SMARTPHOTOCRAFTER: UNIFIED REASONING, GENERATION AND OPTIMIZATION FOR AUTOMATIC PHOTOGRAPHIC EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic photographic enhancement requires more than executing predefined editing instructions. A model must first infer whether an image should be edited, determine what needs to be changed, and estimate the appropriate strength of each adjustment. This problem becomes particularly challenging in real photographs, where technical degradations and aesthetic deficiencies often co-occur, while successful edits must preserve the original content and photographic intent. We introduce **SmartPhotoCrafter**, a unified framework for automatic photographic optimization across restoration, retouching, and their compositions. SmartPhotoCrafter consists of an Image Critic that diagnoses image quality and infers editing decisions, and a Photographic Artist that realizes these decisions through conditional generation. Instead of relying on a text-only cascade, the Artist is conditioned directly on the Critic's contextual and reasoning representations, providing a richer interface between visual assessment and image generation. We further align reasoning and generation through executable editing targets and outcome-aligned reinforcement learning, with task-specific signals that evaluate editing direction, photometric agreement, perceptual quality, and content preservation. Experiments on automatic optimization, mixed restoration-retouching, and conventional restoration tasks demonstrate the effectiveness of the proposed method.

1 INTRODUCTION

Automatic photographic optimization aims to make high-quality photo enhancement accessible without requiring users to diagnose image defects or translate aesthetic goals into editing instructions Yao et al. (2026); Chen et al. (2025); Lin et al. (2025a). Effective enhancement involves more than applying an operation: one must judge whether a visual characteristic needs correction, identify appropriate adjustments, and determine their direction and strength while preserving the scene and its original contents. A user may simply want a better photograph without knowing whether it requires dehazing, exposure correction, or a different tonal balance. Although instruction-driven generative editors have substantially improved the execution of specified requests (Brooks et al., 2023; Liu et al., 2025; Wu et al., 2025a), they typically leave the formulation of those requests to the user. The value of automation is therefore not merely to remove an input prompt, but to make the system responsible for both precise photographic decision making and effective visual execution.

Moreover, technical degradation and aesthetic appearance interact: recovering a hazy photograph can also change its contrast and saturation, while an otherwise clear image may still benefit from tonal refinement. We therefore consider restoration and retouching within a common automatic optimization task. Restoration examples provide paired recovery targets, whereas retouching admits multiple plausible renderings of the same scene. The model should be able to handle both artifact corrections and tonal refinements while preserving the photographic content and intent of the input.

In this paper, we present **SmartPhotoCrafter**, a framework that couples an MLLM-based *Image Critic* with a generative *Photographic Artist*. The Critic analyzes the input and produces photographic reasoning, edit suggestions, and a quality score. We introduce a **latent photographic decision interface** that transfers the Critic's contextual representations directly to the Artist. A textual suggestion such as increasing exposure identifies an operation, but its realization also depends on the

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

photograph's existing tonal distribution and the other adjustments under consideration. Latent conditioning makes the image-aware assessment available to generation without requiring all relevant cues to be explicitly restated in the suggestion. The Critic's textual output specifies the editing decision, while its context and reasoning representations condition the Artist's execution. A hidden-state connection alone, however, does not teach the Artist how to interpret the Critic's representations. We therefore ground the interface in executable editing effects. For retouching suggestions, the Critic's proposed operations are applied to the input to construct online supervision targets. The Artist is adapted to reproduce these effects under the corresponding latent conditions, alongside paired restoration and retouching supervision. This establishes a learned correspondence between the assessment and its visual consequences, rather than assuming that independently trained understanding and generation modules are already compatible.

We further develop **outcome-aligned reinforcement learning** centered on the quality of the generated result. To provide outcome-aligned supervision, we design reward signals for the Critic and Artist according to their respective roles. For the Artist, the rewards measure direction-sensitive semantic compliance, attribute-wise photometric improvement, and reference-relative perceptual consistency. These signals distinguish whether an inferred adjustment is correctly followed, whether it is realized with an appropriate magnitude, and whether the perceptual appearance of the image is preserved. The Critic reuses the same attribute-wise criterion to evaluate the rule-executed effects of its retouching suggestions, while the generated images provide score-ranking feedback. The two policies retain distinct rewards but interact through their conditions, suggested operations, and generated outcomes. Together, these role-specific rewards bridge the gap between photographic decisions and their realized effects, preventing optimization from stopping at plausible editing suggestions and instead grounding both reasoning and generation in the quality of the results.

In summary, our contributions are threefold: (1) **Automatic enhancement across restoration and retouching:** We address automatic photographic enhancement across restoration, retouching, and their combinations. We bring restoration and retouching into a single automatic enhancement workflow, allowing users to improve a photo with one simple request; (2) **Assessment-conditioned photographic generation:** A CriticArtist framework is developed, which conditions image generation on contextual representations from photographic assessment, connecting image diagnosis with its visual realization; and (3) **Photographic outcome alignment:** We design outcome-based photographic feedback that evaluates directional compliance, reference-relative attribute improvement, and perceptual consistency. These signals support coordinated refinement of editing suggestions and generated images through role-specific optimization objectives.

2 Related Work

Photographic enhancement and retouching agents. Photographic enhancement covers both the recovery of degraded images and adjustments to tone, color, and visual appearance. Recent MLLM-based systems make these adjustments more accessible through language-guided planning and execution Dutt et al. (2025); Ye-Bin et al. (2025); Lin et al. (2025a); Yao et al. (2026); Lin et al. (2025b); Wu et al. (2026); Zhang & Yang (2026). SmartPhotoCrafter studies automatic enhancement through a shared generative model for restoration and retouching, with photographic assessment directly conditioning image generation.

Generative image editing. Instruction-based generative editing learns to transform images according to natural language requests. Progress in this direction has been supported by editing datasets and training frameworks that improve instruction coverage, annotation quality, and support for diverse editing tasks (Zhang et al., 2023a; Yu et al., 2025; Wu et al., 2025b;a; Google, 2026). Recent multimodal models further connect visual understanding with image generation, enabling flexible editing through joint image and language conditioning (Liu et al., 2025; Wu et al., 2025a;b). Within photographic editing, dedicated methods explore fine-grained control over camera parameters and their visual effects (Qin et al., 2025; Yang et al., 2025). Our work focuses on automatic photographic enhancement across restoration and retouching, where editing suggestions are inferred from the input image. The generator is conditioned on the Critic's contextual representations. This connection is adapted through paired supervision and executable editing targets.

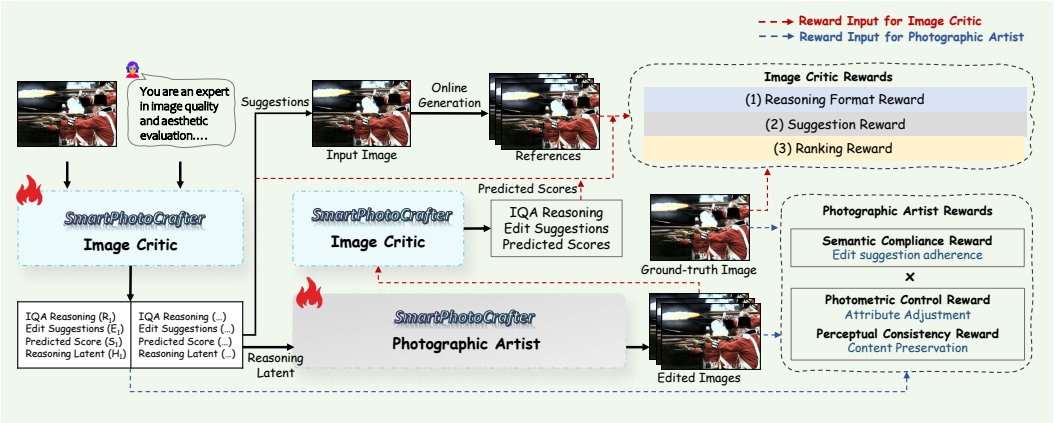


Figure 1: Outcome-aligned reinforcement learning in SmartPhotoCrafter.

Quality assessment and feedback. Image quality assessment provides perceptual criteria for evaluating photographic quality and aesthetic appeal (Talebi & Milanfar, 2018; Ke et al., 2021). Recent multimodal approaches extend quality assessment to language-based scoring, degradation analysis, and reasoning about visual quality (Wu et al., 2024b; Li et al., 2026). SmartPhotoCrafter uses photographic assessment to condition image generation and provides editing-specific feedback on directional compliance, attribute-wise improvement, and perceptual consistency.

Detailed discussion of the related works is provided in Appendix A.

