mean and sample standard deviation. The masked advantage for each supervision type is

$$\widetilde{A}^c_{i,k,t} = \chi^c_{i,k,t} A^c_{i,k}. \tag{12}$$

Thus, reasoning and rating rewards update only their corresponding fields, whereas the format reward applies to the complete output.

3.4 Masked Credit with GRPO

We optimize the Actor with Group Relative Policy Optimization (GRPO) (Shao et al. 2024). Let $\mathcal{I}_c$ contain the valid sampled outputs for supervision type c. Using the masked advantage in Eq. (12), the masked GRPO loss is

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = -\sum_{c\in\mathcal{C}} \frac{1}{|\mathcal{I}_c|} \sum_{(i,k)\in\mathcal{I}_c} \frac{1}{|\Omega^c_{i,k}|} \sum_{t\in\Omega^c_{i,k}} \omega_{i,k,t} \\ \times \min\left(\rho_{i,k,t}\widetilde{A}^c_{i,k,t}, \rho^{\text{clip}}_{i,k,t}\widetilde{A}^c_{i,k,t}\right). \end{aligned} \tag{13}$$

Here, $\rho_{i,k,t}$ is the likelihood ratio between the current Actor π_θ and the rollout policy π_{old}. $\rho^{\text{clip}}_{i,k,t}$ clips this ratio to $[1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}]$. $\omega_{i,k,t}$ corrects the token-level mismatch between the rollout distribution and π_{old} and is capped by $\omega_{\max}$.

4 Experiments

4.1 Experimental Settings

Datasets. We train all controlled models on the training subset of the KonIQ-10k split (Hosu et al. 2020), which contains $7{,}046$ in-the-wild images at 512×384 resolution. We evaluate on the KonIQ-10k test split (2,010 images) as the in-distribution authentic benchmark. For out-of-distribution (OOD) evaluation, we use the authentic distortion datasets SPAQ (11,125 images) (Fang et al. 2020) and LIVE In the Wild (LIVE-W; 1,162 images) (Ghadiyaram and Bovik 2015), as well as the AI-generated image quality dataset AGIQA-3K (2,982 images) (Li et al. 2024). We further include the synthetic distortion datasets KADID-10k (10,125 images) (Lin, Hosu, and Saupe 2019) and CSIQ (866 images) (Larson and Chandler 2010).

Model Backbones. The **Actor** is initialized from Qwen3.5-4B (Qwen Team 2026), selected for its balance of visual reasoning and efficiency. The **Editor** uses FLUX.2 [klein] 4B (Black Forest Labs 2025) with a four-step LoRA for fast and efficient editing. It runs in BF16 with classifier-free guidance 1.0 and sigma schedule $[1.0, 0.75, 0.5, 0.25]$. The **Judge** uses a Qwen3.5-4B checkpoint trained for 5 epochs on KonIQ-7k with rating supervision, selected for the same balance of visual reasoning and efficiency. On a single NVIDIA RTX A6000, the Editor's mean inference time is 0.902 s per image, while Judge inference takes 0.800 s per image. During GRPO optimization, only the Actor is updated; the frozen Editor and Judge provide offline supervision and receive no gradients.

Implementation Details. We train Qwen3.5-4B (Qwen Team 2026) for five epochs on eight 48-GB NVIDIA RTX A6000 GPUs. Each rank processes $N = 6$ images and samples $K = 6$ completions per image. The maximum completion length is 192 tokens. We use AdamW (Loshchilov and Hutter 2019) with $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon_{\text{Adam}} = 10^{-8}$, a learning rate of 10^{-6}, weight decay of 0.1, cosine scheduling without warmup, and gradient-norm clipping at 1.0. We use symmetric policy clipping with $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.20, 0.20)$, no hard clipping of advantages, and token-level importance-ratio truncation at $\omega_{\max} = 2$. Rollouts use temperature 0.7,

	Authentic						AI-generated		Synthetic				Average	
	KonIQ		SPAQ		LIVE-W		AGIQA-3K		KADID-10k		CSIQ		Avg.	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Hand-crafted														
NIQE (2012)	0.533	0.530	0.679	0.664	0.493	0.449	0.560	0.533	0.468	0.405	0.718	0.628	0.575	0.535
BRISQUE (2012)	0.225	0.226	0.490	0.406	0.361	0.313	0.541	0.497	0.429	0.356	0.740	0.556	0.464	0.392
Deep-learning-based														
NIMA (2018)	0.896	0.859	0.838	0.856	0.814	0.771	0.715	0.654	0.532	0.535	0.695	0.649	0.748	0.721
DBCNN (2020)	0.884	0.875	0.812	0.806	0.773	0.730	0.641	0.648	0.497	0.484	0.586	0.572	0.699	0.686
MUSIQ (2021)	0.924	0.929	0.868	0.863	0.789	0.830	0.722	0.630	0.575	0.556	0.771	0.710	0.775	0.753
MANIQA (2022)	0.849	0.834	0.768	0.758	0.849	0.832	0.723	0.636	0.499	0.465	0.623	0.627	0.719	0.692
CLIP-IQA+ (2023)	0.909	0.895	0.866	0.864	0.832	0.805	0.736	0.685	0.653	0.654	0.772	0.719	0.795	0.770
MLLM-based: SFT training														
C2Score (2024)	0.923	0.910	0.867	0.860	0.786	0.772	0.777	0.671	0.500	0.453	0.735	0.705	0.765	0.729
Q-Align (2023)	0.941	0.940	0.886	0.887	0.853	0.860	0.772	0.735	0.674	0.684	0.671	0.737	0.800	0.807
DeQA (2025)	0.953	0.941	0.895	0.896	0.892	0.879	0.809	0.729	0.694	0.687	0.787	0.744	0.838	0.813
MLLM-based: RL training														
Q-Insight (2025a)	0.918	0.895	0.903	0.903	0.870	0.839	0.816	0.766	0.702	0.702	0.685	0.640	0.816	0.791
VQ-R1 (2025)	0.886	0.919	0.867	0.887	0.817	0.869	0.744	0.718	0.635	0.640	0.709	0.721	0.776	0.792
MR-IQA (2026)	0.949	0.931	0.892	0.897	0.899	0.883	0.804	0.732	0.672	0.683	0.767	0.732	0.831	0.810
MLLM-based: Tool-augmented training														
Zoom-IQA† (2026)	0.938	0.922	0.902	0.900	0.887	0.870	0.816	0.765	0.701	0.700	0.797	0.754	0.840	0.819
MR-IQA-2 (Ours)	0.937	0.917	0.900	0.899	0.893	0.863	0.809	0.739	0.667	0.669	0.824	0.785	0.838	0.812
MR-IQA-2* (Ours)	0.925	0.904	0.901	0.900	0.881	0.847	0.826	0.768	0.665	0.676	0.840	0.801	0.840	0.816

Table 1: **Rating performance comparison.** Each dataset reports PLCC↑ and SRCC↑ between predicted and ground-truth ratings. Red and blue denote the best and second-best results, respectively. Baseline entries use reported results, except that VQ-R1 is reproduced with Qwen3-VL-2B (Bai et al. 2025a) because its original results were obtained under a different training protocol. Both MR-IQA-2 variants use Qwen3.5-4B; the unstarred row uses the final E5 credit-mask checkpoint, while MR-IQA-2* freezes the vision encoder and aligner during training. Zoom-IQA† uses additional annotations and combines SFT with RL; all other trainable methods use the same KonIQ training split.

	Pretrained Judge		**Q-Insight as Judge**		**GPT-5.6 Sol as Judge (%)**
Actor \ **Editor**	**FLUX.2 [klein]**	**Mage-Flow**	**FLUX.2 [klein]**	**Mage-Flow**	**FLUX.2 [klein]**
Qwen3.5-4B (Baseline)	−0.137	−0.144	−0.130	−0.080	4.19%
Q-Insight	+0.213	+0.142	+0.138	+0.139	14.41%
MR-IQA-2 (Ours)	**+0.789**	**+0.468**	**+0.561**	**+0.376**	**81.40%**

Table 2: **Reasoning performance validation.** The first four columns report mean quality gain Δs (higher is better; Eq. (5)); GPT-5.6 Sol uses 3-Alternative Forced Choice (3AFC). We compare the Qwen3.5-4B baseline (Qwen Team 2026), Q-Insight (Li et al. 2025a), and MR-IQA-2 (ours). The 4B four-step LoRA Editors are FLUX.2 [klein] (Black Forest Labs 2025) and Mage-Flow (Zhang et al. 2026). Judges are our pretrained Qwen3.5-4B Judge (Qwen Team 2026), Q-Insight (Li et al. 2025a), and GPT-5.6 Sol (OpenAI 2026). Evaluation uses a fixed 5,866-image subset formed by randomly selecting 1,000 images from each test set, except CSIQ, for which all 866 images are used.

top-p 1.0, top-k 20, and a presence penalty of 1.5. Actor-only training takes approximately 2 hours per epoch. For the full MR-IQA-2 framework, four GPUs are allocated to Actor training and four to Editor–Judge inference, increasing the per-epoch time to approximately 12 hours.