3 Method

3.1 Automatic Photographic Optimization

SmartPhotoCrafter is proposed to address automatic photographic optimization, unifying technical restoration and aesthetic retouching. Given an input photograph X and a fixed task-level request shared across inputs, the goal is to generate an enhanced image X_e without image-specific editing instructions. Because the desired edits are not specified in advance and restoration and retouching can interact, generation must selectively address image-specific needs with adjustments of appropriate type and magnitude, while preserving scene content and respecting photographic intent.

The framework comprises an Image Critic f_c with parameters θ and a Photographic Artist f_a with parameters ϕ. The Critic produces a structured response $o = (R, E, S)$, consisting of photographic reasoning R, edit suggestions E, and a quality score S. The Artist generates X_e from X, conditioned on the Critic's contextual hidden representations, realizing restoration and retouching within a shared conditional generative model.

3.2 Latent Photographic Decision Interface

We condition the Photographic Artist on the representations computed during photographic assessment by the Image Critic. Let $h_t^{(L)}$ denote the final-layer hidden state at token position t. The conditioning sequence and generated output are:

$$H_c = \mathrm{Concat}\left(h_0^{(L)}, h_1^{(L)}, \ldots, h_{T-1}^{(L)}\right), \qquad X_e = f_{a,\phi}(X, H_c), \tag{1}$$

where T spans the input context and generated reasoning tokens, and H_c retains the multimodal context alongside the analysis. Unlike textual cascading, this setting exposes the Critic's multimodal context and reasoning directly to the Artist.

The structured response and the latent sequence play complementary roles. The response provides inspectable semantic assessment, while the latent sequence serves as the Artist's generation condition. Unlike executing a prescribed sequence of editing tools, the Artist realizes the inferred adjustments

within one conditional generative process, combining representation-level decision transfer with a shared model for restoration and retouching.

3.3 Learning Decision-Conditioned Photographic Generation

Photographic foundations. We first perform task-specific supervised pretraining of the two modules, starting from pretrained backbones. The Critic learns to assess photographic quality and formulate editing suggestions, while the Artist learns to perform restoration and retouching under concise editing instructions. This provides a photographic editing foundation for subsequent automatic generation. These capabilities alone, however, do not establish how the Critic's contextual representations should control the generated image: an Artist trained on editing instructions has not yet learned to interpret the richer states produced during photographic assessment.

Executable-edit grounding. To learn the correspondence between the two modules, we freeze the pretrained Critic and fine-tune the Artist conditioned on H_c, using paired restoration and retouching examples together with online targets constructed from the Critic's suggestions.

For suggestions supported by the rule-based retouching operators, we execute the proposed edits E on the input X to obtain a reference image:

$$X_{\mathrm{ref}} = \mathcal{T}(X; E), \tag{2}$$

and train the Artist to reproduce its visual effects under the corresponding H_c. Here X_{ref} is an operator-produced target, distinct from the dataset reference X_{gt}. Writing Y for the supervision target associated with an adaptation example, the objective is:

$$\mathcal{L}_{\mathrm{adapt}}(\phi) = \mathbb{E}_{(X,H_c,Y)}[\ell_{\mathrm{FM}}(\phi; X, H_c, Y)]\,, \tag{3}$$

where ℓ_{FM} is the Artist's flow-matching loss, $Y = X_{\mathrm{gt}}$ denotes a restoration target or a selected retouched rendering of the same scene for paired examples, and $Y = X_{\mathrm{ref}}$ is the online operator-grounded reference. Both types of reference provide supervision and are not supplied as inputs at inference. The online construction anchors the latent interface to the visual consequences based on the Critic's own suggestions. For example, a predicted slight exposure increase is paired with an image exhibiting that adjustment under the same assessment condition H_c.

The adaptation establishes a correspondence between photographic assessments and visual realization. However, faithfully realizing an inferred adjustment does not by itself ensure that the resulting photograph is better, motivating the design of outcome-aligned reinforcement learning.

3.4 Outcome-Aligned Reinforcement Learning

Supervised training and decision-conditioned adaptation provide photographic editing priors and connect inferred adjustments to their visual realization. Further refinement, however, requires evaluating how the generated changes improve the photograph: an edit may follow the suggested direction yet overshoot the desired correction, displace attributes already close to the reference, or introduce unwanted deviations in appearance. Moreover, the quality of the generated outcome also depends on the photographic decisions that condition its generation, making decision refinement and generative optimization inherently coupled.

Figure 1 summarizes our outcome-aligned reinforcement learning framework, centered on the Artists generated outcomes. Its reward couples direction-sensitive compliance with attribute-wise photometric improvement and reference-relative perceptual consistency. Crucially, photometric improvement is evaluated relative to both the input and the selected reference: the criterion rewards reducing existing discrepancies while preserving attributes already within tolerance, distinguishing beneficial correction from unnecessary adjustment. Perceptual feedback complements these attribute statistics by encouraging agreement with the reference appearance. To support this generative optimization, the Critic is refined using the same photometric criterion on rule-executed suggestions, together with auxiliary score-ranking feedback from generated images.

3.4.1 Outcome-Aligned Photographic Generation

The Photographic Artist's reward evaluates three complementary aspects of a generated edit: whether its photometric changes follow the Image Critic's suggestions, whether they reduce attribute

discrepancies to the selected reference, and whether the resulting appearance remains perceptually consistent with that reference. We combine these signals as:

$$r_{\mathrm{PA}} = r_{\mathrm{comp}} \left[\lambda_1 r_{\mathrm{photo}}(X, X_e, X_{\mathrm{gt}}) + \lambda_2 r_{\mathrm{perc}}\right], \quad \lambda_1 = 1.0, \qquad \lambda_2 = 0.5. \tag{4}$$

where $r_{\mathrm{photo}}(X, X_e, X_{\mathrm{gt}})$ measures reference-relative attribute improvement for an input image X. The compliance factor modulates the combined photometric and perceptual score according to the fraction of satisfied suggestions, coupling the outcome evaluation to the inferred editing directions.

Direction-sensitive semantic compliance. Let $E_{\mathrm{photo}} = \{e_i\}_{i=1}^{N}$ denote the color- and tone-related suggestions evaluated by this reward. For each suggestion, c_i indicates whether its specified attribute change is satisfied in X_e relative to X. For $N > 0$, we compute

$$c_i = \begin{cases} 1, & \text{if the attribute change specified by } e_i \text{ is satisfied,} \\ 0, & \text{otherwise,} \end{cases} \qquad r_{\mathrm{comp}} = \frac{1}{N} \sum_{i=1}^{N} c_i. \tag{5}$$

This term evaluates directional compliance within the photometric subset rather than the fidelity of all restoration operations. It does not determine whether the amount of change is appropriate relative to the selected target appearance.

Reference-relative photometric improvement. To evaluate the extent of photometric refinement, we measure how an edit changes the discrepancy to the reference in explicit attribute statistics. Let $a_k(\cdot)$ be a computable scalar descriptor for photometric attribute k, such as exposure or contrast. For an evaluated image X_e, define

$$d_k^{\mathrm{in}} = \left|a_k(X_{\mathrm{gt}}) - a_k(X)\right|, \qquad d_k^{\mathrm{out}}(X_e) = \left|a_k(X_{\mathrm{gt}}) - a_k(X_e)\right|. \tag{6}$$

The attribute-wise reward and its aggregation are

$$r_{\mathrm{attr}}^{k}(X, X_e, X_{\mathrm{gt}}) = \begin{cases} 1, & d_k^{\mathrm{in}} < \tau_k \text{ and } d_k^{\mathrm{out}}(X_e) \le \tau_k, \\ 0, & d_k^{\mathrm{in}} < \tau_k \text{ and } d_k^{\mathrm{out}}(X_e) > \tau_k, \\ \dfrac{d_k^{\mathrm{in}} - d_k^{\mathrm{out}}(X_e)}{d_k^{\mathrm{in}} + \epsilon}, & \text{otherwise,} \end{cases} \tag{7}$$

$$r_{\mathrm{photo}}(X, X_e, X_{\mathrm{gt}}) = \frac{1}{K} \sum_{k=1}^{K} r_{\mathrm{attr}}^{k}(X, X_e, X_{\mathrm{gt}}).$$

where τ_k is an attribute-specific tolerance, ϵ provides numerical stability, and K is the number of evaluated attributes. The reward measures relative discrepancy reduction when correction is truly needed. The criterion therefore provides attribute-level feedback on the realized adjustment, rather than checking exact editing parameters.

Reference-relative perceptual consistency. Images with similar global photometric statistics can still differ substantially in appearance. We therefore complement the attribute criterion with

$$r_{\mathrm{perc}} = \exp[-\,\mathrm{LPIPS}(X_e, X_{\mathrm{gt}})], \tag{8}$$

using LPIPS (Zhang et al., 2018) to encourage agreement with the reference in perceptual feature space. This provides a soft consistency signal beyond the selected photometric statistics, rather than a hard guarantee of unchanged structure or identity.