4.2 Rating Performance

Compared methods. Table 1 compares MR-IQA-2 with representative BIQA methods across six benchmarks. The hand-crafted group includes NIQE (Mittal, Soundararajan, and Bovik 2012) and BRISQUE (Mittal, Moorthy, and Bovik 2012). Deep-learning-based methods include NIMA (Talebi and Milanfar 2018), DBCNN (Zhang et al. 2020), MUSIQ (Ke et al. 2021), MANIQA (Yang et al. 2022), and CLIP-IQA+ (Wang, Chan, and Loy 2023). Among MLLM-based methods, SFT approaches include C2Score (Zhu et al. 2024), Q-Align (Wu et al. 2023), and DeQA (You et al. 2025), while RL approaches include Q-Insight (Li et al. 2025a), VQ-R1 (Wu et al. 2025), and MR-IQA (Li et al. 2026). We further compare with the

tool-augmented RL method Zoom-IQA (Liang et al. 2026). For MR-IQA-2, we report an active-vision checkpoint and a frozen-vision variant, denoted by *.

Efficiency and performance. MR-IQA-2 is designed primarily to improve reasoning faithfulness rather than maximize rating. Nevertheless, it achieves competitive rating performance. **(1)** Zoom-IQA (Liang et al. 2026) uses a two-stage SFT–RL pipeline with additional annotations, whereas MR-IQA-2 uses only RL training without additional data. The frozen-vision variant reaches an average PLCC/SRCC of 0.840/0.816, comparable to Zoom-IQA's 0.840/0.819. **(2)** Freezing the vision encoder and aligner is particularly effective on AI-generated and synthetic datasets. The frozen variant achieves the best CSIQ PLCC/SRCC of 0.840/0.801, whereas active visual training yields its main gains on the authentic datasets.

4.3 Faithful Reasoning Performance

Quality Gain as a Proxy for Reasoning Faithfulness. Prior work, including Q-Insight (Li et al. 2025a) and VisualQuality-R1 (Wu et al. 2025), typically evaluates reasoning through human inspection of a small number of examples or indirectly through improvements in rating performance. These practices lack a unified evaluation criterion. In this work, we instead use the image-quality gain produced by reasoning-guided editing as the evaluation criterion (Eq. (6)), rather than rating performance. We assume that faithful reasoning should identify quality-limiting factors whose correction improves the given image.

Cross-editor/judge stability. Table 2 examines whether the Actor overfits to the frozen Editor–Judge during training. We first replace the FLUX.2 [klein] Editor (Black Forest Labs 2025) with Mage-Flow (Zhang et al. 2026). MR-IQA-2 retains larger quality gains than the competing Actors, indicating that its reasoning is not specific to a single Editor. We then replace the pretrained Judge with Q-Insight (Li et al. 2025a) and GPT-5.6 Sol (OpenAI 2026). These alternative Judges preserve the direction of the measured gains. These results indicate that the Actor learns generalizable quality knowledge rather than exploiting idiosyncrasies of the training-time Editor or Judge.

Quality and behavior analysis. Additional analyses (in Appendix Figure 4) provide four observations. (a) Across all datasets, lower-quality images obtain larger quality gains. (b) Changes in sharpness are strongly associated with Judge preference. (c) Higher-quality images tend to remain closer to their originals after editing. (d) Training shifts the Actor toward quality-relevant reasoning tokens.

4.4 Ablation Study

Evaluation. Table 3 reports ablations on rating alignment and reasoning-guided quality improvement. Rating-only training already achieves competitive rating alignment, with an average PLCC/SRCC of $0.836/0.815$. However, its mean quality gain is -0.260, below the baseline result of -0.136. Rating performance alone therefore does not guarantee faithful reasoning. At E5, the variants without and with the credit mask reach average PLCC/SRCC values of $0.834/0.812$ and $0.838/0.812$, respectively. Both produce positive quality gains. The variant without the mask reaches a larger raw gain of $+1.442$, but collapses to one normalized solution. The credit-mask variant retains 23,457 normalized solutions while achieving a gain of $+1.079$. Therefore, raw gain alone can favor a universal edit and does not establish image-conditioned reasoning. The final E5 comparison is diagnostic because the two variants also differ in KL scope; controlled E1 results are reported in Appendix E.

The credit mask is more beneficial early in training. At step 30 on the 200-image validation set, it reaches a PLCC/SRCC of $0.745/0.772$, compared with $0.667/0.617$ without the credit mask and $0.616/0.610$ for the baseline. Its quality gain is lower at this stage, at 0.255 versus 0.387 without the credit mask. This suggests that the credit mask accelerates rating convergence but may temporarily limit reasoning improvement. Once rating alignment stabilizes, its benefit becomes smaller. This behavior warrants further study of the imbalance between rating and reasoning rewards.

4.5 Case Study

Figure 2 compares reasoning-guided edits before and after training. After training, all three edits receive higher Judge ratings, and the Actor proposes more specific interventions. In the dog example, it recommends cropping the distracting foot near the image boundary to improve visual appeal. The examples also expose a semantic-preservation problem: an edited image may no longer retain the meaning of the original. In the first row, the Editor changes a night sky into a blue daytime sky. The boundary between quality enhancement and semantic alteration therefore requires further study.

5 Discussion

Good Evidence, Bad Solution? In this work, we jointly treat `evidence` and `solution` as reasoning. However, as with human perception, a model may correctly identify why an image has poor quality yet fail to propose an effective correction. Our training necessarily encourages the model to discover effective solutions through exploration, raising the question of whether this process also improves its evidence. Maybe the evidence capability does not improve. But we argue that a solution that improves image quality through relevant visual changes indicates an improved understanding of the underlying quality factors. Explicitly modeling the causal relation between evidence and solution may provide a more reliable direction for future work.

6 Limitations and Future Work

This work has two main limitations. First, the current framework does not yet realize fully closed-loop visual reflection. The edited image is evaluated by the Judge but is not returned to the Actor for further visual reasoning. Second, using separate Actor, Editor, and Judge models introduces computational and parameter redundancy. Future work will close the reflection loop by feeding the edited image back to the Actor, enabling direct comparison between the original

Rating alignment

	Authentic						AI-generated		Synthetic				Average	
	KonIQ		SPAQ		LIVE-W		AGIQA-3K		KADID-10k		CSIQ		Avg.	
Method	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Baseline	0.553	0.583	0.750	0.743	0.592	0.564	0.679	0.655	0.660	0.669	0.681	0.657	0.652	0.645
Rating-only	**0.952**	**0.938**	**0.901**	0.899	**0.896**	**0.875**	0.808	**0.740**	0.669	0.676	0.789	0.764	0.836	**0.815**
Without credit mask (E5)	0.932	0.914	0.896	0.897	0.878	0.857	0.809	0.736	**0.676**	**0.684**	0.815	**0.786**	0.834	0.812
Credit mask (E5)	0.937	0.917	0.900	**0.899**	0.893	0.863	**0.809**	0.739	0.667	0.669	**0.824**	0.785	**0.838**	0.812

Subset quality gains

Method	**Original mean**	**Edited mean**	**Mean gain**	**Positive gain (%)**
Baseline	2.844	2.708	−0.136	40.664%
Rating-only	2.844	2.584	−0.260	30.686%
Without credit mask (E5)	2.844	**4.286**	**+1.442**	**99.997%**
Credit mask (E5)	2.844	3.923	+1.079	98.813%

Table 3: **Rating and quality-gain ablations.** The upper panel reports PLCC/SRCC on the complete test sets and their six-dataset average. The lower panel reports Judge scores before and after reasoning-guided editing. Baseline and Rating-only use 5,843 common images; the E5 rows use their full successful audits (28,270 without the mask and 28,044 with the mask). Bold marks the best scalar result before rounding. The higher unmasked gain coincides with single-solution collapse and is not interpreted as improved reasoning.

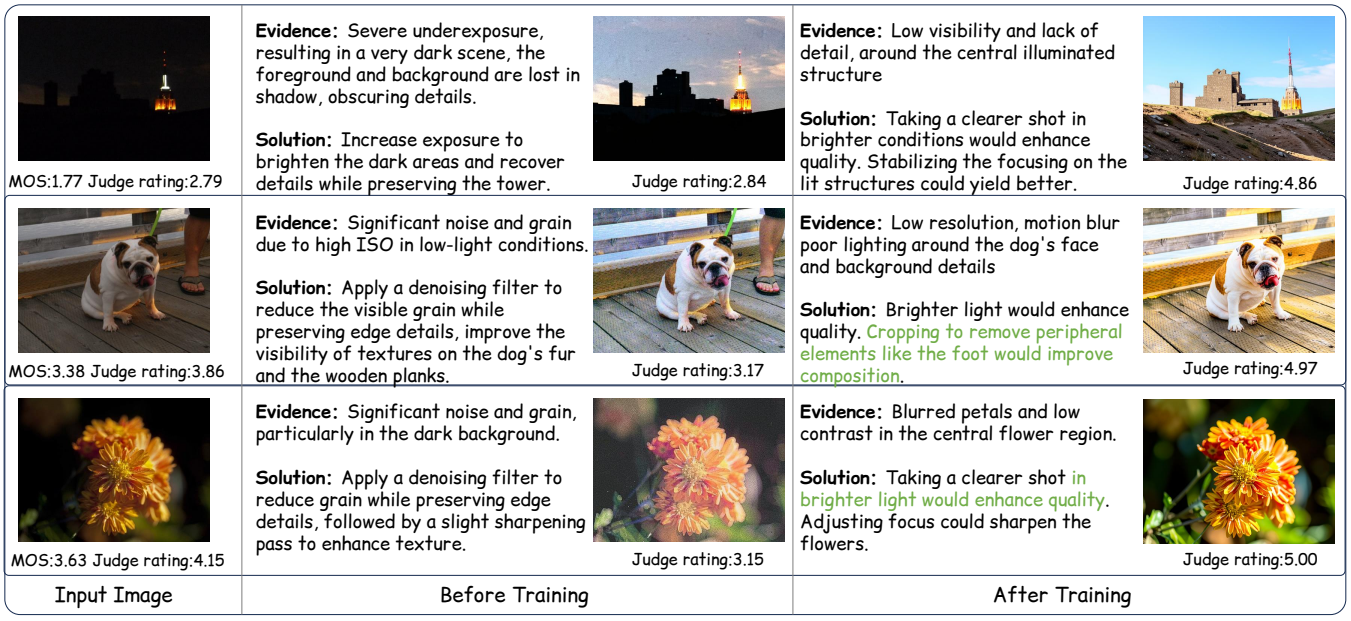

Figure 2: **Qualitative comparison before and after training.** Each row shows the input image and the Actor's reasoning-guided edits before and after training. Post-training reasoning yields more effective interventions and larger Judge-rated quality gains. Green text highlights additional quality factors identified after training.

and edited images. We will also explore smaller models and parameter sharing across modules to improve efficiency.

7 Conclusion

In this work, we aim to supervise the faithfulness of image-quality reasoning, moving BIQA beyond plausible explanations toward visually verifiable understanding. We use quality gain as an operational measure of reasoning faithfulness and propose MR-IQA-2, an Actor–Editor–Judge framework. The Actor's reasoning guides the Editor, while the Judge evaluates the resulting quality change to test whether the identified factors are supported. Fine-grained credit assignment further decouples reasoning and rating supervision to reduce reward ambiguity. MR-IQA-2 retains competitive rating alignment while producing more reliable and actionable quality reasoning. More broadly, our framework provides a new perspective on evaluating reasoning faithfulness, links blind assessment to reference-based verification through intervention-generated image pairs, and suggests a path toward modeling

human visual preferences and aesthetic judgments.

References

Agnolucci, L.; Galteri, L.; Bertini, M.; and Del Bimbo, A. 2024. ARNIQA: Learning Distortion Manifold for Image Quality Assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 189–198.

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

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

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024. TOPIQ: A Top-Down Approach from Semantics to Distortions for Image Quality Assessment. *IEEE Transactions on Image Processing*, 33: 2404–2418.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3677–3686.

Ghadiyaram, D.; and Bovik, A. C. 2015. Live in the wild image quality challenge database. *Online: http://live. ece. utexas. edu/research/ChallengeDB/index. html [Mar, 2017]*.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment. *IEEE Transactions on Image Processing*, 29: 4041–4056.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1): 011006–011006.

Li, C.; Zhang, Z.; Wu, H.; Sun, W.; Min, X.; Liu, X.; Zhai, G.; and Lin, W. 2024. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 6833–6846.

Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; and Zhang, J. 2025a. Q-Insight: Understanding Image Quality via Visual Reinforcement Learning. *arXiv preprint arXiv:2503.22679*.

Li, Y.; Lin, Y.; Sun, Z.; Yang, Y.-H.; Miyoshi, K.; Chu, C.; and Nishida, S. 2026. MR-IQA: A Unified Margin View of Regression and Ranking for Blind Image Quality Assessment. arXiv:2606.29760.

Li, Y.; Sun, Z.; Chen, Y.-j.; and Nishida, S. 2025b. Building Reasonable Inference for Vision-Language Models in Blind Image Quality Assessment. In *International Conference on Neural Information Processing*, 283–295. Springer.

Li, Y.; Yu, Y.; Lin, Y.; Yang, Y.-H.; Chu, C.; and Nishida, S. 2025c. Guiding Perception-Reasoning Closer to Human in Blind Image Quality Assessment. *arXiv preprint arXiv:2512.16484*.

Liang, G.; Wang, J.; Wu, Z.; Zhou, S.; and Loy, C. C. 2026. Zoom-IQA: Image Quality Assessment with Reliable Region-Aware Reasoning. *arXiv preprint arXiv:2601.02918*.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. IEEE.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

OpenAI. 2026. GPT-5.6 Sol Model. https://developers.openai.com/api/docs/models/gpt-5.6-sol.

Qin, G.; Zhang, J.; He, C.; Fu, Y.; Liang, J.; Wu, T.; and Zhang, L. 2026. Tool-IQA: Augmenting Image Quality Assessment with Simple Tools. *arXiv preprint arXiv:2606.16082*.

Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Székely, G. J.; Rizzo, M. L.; and Bakirov, N. K. 2007. Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35(6): 2769–2794.

Talebi, H.; and Milanfar, P. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing*, 27(8): 3998–4011.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 2555–2563.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multi-Scale Structural Similarity for Image Quality Assessment. In *Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers*, volume 2, 1398–1402.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024. Q-Instruct: Improving Low-Level Visual Abilities for Multi-Modality Foundation Models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25490–25500.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Wu, T.; Zou, J.; Liang, J.; Zhang, L.; and Ma, K. 2025. VisualQuality-R1: Reasoning-Induced Image Quality Assessment via Reinforcement Learning to Rank. *Advances in Neural Information Processing Systems*, 38: 88167–88190.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1191–1200.

You, Z.; Cai, X.; Gu, J.; Xue, T.; and Dong, C. 2025. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14483–14494.

You, Z.; Li, Z.; Gu, J.; Yin, Z.; Xue, T.; and Dong, C. 2024. Depicting beyond scores: Advancing image quality assessment through multi-modal language models. In *European Conference on Computer Vision*, 259–276. Springer.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; et al. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Advances in Neural Information Processing Systems*, volume 38.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020. Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.

Zhang, W.; Zhai, G.; Wei, Y.; Yang, X.; and Ma, K. 2023. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14071–14081.

Zhang, X.; Zhang, P.; Zheng, S.; Guo, J.; Jia, Z.; Shen, Y.; Guo, X.; Luo, Y.; Li, J.; Xie, W.; Pu, F.; Zhang, X.; Zhang, K.; Guo, Z.; Bi, T.; Gui, D.; Liu, Z.; Wen, Z.; Zheng, Z.; Yang, S.; Li, X.; Wang, J.; Li, B.; and Lu, Y. 2026. Mage-Flow: An Efficient Native-Resolution Foundation Model for Image Generation and Editing. *arXiv preprint arXiv:2607.19064*.

Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; and Wang, S. 2024. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37: 32611–32629.

In this appendix, we provide technical details and more comprehensive analyses of our proposed method. It comprises Section A, *Backbone, Algorithm and Hyperparameters*; Section B, *Visual Quality Feature Analysis*; Section C, *Discussion on the Editor Proxy*; Section D, *Joint Rating, Reasoning, and Diversity Analysis*; and Section E, *Detailed Credit-Mask Audit*.

A Backbone, Algorithm and Hyperparameters

A.1 Dataset Protocol.