3.4.2 Effect-Grounded Photographic Decision Refinement

Photographic optimization also depends on which adjustments the Critic proposes. We therefore refine its decisions using feedback on executable suggestion effects, complemented by auxiliary signals for input-output score ordering and response structure.

Edit-suggestion reward through executable effects. For restoration examples with annotated degradations, suggestions are evaluated for inclusion of the corresponding corrective operation, such as haze removal for a hazy input. For retouching examples, we do not require a single annotated suggestion sequence. Instead, the proposed color and tone adjustments are executed through $\mathcal{T}$ to obtain the pseudo-edited image X_{ref} defined in Eq. (2). We evaluate its attribute-wise improvement using the same photometric criterion as for the Artist:

$$r_{\mathrm{sugg}}^{\mathrm{retouch}} = r_{\mathrm{photo}}(X, X_{\mathrm{ref}}, X_{\mathrm{gt}}). \tag{9}$$

This evaluates suggestions through their visual consequences rather than their agreement with a unique textual sequence. The Critic and Artist share the attribute-wise criterion but are evaluated through different execution mechanisms: the Critic's suggestions through the rule-executed X_{ref}, and the Artist through its generated image X_e. The former provides a proxy for the effects of the proposed retouching operations; it does not substitute for evaluating their realization by the Artist.

Auxiliary score-ranking feedback. The Critic also re-assesses the Artist's generated images. We then construct a pairwise ranking signal that reflects whether the generated edit leads to an improvement over the input, and define the following aesthetic-aware reward:

$$r_{\mathrm{rank}} = \begin{cases} 1, & \text{if } S(X) < S(X_e) \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $S(X)$ and $S(X_e)$ denote the predicted score of the input image and edited image from the Image Critic, respectively. This term encourages the input-output ordering in the Critic's own assessments, linking its scoring behavior to generated photographs. It is an auxiliary internal ordering signal rather than independent verification that a generated image is better than its input.

Structured-response regularization. A format reward checks the prescribed ordering of reasoning, suggestions, and score in the structured response. Together, the suggestion, ranking, and format terms form the Critic reward r_{IC} used for policy optimization.

3.4.3 Coordinated Policy Updates

We sample groups of Critic responses and generate edits conditioned on their latent representations. The Artist is optimized with DiffusionNFT (Zheng et al., 2026) using r_{PA}, and the Critic with GRPO (Guo et al., 2025) using r_{IC}. The Critic provides the generation conditions and edit suggestions used to evaluate directional compliance, while the generated images are used to compute the Critic's score-ranking reward. The two policies use separate rewards and optimizers. Full optimization objectives are provided in Appendix B.

3.5 Training and Inference

We initialize the Critic with Qwen2.5-VL-7B (Bai et al., 2025) and the Artist with Qwen-Image-Edit (Wu et al., 2025a). Stage I uses task-specific supervised pretraining to establish their respective assessment and editing capabilities. Qwen2.5-VL-72B supplies reasoning and suggestion annotations; available mean opinion scores replace generated scores, and score-token losses are masked for editing samples. The Artist is trained with flow matching on single-edit and combined restoration-retouching pairs. We then perform the frozen-Critic adaptation in Section 3.3, followed by outcome-aligned RL optimization in Section 3.4. Thus, the three stages implement capability acquisition, decision-to-effect alignment, and photographic outcome refinement, respectively. Dataset construction and training settings are detailed in Appendices C and D.

At inference, the Critic assesses the input under the fixed enhancement request, and the Artist generates the output conditioned on (X, H_c). No paired reference, rule-based target construction, or reward evaluation is required. Restoration and retouching share this inference procedure, with the Artist realizing the combined transformation through conditional diffusion sampling.

Table 1: Comparison on automatic photographic enhancement.

Method	Q-Align↑	NIMA↑	MUSIQ↑	DINO↑	CLIP↑	FID↓	LPIPS↓
Instruct-Pix2Pix Brooks et al. (2023)	2.93	5.27	60.48	0.20	0.45	224.90	0.68
FLUX.2-Dev Labs (2025)	<u>3.46</u>	<u>5.67</u>	**72.94**	0.93	0.88	76.26	0.26
Qwen-Image-Edit Wu et al. (2025a)	3.38	5.57	68.63	<u>0.96</u>	<u>0.93</u>	<u>42.81</u>	<u>0.17</u>
OmniGen2 Wu et al. (2025b)	3.37	**5.69**	67.52	0.81	0.85	92.78	0.41
Step1X-Edit Liu et al. (2025)	3.12	5.31	66.34	0.95	0.89	48.48	0.19
MonetGPT Dutt et al. (2025)	3.08	5.32	64.11	<u>0.96</u>	0.92	45.93	0.18
JarvisEvo Lin et al. (2025b)	3.01	5.27	62.71	0.95	0.89	47.63	0.20
Nano Banana 2 Google (2026)	3.43	5.57	69.37	0.95	0.89	47.56	<u>0.17</u>
Ours	**3.52**	5.66	<u>69.52</u>	**0.98**	**0.96**	**27.96**	**0.10**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND EVALUATION PROTOCOLS

Tasks and data. The automatic enhancement suite comprises FiveK retouching examples (Bychkovsky et al., 2011a), synthetically perturbed high-aesthetic AVA images (Murray et al., 2012), and restoration test images. All methods receive an input photograph and a task-level enhancement request, without image-specific corrective instructions. A separate controlled experiment supplies combined restoration-retouching instructions to test execution against the prescribed target. Dataset roles and the details of each evaluation setting are detailed in Appendix D.

Evaluation criteria. We evaluate photographic quality, reference consistency, and human judgments across the experiments. Q-Align, MUSIQ, and NIMA assess image quality and aesthetics without reference images (Wu et al., 2024b; Ke et al., 2021; Talebi & Milanfar, 2018). DINO and CLIP measure feature similarity to the references, while LPIPS and FID assess perceptual and distributional agreement, respectively (Caron et al., 2021; Zhang et al., 2018; Heusel et al., 2017). For prescribed multi-edit tasks and image restoration, we additionally report PSNR and SSIM to evaluate reconstruction fidelity (Wang et al., 2004), with DISTS further assessing structure and texture similarity in restoration. Human evaluation complements these metrics through five-point ratings of photorealistic aesthetics and consistency with the input photograph. The Critic's own scores are used for training feedback, not as evaluation metrics.

Implementation. We initialize the Critic with Qwen2.5-VL-7B and the Artist with Qwen-Image-Edit (Bai et al., 2025; Wu et al., 2025a). Training uses eight NVIDIA A100 GPUs. Data construction and optimization settings are provided in Appendices C and D

4.2 AUTOMATIC PHOTOGRAPHIC ENHANCEMENT

Quality and reference consistency. Table 1 shows that SmartPhotoCrafter achieves the highest Q-Align score and competitive NIMA and MUSIQ scores, together with the strongest reference consistency across DINO, CLIP, LPIPS, and FID. For reference, the ground-truth images achieve MUSIQ and NIMA scores of 70.96 and 5.23, respectively. These results indicate that the photographic adaptation improves both predicted image quality and agreement with the reference enhancements. Figure 2 illustrates improved visibility and tonal balance in dark and haze-affected scenes while retaining the main scene structure.

Human evaluation. We compare SmartPhotoCrafter with JarvisEvo (Lin et al., 2025b), FLUX.2-Dev (Labs, 2025), and Nano Banana 2 (Google, 2026) using ratings from 1 to 5 for photorealistic aesthetics and image consistency, with higher scores indicating better performance. As shown in Figure 3, SmartPhotoCrafter receives ratings of 4-5 for 57.4% of aesthetic judgments, close to FLUX.2-Dev's 58.7% and above Nano Banana 2's 38.0% and JarvisEvo's 9.1%. For image consistency, SmartPhotoCrafter and Nano Banana 2 share the highest proportion of ratings of 4-5 at 75.9%. These results show that SmartPhotoCrafter combines a high proportion of favorable aesthetic ratings with strong content preservation. The evaluation protocol is detailed in Appendix H.