Our training and evaluation settings follow the same configuration as the main paper. We train on the KonIQ-10k training subset (Hosu et al. 2020) and use a fixed set of 200 images randomly sampled from the KonIQ-10k test split to monitor early-stage convergence. Final testing is conducted on the KonIQ-10k test split, SPAQ (Fang et al. 2020), LIVE-W (Ghadiyaram and Bovik 2015), AGIQA-3K (Li et al. 2024), KADID-10k (Lin, Hosu, and Saupe 2019), and CSIQ (Larson and Chandler 2010).

At the time of manuscript completion, we did not have access to Zoom-IQA (Liang et al. 2026) and therefore could not reproduce it under our protocol. Tool-IQA (Qin et al. 2026) follows a different training–testing configuration; consequently, we cannot report directly comparable Tool-IQA results for some experiments.

For the cross Actor–Judge–Editor evaluation reported in Table 2 of the main paper, we use a fixed 5,866-image subset, randomly sampling 1,000 images from each test set except CSIQ, for which all 866 images are used. This subsampling is necessary for computational efficiency: evaluating the full test sets would involve approximately 23,000 images, each requiring four editing operations and three Judge evaluations, resulting in a prohibitively heavy inference cost.

A.2 Candidate Backbones.

The Qwen series has been widely adopted in recent BIQA studies due to its open-source availability, strong multimodal capability, and computational efficiency. For example, existing methods such as Q-Insight (Li et al. 2025a), VQ-R1 (Wu et al. 2025), and Zoom-IQA (Liang et al. 2026) employ Qwen2.5-VL-7B (Bai et al. 2025b) as their backbone model. However, despite its strong performance, the 7B-scale model still introduces considerable computational overhead, which limits its practical deployment and training efficiency.

Therefore, in this work, we investigate more lightweight backbone configurations with 2B and 4B parameters. Considering future scalability and the continuous evolution of multimodal large language models, we further evaluate the latest Qwen3-VL and Qwen3.5-VL architectures. We conduct comprehensive backbone comparisons among 2B/4B variants of Qwen3 and Qwen3.5 to identify an optimal trade-off between training efficiency and performance.

Qwen3-VL versus Qwen3.5. The T1–T2 comparison in Table 4 shows that, under matched frozen-vision DAPO settings, Qwen3.5-2B outperforms Qwen3-VL-2B on five of six datasets for both PLCC and SRCC, and on all six in MAE. It also requires only $0.81\times$ the training time per epoch.

2B versus 4B. The T6–T7 comparison in Table 4 shows that the 4B model converges faster in rating performance. By Epoch 2, T7 reaches a validation PLCC/SRCC/MAE of 0.929/0.919/0.566, compared with 0.915/0.904/0.744 for the 2B T6; by Epoch 3, T7 further improves to 0.935/0.924/0.228. The 4B model also produces more diverse outputs, with 83 distinct ratings, 200 unique completions, and a 41.5% U-score, compared with a 39.0% U-score for T6. We further observe that its response lengths are better aligned with our objective of eliciting detailed quality reasoning.

A.3 Optimization Algorithms: GRPO or DAPO?

Compared with GRPO (Shao et al. 2024), DAPO (Yu et al. 2025) dynamically resamples low-variance groups, thereby increasing within-group reward diversity and preventing policy advantages from vanishing prematurely. We therefore adopted DAPO in most of our early experiments.

Across three epochs, 3,282 of 20,832 GRPO learner groups (15.75%) have zero within-group reward variance and therefore zero policy advantage. DAPO resampling reduces the proportion of zero-advantage learner groups to 0%. Comparing T2 and T4 in Table 4, DAPO improves the six-dataset average PLCC/SRCC by 0.006/0.008 and reduces MAE by 0.138 at Epoch 3, with improvements on five, six, and six datasets, respectively.

Despite its better performance, DAPO requires $1.866\times$ more sampled trajectories and increases the training time per epoch from 0.81 to 1.13 hours (approximately 39.5%). This additional training-time cost is not affordable in our training environment, especially when scaling the model to 4B parameters or beyond. Therefore, we use GRPO as the default optimization algorithm in the final framework.

A.4 Hyper-Parameter Settings.

Group Size: 6 versus 48. MR-IQA (Li et al. 2026) reports its best performance with local groups of six, where rewards are computed by comparing samples only within each six-image group. Table 4 compares this local-6 setting (T6) with a larger group of 48 (T5). The large group achieves better in-domain validation PLCC/SRCC/MAE (0.929/0.921/0.261 versus 0.917/0.906/0.587), whereas local-6 achieves higher six-dataset average PLCC/SRCC (0.835/0.813 versus 0.828/0.809). Because the local-6 setting provides stronger cross-dataset correlation, we adopt a group size of six.

Vision Modules: Frozen versus Active. The T2–T5 comparison in Table 4 shows that active visual training performs better on authentic datasets, achieving PLCC/SRCC values of 0.952/0.938 on KonIQ and 0.897/0.877 on LIVE-W, compared with 0.925/0.904 and 0.881/0.847 for the frozen variant. In contrast, freezing the vision encoder and aligner performs better on AGIQA-3K and the synthetic datasets: it achieves 0.826/0.768 on AGIQA-3K, 0.665/0.676 on KADID-10k, and 0.840/0.801 on CSIQ, compared with 0.816/0.747, 0.661/0.668, and 0.771/0.746 under active visual training. Consequently, the frozen variant yields

Variant	Iter.	Time (h/ep.)	Val. P/S/M	U-score (%)	Gen. P/S/M
(a) Qwen3-VL-2B					
T1: DAPO i1, frozen, global	1	1.38	0.831/0.817/0.868	12.0	0.809/0.787/0.936
(b) Qwen3.5-2B					
T2: DAPO i1, frozen, global	1	1.13	0.890/0.881/0.691	36.0	**0.835**/0.808/0.740
T3: DAPO i4, frozen, global	4	2.47	0.910/0.904/0.265	21.5	0.834/0.809/0.541
T4: GRPO i1, frozen	1	0.81	0.886/0.883/0.845	34.0	0.829/0.800/0.878
T5: DAPO i1, visual, global	1	1.30	**0.929/0.921/0.261**	35.0	0.828/0.809/**0.513**
T6: DAPO i1, visual, local-six	1	1.54	0.917/0.906/0.587	39.0	**0.835**/**0.813**/0.684
(c) Qwen3.5-4B, local-six					
T7: DAPO i1, visual, local-six	1	2.56	0.935/0.924/0.228	41.5	–

Table 4: Actor-only ablations at their reported endpoints. P/S/M denotes PLCC/SRCC/MAE; Val. is the aligned 200-image set, and Gen. is the six-dataset average. U-score measures distinct validation ratings, and Time is the mean wall-clock time per epoch. Bold marks the best alignment result among the controlled Qwen3.5-2B variants; higher PLCC/SRCC and lower MAE are better. All variants use E3.

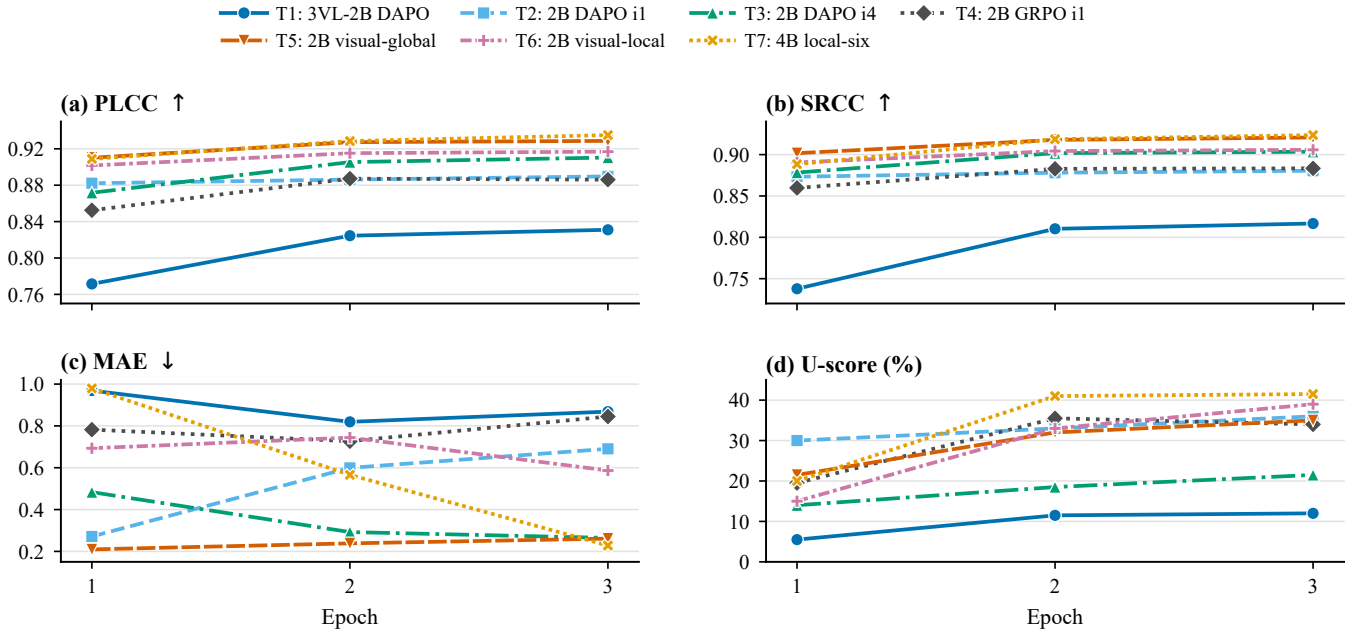

Figure 3: Validation trajectories through E3 on the common 200-image set. T7 denotes the 4B local-six run. Curves are single fixed-seed runs, so no error bands are shown. Higher PLCC/SRCC and lower MAE are better, while U-score diagnoses rating diversity. Color, marker, and line style consistently identify T1–T7.

higher six-dataset average PLCC/SRCC (0.840/0.816 versus 0.833/0.812), indicating stronger overall generalization.

Rating Diversity. Zoom-IQA (Liang et al. 2026) reports score-space collapse when using a ranking reward, with only a 2.04% unique-score ratio. Across T1–T7 in Table 4, the E3 U-scores remain between 12.0% and 41.5%, corresponding to 24–83 distinct ratings on the 200-image validation set. Thus, we do not observe rating-diversity collapse under our current settings.

Repeated Sampling Updates. MR-IQA (Li et al. 2026) reuses each rollout sample for four policy updates. We evaluate whether these repeated updates remain effective by comparing T2 and T3 in Table 4. Four updates improve validation PLCC/SRCC/MAE from 0.890/0.881/0.691 to 0.910/0.904/0.265. However, the six-dataset average PLCC/SRCC remain nearly unchanged (0.835/0.808 versus 0.834/0.809), while training time increases from 1.13 to 2.47 hours per epoch. Therefore, one policy update per rollout is sufficient under our current settings.

B Visual Quality Feature Analysis

In our framework, the FLUX.2 [klein] Editor (Black Forest Labs 2025) produces an edited image conditioned on the quality reasoning for each input. The resulting original–edited pair bridges single-image quality assessment with reference-based quality assessment, enabling us to analyze how the visual intervention relates to the predicted qual-

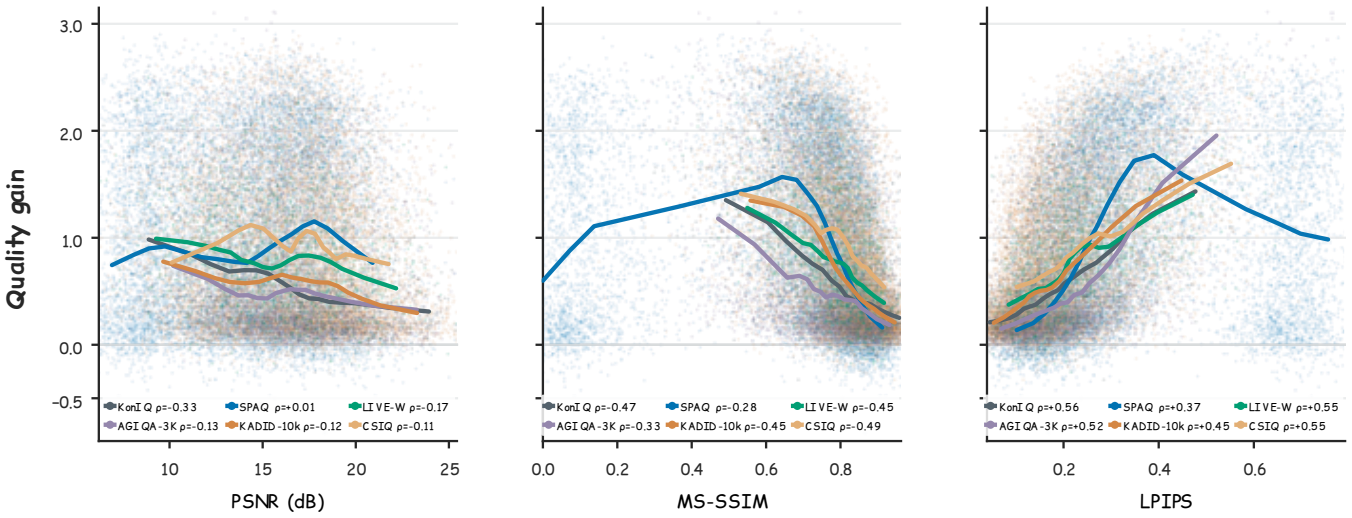

Figure 4: **Quality gain versus visual change scale.** The panels relate the Judge-observed quality gain Δs to PSNR, MS-SSIM (Wang, Simoncelli, and Bovik 2003), and LPIPS (Zhang et al. 2018) between the original and edited images. Colors denote datasets, curves show smoothed trends, and ρ denotes Spearman correlation.

ity improvement. In this section, we investigate three questions: (1) whether larger visual changes lead to greater quality gains; (2) how the distributions of low-level visual features change after editing; and (3) what behavioral preferences emerge from the overall framework.

B.1 Visual Change and Quality Gains

Visual change and quality improvement. Figure 4 relates the Judge-observed quality gain to three reference-based measures of the visual change between the original and edited images. Quality gain is generally negatively correlated with PSNR ($\rho = -0.33$ to $+0.01$) and MS-SSIM ($\rho = -0.49$ to -0.28), and positively correlated with LPIPS ($\rho = 0.37$ to 0.56). Because larger changes correspond to lower PSNR/MS-SSIM and higher LPIPS, these consistent directions indicate that larger editing changes are generally associated with greater image-quality improvements. SPAQ deviates from the other datasets across multiple measures. Its quality-gain correlation is nearly zero for PSNR ($\rho = +0.01$) and is the weakest in magnitude for both MS-SSIM ($\rho = -0.28$) and LPIPS ($\rho = +0.37$). A similar pattern appears in rating performance: while the rating correlations on the other datasets change during training, the SPAQ correlation remains close to 0.90 with only minor variation. This stability suggests that SPAQ is less sensitive to the training-induced changes observed on the other benchmarks.

B.2 Low-Level Visual Feature Patterns

Human-like Judge feedback. Figure 5 shows that the feature-dependent curves produced by the original-image Judge broadly follow the corresponding human-MOS curves across the six datasets. Their similar shapes and directions, particularly for sharpness and contrast, indicate that the Judge captures human-like quality preferences over low-level visual attributes. This agreement supports its use as a qualified source of human-like feedback for the original–edited image pairs.

Cross-dataset feature patterns. Among luminance, contrast, saturation, sharpness, colorfulness, and entropy, sharpness exhibits the clearest cross-dataset trend. Both human MOS and Judge scores generally increase with sharpness, whereas the other attributes show weaker, nonlinear, or dataset-dependent patterns. This result suggests that sharpness is a broadly shared quality cue, while no single low-level attribute fully explains perceptual quality across all datasets.

Edited-feature distributions. After editing, the images retain broad distributions across all six low-level attributes rather than collapsing to fixed feature values. The framework therefore does not appear to overfit a single low-level pattern when improving image quality. Nevertheless, Figure 5 examines each attribute marginally. The joint distribution of low-level attributes, and its relation to the human visual system and learned perceptual representations such as CLIP-IQA (Wang, Chan, and Loy 2023), warrants further investigation.