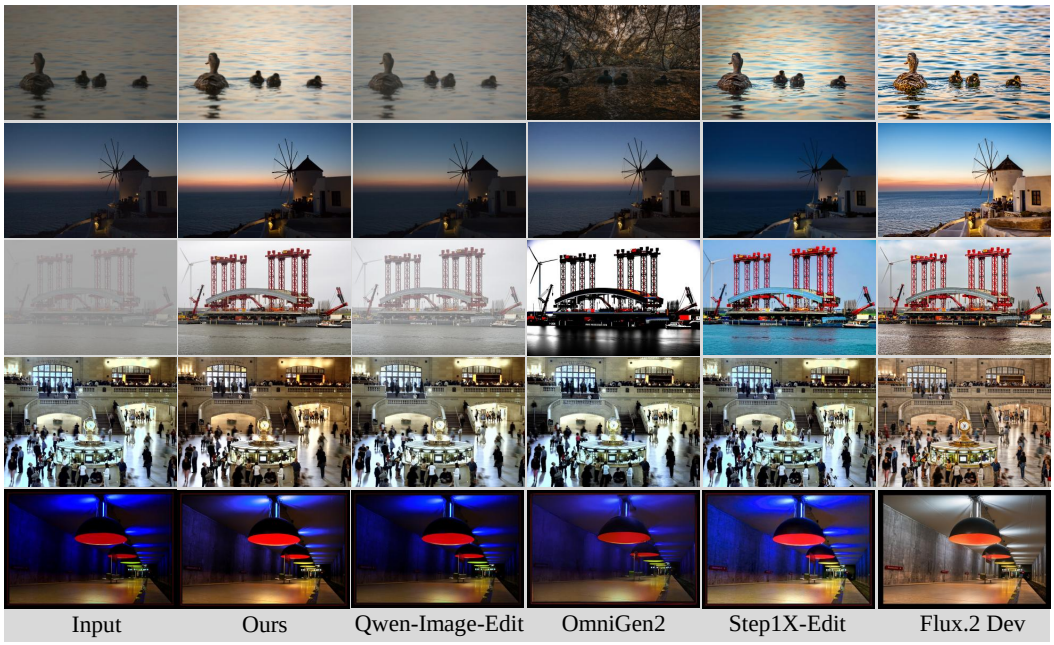


Figure 2: Visual comparison of automatic photographic enhancement task across different methods.

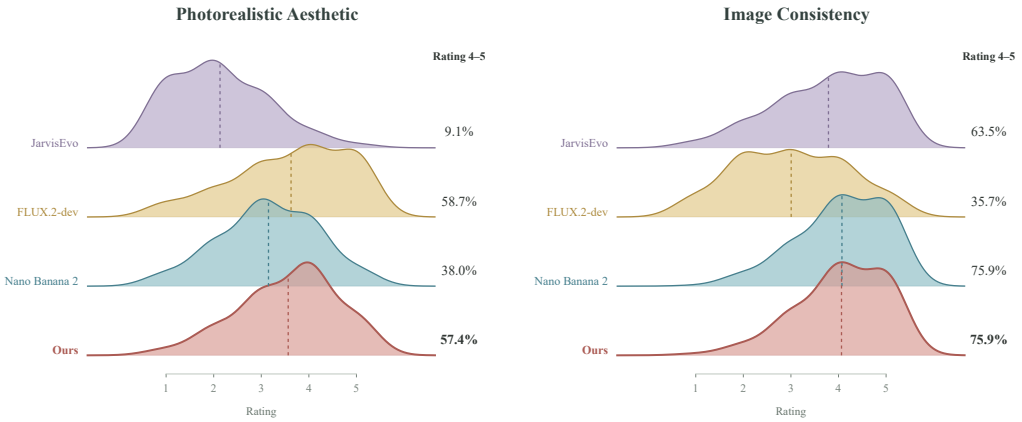


Figure 3: Human evaluation of automatic photographic enhancement. Rating distributions are shown for photorealistic aesthetics (left) and image consistency (right).

4.3 Ablation Study

Decision interface. We compare the two conditioning interfaces after Stage II, before reinforcement learning. As shown in Table 2, the latent decision interface improves NIMA from 5.51 to 5.57 and MUSIQ from 67.46 to 67.82, while reducing FID from 36.32 to 30.61. Q-Align increases slightly, while DINO and CLIP remain unchanged. These results support direct latent conditioning as a more effective interface for transferring photographic decisions from the Critic to the Artist than re-encoding textual suggestions.

Outcome alignment. Compared with Supervised only, the Full model improves all reported metrics, increasing MUSIQ from 67.82 to 69.52 and reducing FID from 30.61 to 27.96. To isolate the contribution of attribute-wise photometric supervision, we remove r_{photo} from the Artist reward while keeping the Critic reward unchanged. The variant w/o r_{photo} also improves quality scores, with MUSIQ reaching 68.25 and NIMA reaching 5.58. However, FID worsens to 38.51 and CLIP

Table 2: Ablation study of the photographic decision interface and outcome-aligned reinforcement learning.

Variant	Q-Align↑	NIMA↑	MUSIQ↑	DINO↑	CLIP↑	FID↓
Decision Interface						
Textual cascading	3.23	5.51	67.46	0.97	0.95	36.32
Latent decision interface	3.24	5.57	67.82	0.97	0.95	30.61
Outcome Alignment						
Supervised only	3.24	5.57	67.82	0.97	0.95	30.61
w/o r_{photo}	3.36	5.58	68.25	0.97	0.94	38.51
Full model	3.52	5.66	69.52	0.98	0.96	27.96

Table 3: Comparison on multi-edit instruction adherence.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	DINO ↑	CLIP ↑
Instruct-Pix2Pix	13.84	0.49	0.52	186.41	0.49	0.60
FLUX.2-Dev	15.93	0.71	0.20	46.70	0.96	0.90
Qwen-Image-Edit	14.76	0.44	0.24	52.89	0.93	0.90
OmniGen2	12.56	0.33	0.47	97.36	0.82	0.82
Step1X-Edit	17.05	0.62	0.21	38.91	0.94	0.88
SmartPhotoCrafter	**21.05**	**0.82**	**0.09**	**22.93**	**0.97**	**0.96**

decreases from 0.95 to 0.94, while DINO remains unchanged. This contrast shows that gains in quality scores do not necessarily translate into closer agreement with the references. Including r_{photo} improves both quality scores and reference consistency, supporting the contribution of explicit photometric supervision to outcome-aligned RL.

4.4 Multi-edit instruction adherence

We evaluate instruction-following capability in a multi-edit setting, focusing on whether the model can accurately execute a combination of restoration and retouching edits with explicit instructions. As shown in Table 3, SmartPhotoCrafter achieves the best results across all reported metrics. The higher PSNR and SSIM and lower LPIPS indicate closer agreement with the reference realizations of the prescribed edits, while the lower FID indicates closer agreement with the reference image distribution. These results support the Artist's ability to realize the tested combinations of restoration and retouching operations.

Additional results and analysis. Appendix E reports deblurring and dehazing results, and Appendix G provides further examples of automatic enhancement and combined edits. Appendix F compares textual cascading with latent conditioning. Appendix I presents complete Critic responses alongside generated results, including examples of inconsistent diagnoses and editing suggestions.

5 Conclusion

We presented SmartPhotoCrafter, a framework for automatic photographic optimization that connects image-conditioned assessment with generative restoration and retouching. Its latent photographic decision interface is grounded in executable editing effects, and outcome-aligned reinforcement learning refines generation through directional compliance, attribute-wise photometric improvement, and perceptual consistency. Across the evaluated benchmarks, the method achieves competitive no-reference quality and strong reference-relative fidelity, including combined restoration-retouching transformations. The reward ablation further shows that photographic supervision, rather than RL alone, matters for balancing the evaluated criteria. These findings motivate studying automatic enhancement as both photographic decision making and precise visual execution.

AI USE STATEMENT

We used Qwen2.5-VL-72B to generate photographic analyses, editing suggestions, and initial quality scores for training data construction, as described in Appendix C. Besides, LLM-based assistants were used solely to improve the clarity, grammar, and readability of the manuscript. They were not used to generate research ideas, design the methodology or experiments, or interpret the experimental results. All AI-assisted edits were reviewed and verified by the authors.

ETHICS STATEMENT

This work studies automatic photographic restoration and retouching. We use existing image datasets and construct additional supervision through synthetic edits and model-generated annotations. The use and redistribution of these data should respect the applicable licenses and the privacy of depicted individuals. Generative enhancement may alter fine details or introduce plausible but inaccurate content, so enhanced images should not be treated as faithful records in contexts requiring documentary accuracy. Our evaluation includes human judgments of photographic aesthetics and image consistency. Aesthetic preferences vary across individuals and contexts; the reported ratings and quality scores do not represent universal standards of photographic quality.

REPRODUCIBILITY STATEMENT

Section 3.5 describes the training stages and inference procedure. Appendix C details the data sources, construction pipeline, annotation prompt, and supervision used at each stage. Appendix B provides the policy optimization objectives, while Appendix D reports dataset splits, training settings, and evaluation criteria. Section 4 presents the main comparisons and ablation studies, with additional restoration results in Appendix E and the human evaluation protocol in Appendix H. These descriptions document data construction, model training, and evaluation.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011a.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pp. 97–104. IEEE, 2011b.

Zhuoxuan Cai, Jian Zhang, Xinbin Yuan, Peng-Tao Jiang, Wenxiang Chen, Bowen Tang, Lujian Yao, Qiyuan Wang, Jinwen Chen, and Bo Li. Q-ponder: A unified training pipeline for reasoning-based visual quality assessment. *arXiv preprint arXiv:2506.05384*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Haoyu Chen, Keda Tao, Yizao Wang, Xinlei Wang, Lei Zhu, and Jinjin Gu. Photoartagent: Intelligent photo retouching with language model-based artist agents. *arXiv preprint arXiv:2505.23130*, 2025.