B.3 Framework Behavior Preference

Source quality, visual change, and quality gain. Figure 6(a) shows that lower-MOS images generally receive larger Judge-rated quality gains. Panel (c) further shows that lower-MOS images tend to have lower similarity between the original and edited images, indicating larger visual changes. Together with the observation in Section B.1 that larger visual changes are associated with higher quality gains, these results suggest a plausible behavior: for lower-quality inputs, the model tends to modify more visual attributes, which in turn leads to greater quality improvements. This behavior aligns with the intuition that lower-quality images provide more room for correction and supports the intended opera-

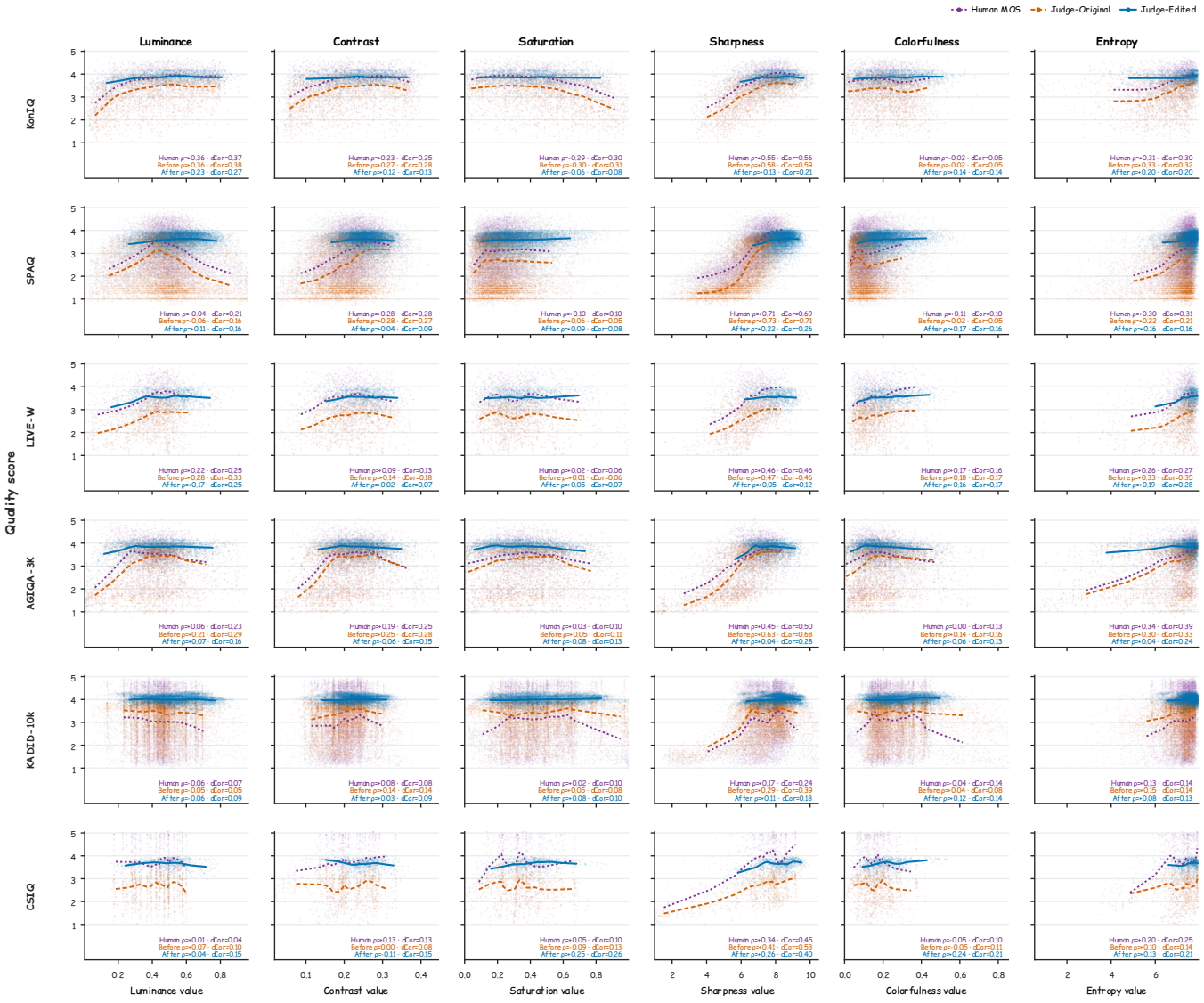

Figure 5: **Quality scores versus low-level image features.** Rows denote datasets and columns denote luminance, contrast, saturation, sharpness, colorfulness, and entropy. Purple and orange show human MOS and Judge scores for original images, while blue shows Judge scores for edited images. Curves show smoothed trends; ρ and dCor denote Spearman and distance correlations (Székely, Rizzo, and Bakirov 2007).

tion of our framework, although the observed associations do not by themselves establish causality.

Preferences over feature changes. Unlike Section B.2, which examines the absolute values of low-level attributes, Figure 6(b) relates their changes after editing to the resulting quality gain. Changes in sharpness retain the strongest and most consistent relationship with quality improvement. For entropy and saturation, the authentic and synthetic datasets form different response patterns, indicating that the preferred corrections depend on the degradation domain. SPAQ again exhibits signals that are less consistent with the other datasets, in agreement with the dataset-specific behavior observed in Section B.1.

Behavioral adaptation and concentration. Figure 6(d) reflects how the model adjusts its reasoning vocabulary in response to Judge feedback. After training, terms such as *natural* and *realistic* become prominent, revealing a learned preference for perceptually plausible corrections. At the same time, *restore* alone accounts for 20.6% of the displayed token frequency, and the top-10 token share rises to 44.1%. This concentration suggests an emerging tendency to overfit a small set of solution patterns, despite the broader behavioral alignment induced by the framework.

C Discussion on the Editor Proxy

This work does not disentangle the Editor's instruction-following capability from semantic preservation, creating

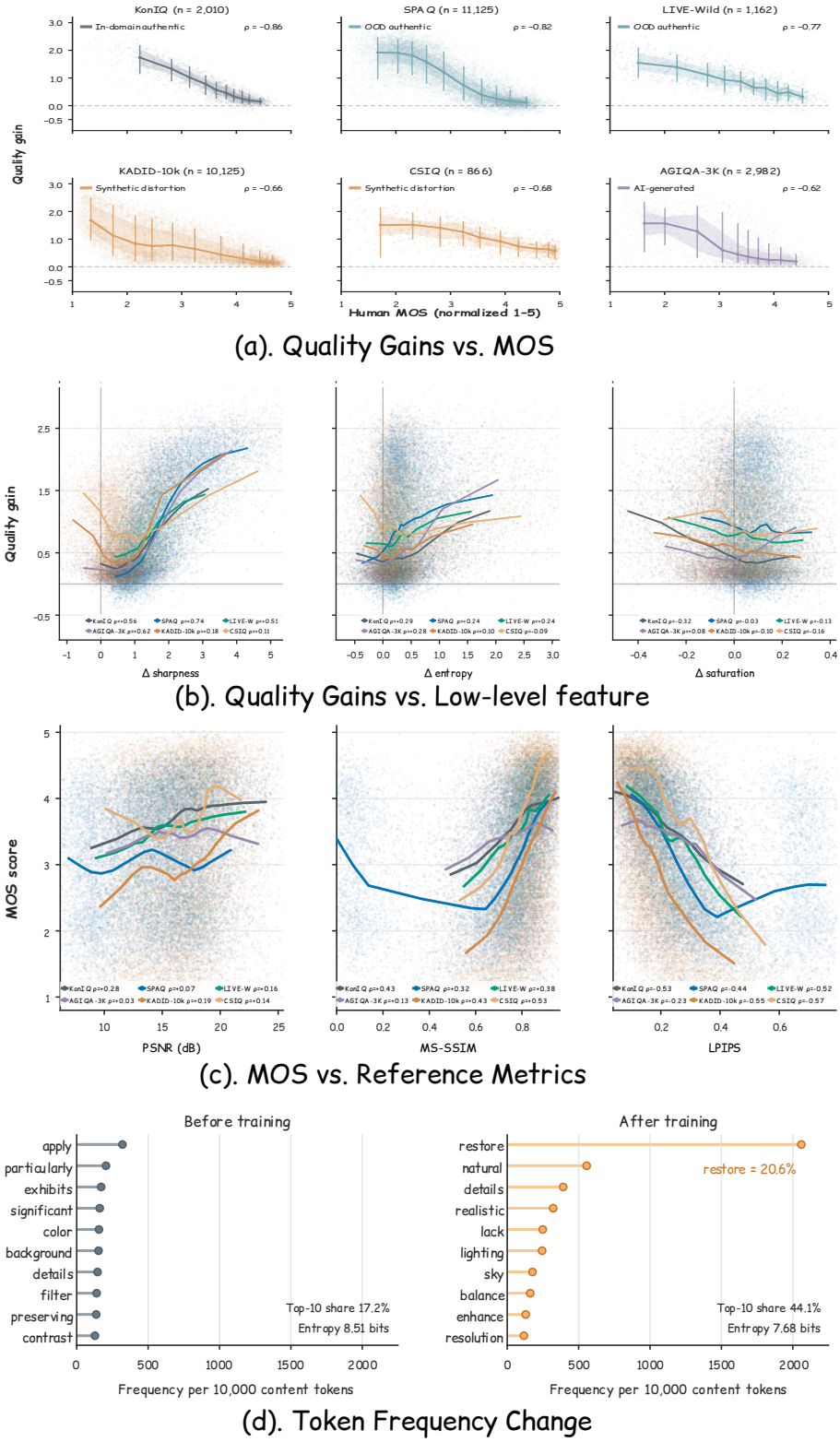

Figure 6: **Behavioral analysis of reasoning-guided editing.** (a) Judge-rated quality gain versus human MOS across six benchmarks. (b) Quality gain versus changes in low-level image features. (c) Human MOS versus reference-based similarity metrics between the original and edited images. (d) Frequent reasoning tokens before and after training.

two attribution risks. First, quality may degrade because of Editor limitations rather than an incorrect Actor instruction. Second, the edited image may become semantically inconsistent with the input, weakening quality change as a proxy for reasoning faithfulness. Future work should use independent human or human-like evaluation to assess instruction following, semantic consistency, and perceptual quality separately, yielding more reliable feedback and a more faithful framework.

D Joint Rating, Reasoning, and Diversity Analysis

Evaluation protocol. We jointly examine rating generalization, functional reasoning, and language diversity. Rating results are six-dataset averages over 28,270 test images. Functional reasoning is measured by the Judge-score gain Δs on the same 5,843 image keys for every Actor, using a fixed FLUX Editor and E5 Judge. This common-case evaluation excludes 23 Mask generations that reached the length limit. Language diversity is evaluated on 23,599 common successful inputs. Evidence and solution uniqueness are exact-string statistics, with the top-1 share measuring the concentration of the most frequent output. Results should therefore be compared within each metric and its stated sample scope.

Dataset-level rating performance. Table 5 provides the complete cross-dataset results. In the controlled E1 comparison, Without Mask performs best on KonIQ, SPAQ, and AGIQA-3K and has the lowest average MAE. Mask, no KL has lower average correlation. Adding KL to Mask recovers most of this gap and gives the strongest LIVE-W result. At E5, Without Mask is stronger on KADID-10k and slightly higher in CSIQ SRCC, whereas Mask has the best average PLCC and is stronger on most authentic and AI-generated benchmarks. Despite their similar average correlation, Without Mask has a substantially higher MAE (1.054 versus 0.543), exposing a calibration failure that PLCC and SRCC alone do not capture.

Rating and functional reasoning. Table 6 extends the main ablation in Table 3 with output-level diagnostics. The rating-only Actor attains strong rating alignment but produces edits that reduce quality on average. Among the controlled E1 variants, Without Mask has the strongest rating correlation, whereas Mask gives the largest quality gain. Mask + KL recovers most of the rating gap while retaining a positive mean gain. These results separate rating accuracy from the functional usefulness of reasoning.

Evidence and solution diversity. Without Mask produces only three distinct evidence strings, with one string accounting for 99.992% of the matched outputs. Its evidence has therefore collapsed even though its solutions remain more varied. Mask increases the number of unique evidence strings to 291 and yields the highest solution uniqueness. Adding KL further raises evidence uniqueness to 5,668 and lowers its top-1 share to 13.191%, at the cost of some solution diversity and edit gain. Evaluating evidence and solution separately is thus necessary: rating correlation, full-output validity, or solution diversity alone can conceal a collapsed reasoning field. Exact uniqueness remains a lexical proxy and does not by itself establish semantic diversity or faithfulness. At E5, Without Mask retains diverse evidence (74.719% unique; 0.556% top-1), while its solution uniqueness falls to 0.004% and its solution top-1 share reaches 100%. This field-wise mismatch shows why evidence and solution must be audited separately.

Final E5 balance. The E5 rows in Table 6 reinforce this multi-objective interpretation. Without Mask E5 reaches a raw gain of $+1.431$ but maps all 28,270 inputs to one normalized solution in the same template family despite retaining diverse evidence. Mask E5 obtains an average PLCC/SRCC of $0.838/0.812$, a mean quality gain of $+1.071$, and 19,618 unique evidence strings on the matched audit. It also produces 23,457 normalized solutions among 28,044 successful full-audit cases, with no detected house-template outputs. The higher Without Mask reward therefore reflects a universal edit rather than image-conditioned solution reasoning. Because the final runs change both credit routing and KL topology and use one seed per setting, this comparison is diagnostic rather than a pure causal estimate of the credit mask. Section E provides the corresponding training and validation dynamics.

E Detailed Credit-Mask Audit

E.1 Audit Scope and Protocol

We audit two complete five-epoch runs with the same Actor, data, prompt, and GRPO settings. **Mask** routes each reward to its supervised output with output-specific KL regularization; **Without Mask** broadcasts rewards over the complete output with one global KL term. Both use a KL coefficient of 0.02. The diagnostic reasoning reward bounds the raw Judge-score change:

$$R_{i,k}^{\mathrm{diag}} = \mathrm{sgn}(\Delta s_{i,k}) \left(1 - \mathrm{e}^{-\Delta s_{i,k}^2/2}\right). \tag{14}$$

The Editor consumes only the solution without an explicit semantic guardrail. Both runs use six samples per image, 160-token completions, and frozen vision encoder and aligner. Each contains 1,455 steps with 144 trajectories per step (209,520 per run; 419,040 total). No step or record is missing, duplicated, invalid, or non-finite. At Without Mask steps 711 and 713, all samples are Actor-ineligible, so absent Judge changes reflect eligibility rather than numerical or service failure. Because credit and KL routing change jointly and each setting has one seed, this audit is descriptive rather than causal.

E.2 Training Dynamics

Figure 7 reports exact four-rank means at the retained checkpoints. Both configurations improve rating reward. Without Mask produces larger reasoning reward and Judge gain, but also shorter outputs and solution collapse. Mask changes more gradually: its reasoning reward rises from 0.207 to 0.307 and Judge gain from 0.586 to 0.840. Rating rewards converge by E5. KL losses are not directly comparable because their token scopes differ. For Without Mask, mean reasoning reward and Judge gain increase from 0.463 and

Table 5: **Dataset-level rating performance of the credit-routing variants.** Dataset and Average entries report PLCC/SRCC, where Average is the unweighted average over the six datasets. The three E1 variants share the same source checkpoint; the two E5 runs are trained from the native parent for five epochs. The first two E1 variants use no KL. Bold marks the best result before rounding in each column.

Method	KonIQ	SPAQ	LIVE-W	AGIQA-3K	KADID-10k	CSIQ	Average	MAE ↓
Without Mask (E1)	**0.952/0.938**	**0.901**/0.898	0.897/0.877	**0.816/0.747**	0.661/0.668	0.771/0.746	0.833/**0.812**	**0.492**
Mask, no KL (E1)	0.945/0.931	0.894/0.892	0.892/0.867	0.805/0.719	0.641/0.650	0.758/0.727	0.822/0.798	0.706
Mask + KL (E1)	0.951/0.938	0.896/0.895	**0.903/0.880**	0.801/0.731	0.653/0.664	0.784/0.758	0.831/0.811	0.676
Without Mask (E5)	0.932/0.914	0.896/0.897	0.878/0.857	0.809/0.736	**0.676/0.684**	0.815/**0.786**	0.834/0.812	1.054
Mask (E5)	0.937/0.917	0.900/**0.899**	0.893/0.863	0.809/0.739	0.667/0.669	**0.824**/0.785	**0.838**/0.812	0.543

Table 6: **Joint rating, functional reasoning, and diversity analysis.** P/S is the six-dataset average PLCC/SRCC. Δs and Pos. are measured on 5,843 common images. Diversity columns report percentages; lower top-1 means less concentration. Unmarked diversity values use 23,599 matched inputs. † marks full E5 audits (28,270 Without Mask; 28,044 Mask).

	Rating	Functional reasoning		Language diversity			
Method	P/S ↑	Mean Δs ↑	Pos. (%) ↑	Evidence unique (%) ↑	Evidence top-1 (%) ↓	Solution unique (%) ↑	Solution top-1 (%) ↓
Baseline	0.652/0.645	−0.136	40.664	–	–	–	–
Rating-only	0.836/**0.815**	−0.260	30.686	–	–	–	–
Without Mask (E1)	0.833/0.812	+0.837	96.954	0.013	99.992	50.621	0.822
Mask, no KL (E1)	0.822/0.798	+0.917	97.707	1.233	21.196	**94.508**	**0.085**
Mask + KL (E1)	0.831/0.811	+0.788	95.054	24.018	13.191	75.228	0.191
Without Mask (E5)	0.834/0.812	**+1.431**	**99.997**	74.719†	0.556†	0.004†	100.000†
Mask (E5)	**0.838**/0.812	+1.071	98.813	**83.131**	**0.407**	83.644†	0.182†

1.189 over steps 1256–1355 to 0.478 and 1.232 over steps 1356–1455. Yet their final-window slopes are -0.0108 and -0.0246; the rating-reward slope is -0.0006. The endpoint therefore masks a plateau and slight decline.