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yuning Cui, {Syed Waqas} Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and {Fahad Shahbaz} Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *13th International Conference on Learning Representations, ICLR 2025*, 13th International Conference on Learning Representations, ICLR 2025, pp. 57335–57356. International Conference on Learning Representations, ICLR, 2025.

Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pp. 870878, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240531. URL https://doi.org/10.1145/3240508.3240531.

Niladri Shekhar Dutt, Duygu Ceylan, and Niloy J Mitra. Monetgpt: Solving puzzles enhances mllms' image retouching skills. *ACM Transactions on Graphics (TOG)*, 44(4):1–12, 2025.

Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3677–3686, 2020.

Armando Fortes, Tianyi Wei, Shangchen Zhou, and Xingang Pan. Bokeh diffusion: Defocus blur control in text-to-image diffusion models. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Trans. Graph.*, 36(4), July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073592. URL https://doi.org/10.1145/3072959.3073592.

Google. Nano banana 2: Combining pro capabilities with lightning-fast speed, 2026. URL https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li,

Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12626–12636, 2025.

Weiqi Li, Xuanyu Zhang, Shijie Zhao, Yabin ZHANG, Junlin Li, Li zhang, and Jian Zhang. Q-insight: Understanding image quality via visual reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=Bds54EfR9x.

Zinuo Li, Xuhang Chen, Shuqiang Wang, and Chi-Man Pun. A large-scale film style dataset for learning multi-frequency driven film enhancement. In *IJCAI*, volume 2023, pp. 1160–1168, 2023.

Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2019.

Yunlong Lin, Zixu Lin, Kunjie Lin, Jinbin Bai, Panwang Pan, Chenxin Li, Haoyu Chen, Zhongdao Wang, Xinghao Ding, Wenbo Li, et al. Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. *arXiv preprint arXiv:2506.17612*, 2025a.

Yunlong Lin, Linqing Wang, Kunjie Lin, Zixu Lin, Kaixiong Gong, Wenbo Li, Bin Lin, Zhenxi Li, Shiyi Zhang, Yuyang Peng, et al. Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization. *arXiv preprint arXiv:2511.23002*, 2025b.

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

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012a.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a completely blind image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012b.

Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2408–2415. IEEE, 2012.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Xinran Qin, Zhixin Wang, Fan Li, Haoyu Chen, RenJing Pei, WenBo Li, and XiaoChun Cao. Camedit: Continuous camera parameter control for photorealistic image editing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Zia-ur Rahman, Daniel J Jobson, and Glenn A Woodell. Multi-scale retinex for color image enhancement. In *Proceedings of 3rd IEEE international conference on image processing*, volume 3, pp. 1003–1006. IEEE, 1996.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pp. 184–201. Springer, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Tim Seizinger, Florin-Alexandru Vasluianu, Marcos V Conde, Zongwei Wu, and Radu Timofte. Bokehlicious: Photorealistic bokeh rendering with controllable apertures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8908–8917, 2025.

Debashis Sen and Sankar K Pal. Automatic exact histogram specification for contrast enhancement and visual system based quantitative evaluation. *IEEE Transactions on image processing*, 20(5): 1211–1220, 2010.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. 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*, pp. 3667–3676, 2020.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. doi: 10.1109/TIP.2018.2831899.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

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

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, Geng Xue, Wenxiu Sun, Qiong Yan, and Weisi Lin. 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 (CVPR)*, pp. 25490–25500, June 2024a.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: teaching lmms for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54015–54029, 2024b.

Qiucheng Wu, Jing Shi, Simon Jenni, Kushal Kafle, Tianyu Wang, Shiyu Chang, and Handong Zhao. Retouchiq: Mllm agents for instruction-based image retouching with generalist reward. *arXiv preprint arXiv:2602.17558*, 2026.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. *arXiv preprint arXiv:2505.14460*, 2025c.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 35(2), February 2016. ISSN 0730-0301. doi: 10.1145/2790296. URL https://doi.org/10.1145/2790296.

Qirui Yang, Yang Yang, Ying Zeng, Xiaobin Hu, Bo Li, Huanjing Yue, Jingyu Yang, and Peng-Tao Jiang. Cameramaster: Unified camera semantic-parameter control for photography retouching, 2025. URL https://arxiv.org/abs/2511.21024.

Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021.

Mingde Yao, Zhiyuan You, King-Man Tam, Menglu Wang, and Tianfan Xue. Photoagent: Agentic photo editing with exploratory visual aesthetic planning, 2026. URL https://arxiv.org/abs/2602.22809.

Ruijie Ye, Jiayi Zhang, Zhuoxin Liu, Zihao Zhu, Siyuan Yang, Li Li, Tianfu Fu, Franck Dernoncourt, Yue Zhao, Jiacheng Zhu, et al. Agent banana: High-fidelity image editing with agentic thinking and tooling. *arXiv preprint arXiv:2602.09084*, 2026.

Moon Ye-Bin, Roy Miles, Tae-Hyun Oh, Ismail Elezi, and Jiankang Deng. Retouchllm: Training-free code-based image retouching with vision language models, 2025. URL https://arxiv.org/abs/2510.08054.

Zhiyuan You, Xin Cai, Jinjin Gu, Tianfan Xue, and Chao Dong. Teaching large language models to regress accurate image quality scores using score distribution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14483–14494, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

Huanjing Yue, Yijia Cheng, Yan Mao, Cong Cao, and Jingyu Yang. Recaptured screen image demoiréing in raw domain. *IEEE Transactions on Multimedia*, 25:5589–5600, 2022.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12753–12763, June 2025.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023a.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Ling Zhang, Yinghao He, Qing Zhang, Zheng Liu, Xiaolong Zhang, and Chunxia Xiao. Document image shadow removal guided by color-aware background. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1818–1827, June 2023b.

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

Shuo Zhang and Xinyu Yang. RetouchAgent: Towards interactive and explainable image retouching with MLLM agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35): 29901–29910, March 2026. doi: 10.1609/aaai.v40i35.40237. URL https://ojs.aaai.org/index.php/AAAI/article/view/40237.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026.

Chengxuan Zhu, Qingnan Fan, Qi Zhang, Jinwei Chen, Huaqi Zhang, Chao Xu, and Boxin Shi. Bokehdiff: Neural lens blur with one-step diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9508–9518, 2025.

Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong V Wang, James Zou, et al. 4kagent: agentic any image to 4k super-resolution. *arXiv preprint arXiv:2507.07105*, 2025.

A Related Work and Task Scope

A.1 Automatic Enhancement and Editing Agents

Automatic photographic enhancement requires choosing useful adjustments as well as executing them. Conventional retouching methods study tonal mapping, contrast enhancement, and color correction through image statistics or learned input-output mappings (Bychkovsky et al., 2011b; Rahman et al., 1996; Sen & Pal, 2010). Deep models extend this line of work through supervised photo adjustment, adversarial learning, and efficient learned image transformations (Chen et al., 2018; Deng et al., 2018; Yan et al., 2016; Gharbi et al., 2017). These methods establish important approaches to improving photographic appearance, while language-based systems make editing decisions and operations accessible through natural-language interaction.

Recent MLLM-based systems connect visual analysis with editing plans, executable code, or tool operations (Dutt et al., 2025; Ye-Bin et al., 2025; Lin et al., 2025a; Chen et al., 2025). Iterative agents further use reflection and visual feedback to revise their decisions (Zhang & Yang, 2026; Yao et al., 2026), while learning-based approaches refine editing policies and evaluation signals (Lin et al., 2025b; Wu et al., 2026). Related agent frameworks also study broader image-editing and restoration workflows (Ye et al., 2026; Zuo et al., 2025). These systems provide comparison points for both automatic edit selection and the execution of photographic adjustments.

SmartPhotoCrafter conditions the Artist on the Critic's contextual representations and learns their correspondence with visual changes during adaptation. The distinction is the interface through which photographic assessment reaches the generator. Comparisons with complete editing systems evaluate the resulting enhancement behavior, while the interface ablation and qualitative examples compare textual cascading with latent conditioning (Table 2 and Appendix F).

A.2 Generative Editing, Restoration, and Retouching

Diffusion and latent generative models provide a foundation for high-quality image synthesis and editing (Ho et al., 2020; Song et al., 2021; Rombach et al., 2022; Podell et al., 2024; Labs, 2024). Instruction-conditioned editors combine image content with language to specify visual transformations, including Instruct-Pix2Pix, Step1X-Edit, Qwen-Image-Edit, and OmniGen2 (Brooks et al., 2023; Liu et al., 2025; Wu et al., 2025a;b). These frameworks support diverse editing requests and motivate the use of a shared generator for photographic transformations. Camera-oriented models further investigate control over photographic attributes such as tone, camera settings, and depth of field (Qin et al., 2025; Yang et al., 2025; Fortes et al., 2025).

Restoration and retouching emphasize different aspects of photographic improvement. Restoration recovers visibility and detail from degraded inputs; FoundIR, MoCE-IR, and AdaIR provide relevant restoration baselines (Li et al., 2025; Zamfir et al., 2025; Cui et al., 2025). Retouching adjusts exposure, contrast, color, and related attributes to improve appearance, with paired photographic datasets and learned enhancement methods supporting this task (Bychkovsky et al., 2011a; Chen et al., 2018; Gharbi et al., 2017). In real photographs, degradation removal and appearance adjustment can be needed together. SmartPhotoCrafter therefore trains a shared Artist on restoration, retouching, and their combinations, then adapts it to assessments inferred by the Critic. Executable editing targets establish the connection between those assessments and their visual effects, and outcome-aligned reinforcement learning evaluates the generated results.

A.3 Quality Assessment as Conditioning and Feedback

Image quality assessment supplies perceptual criteria for evaluation and enhancement. Full-reference measures compare an output with a reference through pixel, structural, or feature differences (Hore & Ziou, 2010; Wang et al., 2004; Zhang et al., 2011; 2018). Traditional no-reference approaches estimate quality from natural-scene statistics (Mittal et al., 2012b;a), while learned predictors capture quality and aesthetic judgments using neural representations (Su et al., 2020; Ke et al., 2021; Talebi & Milanfar, 2018). These approaches provide complementary measurements of the photographic result.

Multimodal language models extend quality assessment to textual descriptions, instruction-based analysis, and score prediction (Wu et al., 2024a;b; You et al., 2025). Reasoning-oriented approaches further connect scoring with degradation analysis or relative quality judgments (Li et al., 2026; Cai et al., 2025; Wu et al., 2025c). Such outputs can identify the attributes responsible for a low score and provide context for deciding which edits to apply.

SmartPhotoCrafter uses photographic assessment both to condition generation and to provide editing-specific training feedback. Its rewards evaluate directional compliance, attribute-wise improvement, and perceptual consistency, with separate feedback for the Critic's suggestions and the Artist's generated images. External quality predictors, reference-based metrics, and human ratings evaluate the resulting system; the Critic's own scores serve as internal training signals. Appendix D details the evaluation criteria and the interpretation of reference-based comparisons.

B Policy Optimization Objectives

Photographic Artist. Let $c = (X, H_c)$ denote the editing condition and x_0 a generated sample from the behavior Artist policy. DiffusionNFT maps the photographic reward to an optimality probability $p \in [0, 1]$ and minimizes:

$$\begin{aligned}\mathcal{L}_{\mathrm{PA}}(\phi) = \mathbb{E}_{c,x_0\sim\pi_{\mathrm{old}}(\cdot|c),t}\Big[p\,\|v^+_\phi(x_t,c,t) - v\|_2^2 \\ +(1-p)\,\|v^-_\phi(x_t,c,t) - v\|_2^2\Big],\end{aligned} \tag{11}$$

where x_t and v denote the perturbed sample and target velocity; the implicit fields are:

$$\begin{aligned}v^+_\phi(x_t,c,t) &= (1-\beta_{\mathrm{NFT}})v^{\mathrm{old}}(x_t,c,t) + \beta_{\mathrm{NFT}}v_\phi(x_t,c,t),\\ v^-_\phi(x_t,c,t) &= (1+\beta_{\mathrm{NFT}})v^{\mathrm{old}}(x_t,c,t) - \beta_{\mathrm{NFT}}v_\phi(x_t,c,t).\end{aligned} \tag{12}$$

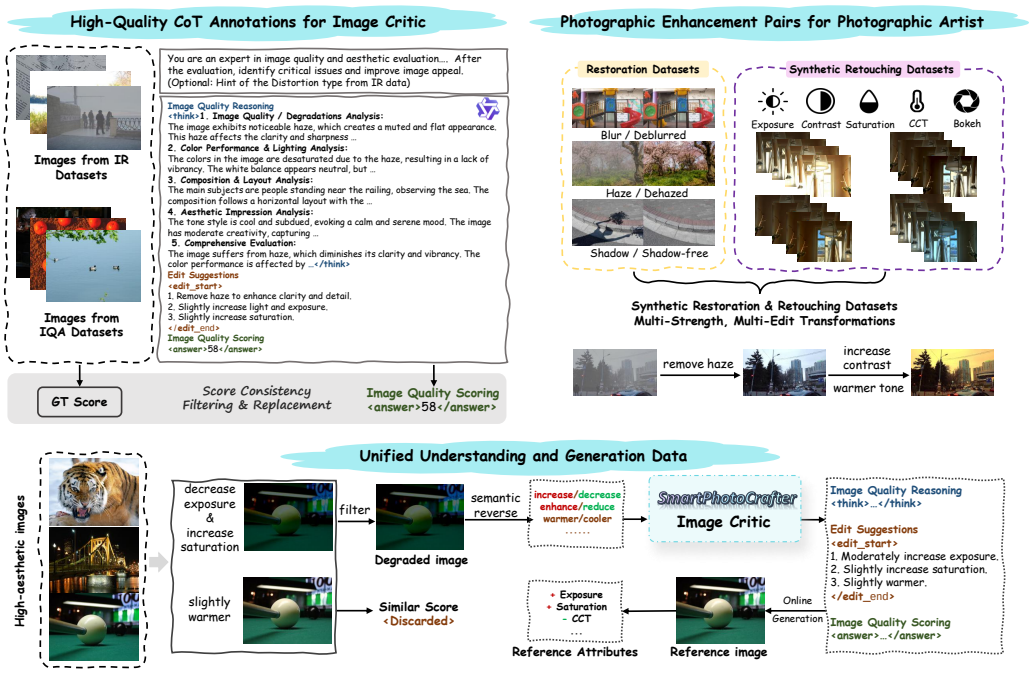


Figure 4: Data construction pipeline of SmartPhotoCrafter. The pipeline constructs photographic analyses, edit suggestions, and quality-score annotations for the Image Critic; paired restoration, retouching, and combined transformations for the Photographic Artist; and assessment-conditioned editing examples for subsequent adaptation and optimization.

Image Critic. Given an input X, sample a group of G responses $\{o_i\}_{i=1}^{G}$ from the behavior policy $\pi_{\theta_{\text{old}}}$. Let $r_i = r_{\text{IC}}(o_i)$ denote the reward associated with response i, with dependence on the input, reference, and resulting feedback omitted for brevity. The group-relative advantage is:

$$A_i = \frac{r_i - \mu_r}{\sigma_r}, \tag{13}$$

where μ_r and σ_r denote the group's mean reward and standard deviation. GRPO maximizes:

$$J_{\text{IC}}(\theta) = \mathbb{E}_{X,\{o_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(\rho_i(\theta)A_i, \text{clip}(\rho_i(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}})A_i\Big) - \beta_{\text{KL}} D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\Bigg], \tag{14}$$

where ρ_i is the importance-sampling ratio, ϵ_{clip} controls clipping, and β_{KL} weights the KL penalty.

The coefficients β_{KL} and β_{NFT} govern reference-policy regularization and velocity-field interpolation, respectively. GRPO updates the response policy with the Critic reward, while DiffusionNFT updates the generative policy with the Artist reward.

C Data Construction and Supervision

The data pipeline serves three learning requirements: assessing photographic deficiencies, realizing controllable transformations, and connecting the assessment to the Artist's conditioning interface. Figure 4 summarizes the construction of Critic annotations, restoration-retouching pairs, and data for connecting the two modules. We describe the supervision by its role and provenance. The sample counts below describe the supervision used at each training stage.

C.1 Critic Supervision: Assessment, Suggestions, and Scores

The Critic's foundation data combine IQA images from KonIQ-10K, SPAQ, and KADID-10K with restoration and editing examples (Hosu et al., 2020; Fang et al., 2020; Lin et al., 2019). Editing-related sources include GoPro, FoundIR, and ISTD, exposing the Critic to different distortion types (Nah et al., 2017; Li et al., 2025; Wang et al., 2018). The Critic also uses the restoration sources listed in the following Artist-supervision subsection to construct distortion-aware annotations; KADID-10K supplies additional distortion labels. Qwen2.5-VL-72B generates photographic analyses, editing suggestions, and quality scores. Where a mean opinion score is available, it replaces the generated score. For editing samples, score-token losses are masked, so automatically generated scores do not become unrestricted scoring supervision. Restoration annotations also provide distortion categories; samples whose generated analysis conflicts with those categories are filtered.

Bokeh-oriented retouching examples additionally teach the Critic to assess whether background blur is appropriate for the scene and subject. The full annotation prompt is provided in Table 5.