E.3 Checkpoint Validation

Figure 8 evaluates the retained checkpoints on the same 200 images. PLCC and SRCC improve for both runs. Without Mask keeps a Judge gain near 1.18, while normalized-solution uniqueness falls from 91.88% at E1 to about 0.5% at E2. The audit further shows that the semantic dwelling-template rate reaches 100% at E1. Semantic collapse therefore precedes near-exact lexical collapse.

E.4 Cross-Dataset Outcomes

At E5, Without Mask obtains a larger pooled quality gain than Mask (1.431 versus 1.071), although their six-dataset PLCC/SRCC averages remain close (0.834/0.812 versus 0.838/0.812). Larger gain alone therefore does not establish healthier reasoning. All 28,270 Without Mask outputs share one normalized solution; Mask retains 23,457 among 28,044 successful outputs.

E.5 Collapse Milestones

Table 7 distinguishes semantic, phrase-level, and near-exact collapse. Under Without Mask, the semantic dwelling family exceeds 90% at step 265, whereas one normalized sentence does not exceed 90% until step 551. Duplicate matching would detect the failure 286 steps late. Mask avoids dwelling collapse, although its super-resolution and smoothing phrase backbone remains above 90% from step 407 onward.

Routing and detector	≥ 50%	≥ 90%	Sustained
Without Mask: semantic dwelling	236	265	948–E5
Without Mask: strict target phrase	356	475	–
Without Mask: modal normalized solution	520	551	973–E5
Without Mask: canonical sentence	1,133	1,146	–
Mask: super-resolution + smoothing	214	260	407–E5

Table 7: **Collapse milestones in global optimizer steps.** Semantic detectors identify shared concepts before normalized near-duplicate matching identifies one sentence.

E.6 Output-Length Evolution

Panel (d) of Figure 7 shows that the mean training completion decreases from 89.39 to 85.26 tokens under Without Mask but rises from 100.57 to 109.00 under Mask. On validation outputs, the corresponding solution lengths change from 32.56 to 18.00 and from 37.85 to 50.94. The former contraction is consistent with convergence to one reusable instruction.

E.7 Rating and Diversity Diagnostics

Without Mask retains a KonIQ PLCC/SRCC of 0.932/0.914 despite solution collapse. Ratings still vary: 2,009 of 2,010

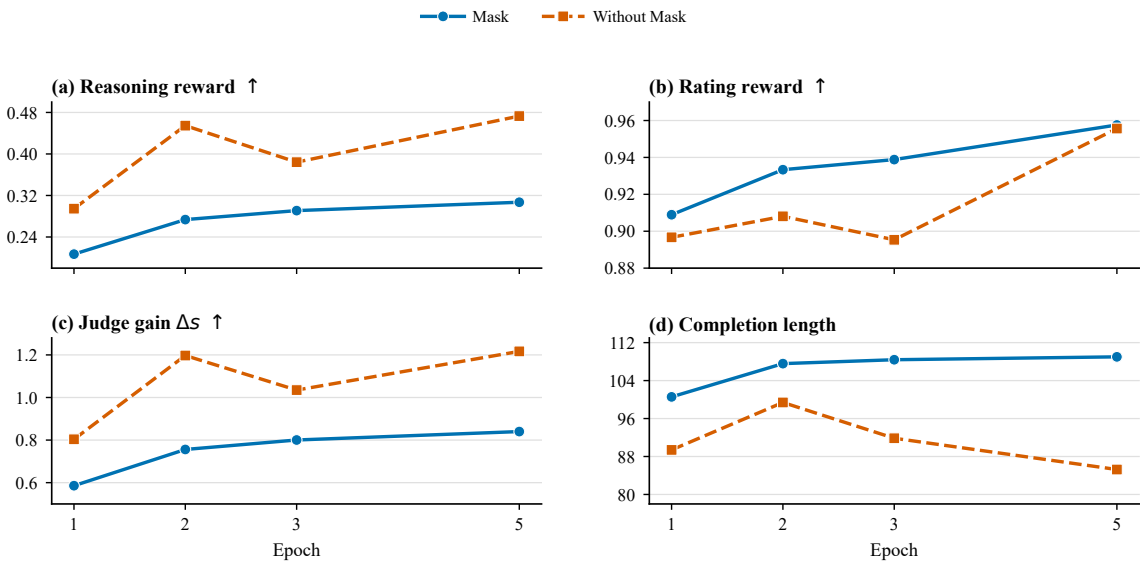

Figure 7: **Training dynamics at the retained checkpoints.** Each point is the exact four-rank mean for E1, E2, E3, or E5. Without Mask produces larger reasoning reward and Judge gain, whereas rating reward converges similarly. Its concurrent reduction in completion length is consistent with convergence toward a high-reward solution template.

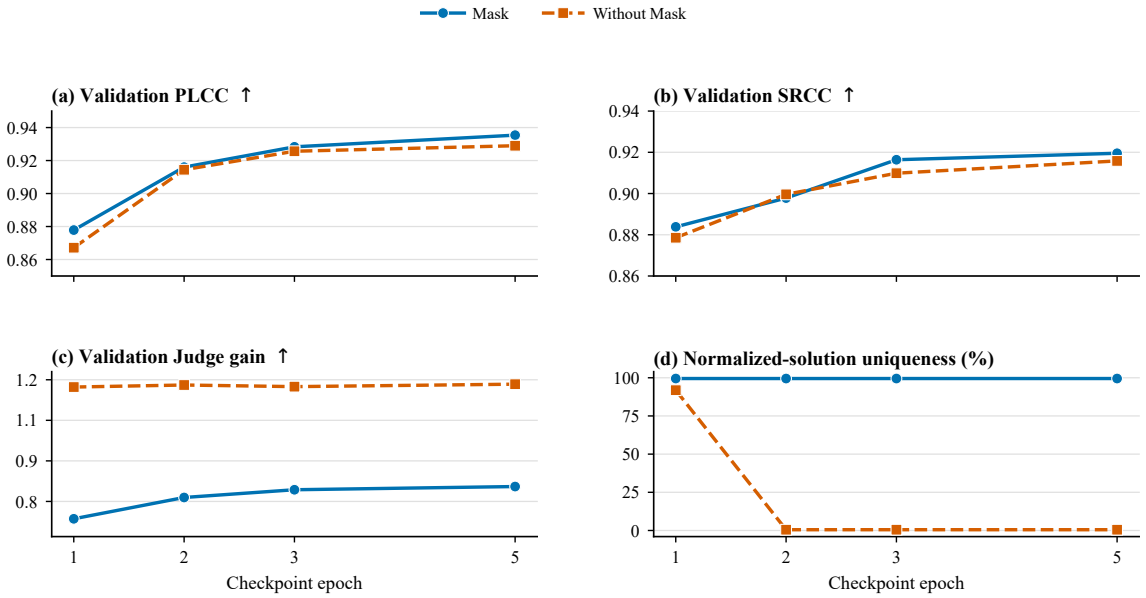

Figure 8: **Checkpoint validation dynamics.** Panels (a)–(c) report PLCC, SRCC, and Judge gain on the complete 200-image validation set. Panel (d) reports normalized-solution uniqueness. Without Mask maintains rating alignment and high Judge gain while its solutions collapse.

complete JSON outputs are unique, although all 2,010 solutions are identical after normalization. Rating correlation and full-output uniqueness therefore miss solution collapse. Evaluation must inspect the supervised output alongside image–solution relevance and semantic preservation.

E.8 Representative Validation Outputs

The Without Mask E5 Actor produces different evidence and ratings for the three examples, but all solutions normalize to one dwelling instruction. Sample 1 still receives 0.865 reward because the fixed intervention raises the Judge score from 2.37 to 4.37. Thus, the reward measures edited-image quality without establishing source-image faithfulness.

Sample	Rating	J_0	J_1	Δs	R
1	1.68	2.37	4.37	2.00	0.865
2	3.32	3.84	4.32	0.48	0.109
3	2.65	3.02	4.36	1.34	0.593

Table 8: Representative Without Mask validation outputs. Different evidence and ratings lead to the same normalized solution.

E.9 Causal Scope

The comparison changes credit and KL routing jointly with one seed per setting. It establishes a reproducible failure mode for Without Mask/global KL, but not the responsible change. Causal attribution requires a 2×2 grid of Without Mask versus Mask and global versus component KL, with multiple seeds.