The structured response contains reasoning, suggestions, and a score. Its assessment covers technical degradation, color and lighting, composition, and aesthetic impression. The breadth of this analysis should not be confused with the generator's supported action space: discussing composition does not establish cropping or viewpoint-editing capability. Annotation quality depends on the teacher and filtering process, and is not equivalent to expert verification of every generated statement.

C.2 Artist Supervision: Single and Combined Transformations

The Artist uses paired restoration data with task-specific instructions such as blur, haze, or moiré removal. The restoration sources include FoundIR for deblurring, dehazing, and low-light enhancement; RealBlur for deblurring; TMM22 for moiré removal; LOL and LOL-v2 for low-light enhancement; and ISTD, RDD, and SRD for shadow removal (Li et al., 2025; Rim et al., 2020; Yue et al., 2022; Wei et al., 2018; Yang et al., 2021; Wang et al., 2018; Zhang et al., 2023b; Qu et al., 2017). Color- and tone-related retouching pairs are generated from FilmSet photographs using parameterized exposure, contrast, saturation, and correlated color temperature adjustments at multiple strengths (Li et al., 2023). Depth-of-field supervision uses the available multi-level blur pairs from RealBokeh. For BokehDiff, we select all-in-focus images and use a pretrained bokeh model to synthesize variants at different blur strengths, forming additional paired examples (Seizinger et al., 2025; Zhu et al., 2025). These data teach controllable editing before latent-conditioned adaptation.

Combined examples apply retouching transformations to restoration pairs, expanding training from individual operations to coupled restoration-retouching edits. Random color and tone adjustments augment restoration examples with multi-edit instructions, allowing the Artist to learn how degradation removal interacts with photographic appearance changes.

C.3 Interface-Grounding and Reference-Guided Refinement

Stage II retains paired restoration and retouching examples and adds FiveK pairs and an AVA-derived set. High-MOS AVA photographs serve as references; synthetic exposure, contrast, saturation, and color-temperature perturbations construct paired inputs (Bychkovsky et al., 2011a; Murray et al., 2012). These are paired synthetic examples, not reference-free preference observations.

For supported retouching suggestions, the Critic also produces an online target through rule-based execution: $X_{\mathrm{ref}} = \mathcal{T}(X; E)$. The Artist is trained under the associated H_c to reproduce the operator-produced effect. This online target is distinct from the dataset reference X_{gt}. In Stage III, the Artist's generated image X_e is evaluated with the photographic reward, while the Critic's suggestion-effect reward reuses the attribute-wise criterion on X_{ref}. The Artist's generated outputs also enter the Critic's ranking feedback, connecting the two policies while keeping the supervision targets distinct.

Table 4: Training samples and supervision across the three stages.

Stage / module	Samples	Supervision
I / Critic	≈80K	Image analysis, suggestions, and quality scores
I / Artist	≈160K	Restoration and single-/multi-edit paired targets
II / Artist	≈30K	Paired targets and online executable-edit targets
III / both modules	≈18K	Reference-guided photographic feedback

Counts refer to examples used within each stage; source photographs may recur across stages.

C.4 Photographic Annotation Structure

The annotation prompt organizes assessment into technical degradation, color and lighting, composition, and aesthetic impression. It then requests concise editing suggestions and an overall quality score. The suggestion guidelines focus on major quality issues and explicitly ask the teacher to disregard minor deviations and intentional artistic choices. This structure provides photographic context; the Artist supports restoration and retouching operations.

The response places analysis in `<think>...</think>`, suggestions in `<edit_start>...</edit_end>`, and the quality score in `<answer>...</answer>`. Scores use a 0-100 scale. For IQA samples, the dataset MOS replaces the generated score; for editing samples, the score-token loss is masked. This lets the Critic learn reasoning and suggestions without treating teacher-generated scores as independent quality labels.

D Experimental Settings and Evaluation Criteria

D.1 Evaluation Tasks and Data

We follow existing training/test splits when available; otherwise, we divide the source photographs into training and test sets at a 9:1 ratio. Source photographs and their derived variants are kept within the same split to prevent overlap between training and evaluation. The automatic enhancement and prescribed multi-edit test sets each contain approximately 300 examples drawn from multiple datasets. For FiveK, we use the Expert-C retouches as references.

The automatic enhancement suite combines photographic retouching, synthetic appearance recovery, and technical restoration. FiveK provides retouching pairs (Bychkovsky et al., 2011a), while high-aesthetic AVA images are perturbed in exposure, contrast, saturation, and color temperature to form paired inputs and references (Murray et al., 2012). The restoration portion uses degraded photographs with clean targets. In this setting, SmartPhotoCrafter infers the edit suggestions from the input under a common enhancement request.

The combined-edit test set is constructed from blur- and haze-affected examples with randomly sampled retouching operations. Each test instruction specifies degradation removal together with the selected photographic adjustments. For example, an instruction can request blur removal together with a decrease in exposure and an increase in saturation. This setting measures whether a prescribed combination is realized accurately, whereas automatic enhancement additionally requires selecting the changes. These tasks share transformations but evaluate different capabilities.

D.2 Training Configuration

Table 6 summarizes the optimization schedule. Stage I performs independent task-specific supervised pretraining of the Critic and Artist; Stage II freezes the Critic and adapts the Artist to its contextual representations. Stage III uses separate policy optimizers in the same training loop. Training is conducted on eight NVIDIA A100 GPUs, and Stage III takes approximately three days.

D.3 Evaluation Metrics

Q-Align, MUSIQ, and NIMA estimate image quality and aesthetic appeal without a paired reference (Wu et al., 2024b; Ke et al., 2021; Talebi & Milanfar, 2018). Q-Align uses multimodal language

Table 5: Full prompt for photographic assessment, editing suggestions, and scoring.

Component	Prompt content
System Prompt	You are an expert in image quality and aesthetic evaluation. Your task is to perform a comprehensive, multi-level analysis of the given image. For every judgment, follow a detailed, step-by-step chain-of-thought reasoning. Ensure your analysis is structured, logical, and fully justified using professional terminology and clear reasoning. After the evaluation, identify critical issues that should be addressed to improve image appeal.
Quality / Degradations Analysis	*Note: These attributes negatively affect image perception and are mostly interference elements that need removal.* 1. Blur / Clarity: motion blur, defocus, etc. 2. Noise: digital noise, compression artifacts, etc. 3. Distortion and Artifacts: stretching, moiré, lens distortion, etc. 4. Visual Interference: glare, shadow occlusion, weather effects, etc.
Color Performance / Lighting Analysis	*Note: These are basic low-level attributes that determine quantifiable visual information*. 1. Saturation 2. White balance / Color temperature 3. Contrast 4. Exposure / Light 5. Dynamic Range 6. Shadow quality (some shadows may be intentional artistic design)
Composition Analysis	*Note: These are technical factors related to static composition and camera parameter settings, crucial for aesthetic evaluation.* 1. Composition and Layout 2. Focus 3. Camera viewpoint & horizon 4. Depth of field (Bokeh-like Effect, Background Blur Level) 5. Photography style (cinematic, portrait, landscape, etc.)
Aesthetic Impression Analysis	*Note: These high-level attributes reflect perceptual and semantic aesthetic impressions.* 1. Tone style (vibrant, warm, cool, filmic, etc.) 2. Creativity 3. Emotional expression and storytelling 4. Semantic richness 5. Environment and background integration 6. Context adaptability
Comprehensive Evaluation	Summarize the above analysis and provide concrete improvement suggestions. Clearly explain the reasoning and logic. Structure the output into numbered sections: 1. Image Quality / Degradations Analysis + distortion severity 2. Color Performance & Lighting Analysis 3. Composition Analysis 4. Aesthetic Impression Analysis 5. Concrete and concise suggestions for image improvement Wrap everything in **<think>...</think>**. After that, identify the most severe issues and summarize practical, feasible suggestions in **<edit_start>...</edit_end>**. Finally, provide the overall quality and aesthetic score (0-10 score range, here mapped to 0-100) in **<answer>...</answer>** without extra words.
Improvement Suggestions Guidelines	Provide concise, practical, and feasible improvement suggestions, organized into numbered categories. Focus only on major quality issues that significantly degrade image perception or aesthetic appeal. Ignore minor deviations or intentional artistic choices.

Table 6: Stage-specific AdamW learning rates and optimization objectives.

Stage	Critic learning rate	Artist learning rate	Objective
I	10^{-5}	10^{-5}	Independent supervised pretraining
II	Frozen	10^{-5}	Reasoning-conditioned adaptation
III	10^{-5}	3×10^{-4}	Outcome-aligned RL

modeling to predict text-defined quality levels. DINO and CLIP compare feature representations, LPIPS evaluates perceptual distance, and FID measures distributional discrepancy from the reference set (Caron et al., 2021; Zhang et al., 2018; Heusel et al., 2017). PSNR and SSIM measure reconstruction agreement, while DISTS provides a complementary structure-and-texture similarity measure for restoration (Wang et al., 2004). These metrics assess complementary output properties.

For restoration and prescribed editing, reference-based metrics evaluate agreement with the target transformation. For retouching, high-quality references provide examples of the desired photographic appearance. We assess agreement with these references alongside no-reference quality scores and human ratings of photorealistic aesthetics and input consistency.

E Technical Image Restoration Analysis

Table 7: Restoration results for deblurring and dehazing. Bold values indicate the best result for each metric within the corresponding task.

Method	Deblurring					Dehazing				
	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	DISTS↓	PSNR↑	SSIM↑
FLUX.2-Dev (Labs, 2025)	0.20	53.38	0.12	17.18	0.60	0.09	36.99	0.07	21.17	**0.88**
Qwen-Image-Edit (Wu et al., 2025a)	0.11	33.71	0.07	24.53	0.73	0.23	56.92	0.15	14.35	0.38
FoundIR (Li et al., 2025)	0.16	50.12	0.11	**25.92**	**0.77**	0.13	39.81	0.11	17.73	0.82
MoCE-IR (Zamfir et al., 2025)	0.21	50.68	0.13	21.41	0.71	0.16	39.34	0.12	15.12	0.72
AdaIR (Cui et al., 2025)	0.19	51.35	0.12	22.83	0.74	0.19	44.28	0.15	15.99	0.74
SmartPhotoCrafter	**0.07**	**21.85**	**0.05**	24.73	0.76	**0.05**	**17.23**	**0.05**	**24.08**	0.87

Deblurring aggregates FoundIR-Blur, RealBlur, and GoPro; dehazing uses FoundIR-Haze. Each task is evaluated with the same five restoration metrics.

We further evaluate the restoration capability of SmartPhotoCrafter on two representative low-level image restoration tasks, including deblurring and dehazing. As shown in Table 7, on deblurring, SmartPhotoCrafter achieves the lowest LPIPS, DISTS, and FID among the tested methods, although FoundIR obtains higher PSNR and SSIM (25.92 dB and 0.77 versus 24.73 dB and 0.76). On dehazing, SmartPhotoCrafter leads in PSNR and the three reference-relative perceptual/distributional metrics, while FLUX.2-Dev has slightly higher SSIM (0.88 versus 0.87). These results suggest that, beyond automatic photographic enhancement, SmartPhotoCrafter also demonstrates competitive performance under classical image restoration scenarios.

F Text-Mediated and Latent-Conditioned Generation

The text-mediated cascade serializes the Critic's suggestions as an instruction for the Artist. SmartPhotoCrafter instead conditions generation on the Critic's contextual representations, after adapting the Artist to their associated visual effects. Both routes are automatic when the suggestions are inferred from the input photograph. Figure 5 compares how these routes translate the assessment into visual changes.



Suggestion: decrease light&exposure



Suggestion: slightly increase contrast



Suggestion: moderately increase light&exposure; slightly increase contrast; slightly increase saturation

Figure 5: Qualitative comparison between Instruction-based and Automatic Photographic Enhancement. The 'Instruction-based' results are generated via a sequential cascaded pipeline, which first analyzes the input image to derive editing instructions, and then generates the enhanced result conditioned on these textual instructions.

Visual effects of the conditioning interface. Figure 5 illustrates how the two routes realize the Critic's suggestions. In the bird example, both routes reduce exposure, but they produce different brightness levels in the plumage and background. The autumn scene shows different strengths of contrast and color adjustment under the same suggestion to slightly increase contrast. In the coastal scene, the combined request for increased exposure, contrast, and saturation produces changes in the visibility of the rocks and trees, with different tonal balances across the two outputs.

These examples illustrate differences in edit strength and spatial effects under shared suggestions. Table 2 complements this comparison with the Stage II ablation of the two conditioning routes and their associated Artist adaptation.

G More qualitative results

Automatic photographic enhancement. Figure 6 shows automatic enhancement across haze-affected, dark, and color-rich scenes. In the bridge and woodland examples, the outputs reduce the washed-out appearance and increase contrast, making the bridge structure, tree trunks, and foreground more distinct. The woodland result also retains the warm sunlight. In the squirrel and child examples, adjustments to exposure and contrast improve subject visibility and separation from the background while preserving the subject's pose and surrounding scene.

Across these examples, the changes depend on the input: haze-affected scenes benefit from visibility recovery, whereas darker photographs mainly require tonal adjustment. The results illustrate how restoration and retouching can work together without replacing the main scene content. They

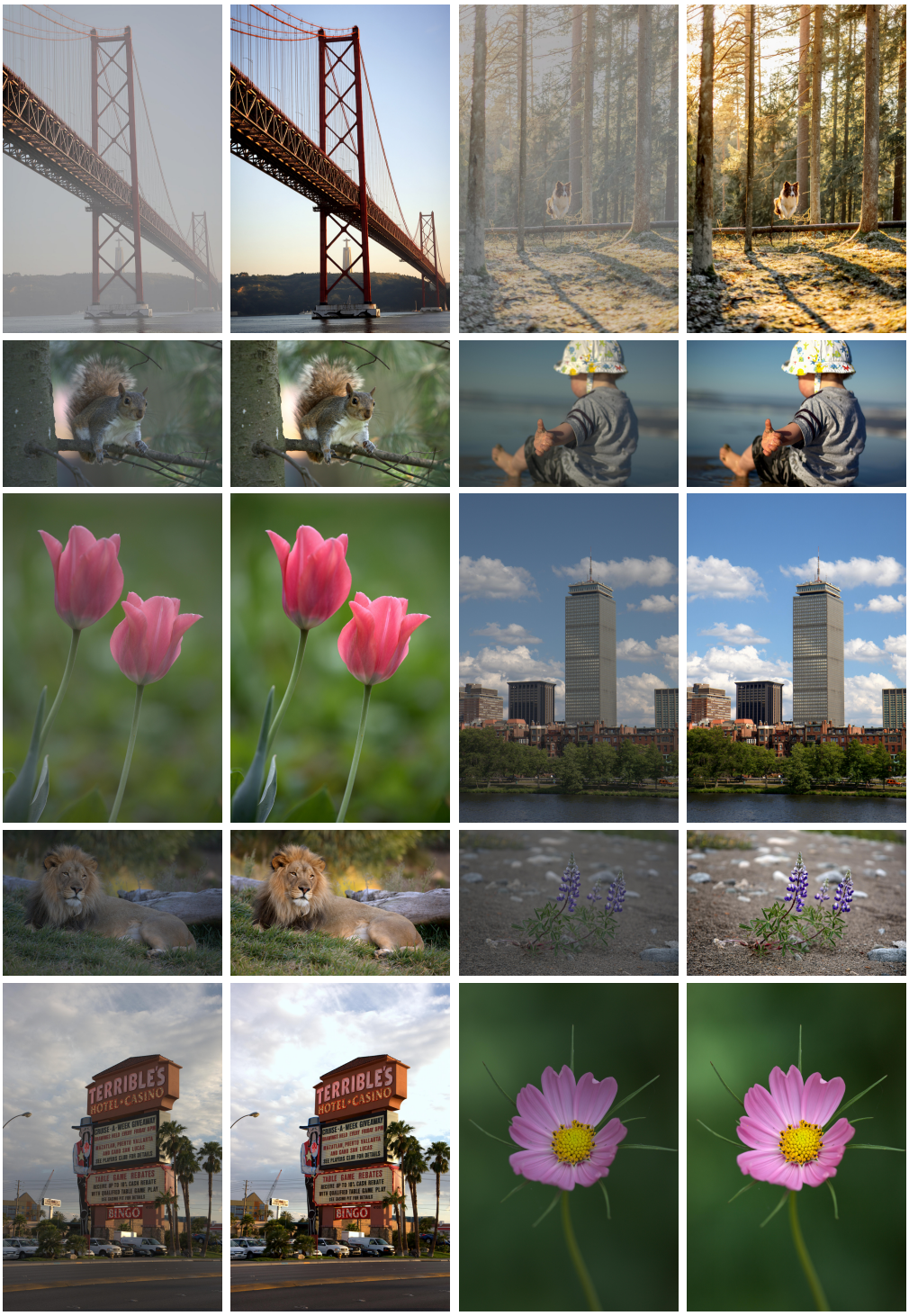


Figure 6: More qualitative results on automatic photographic enhancement.

complement the aggregate quality and reference-consistency scores by showing which photographic attributes change and how those changes affect the appearance of the subject and background.

Combined restoration and retouching. Figure 7 shows SmartPhotoCrafter outputs under prescribed combinations of exposure, saturation, color-temperature, contrast, bokeh, and haze-removal

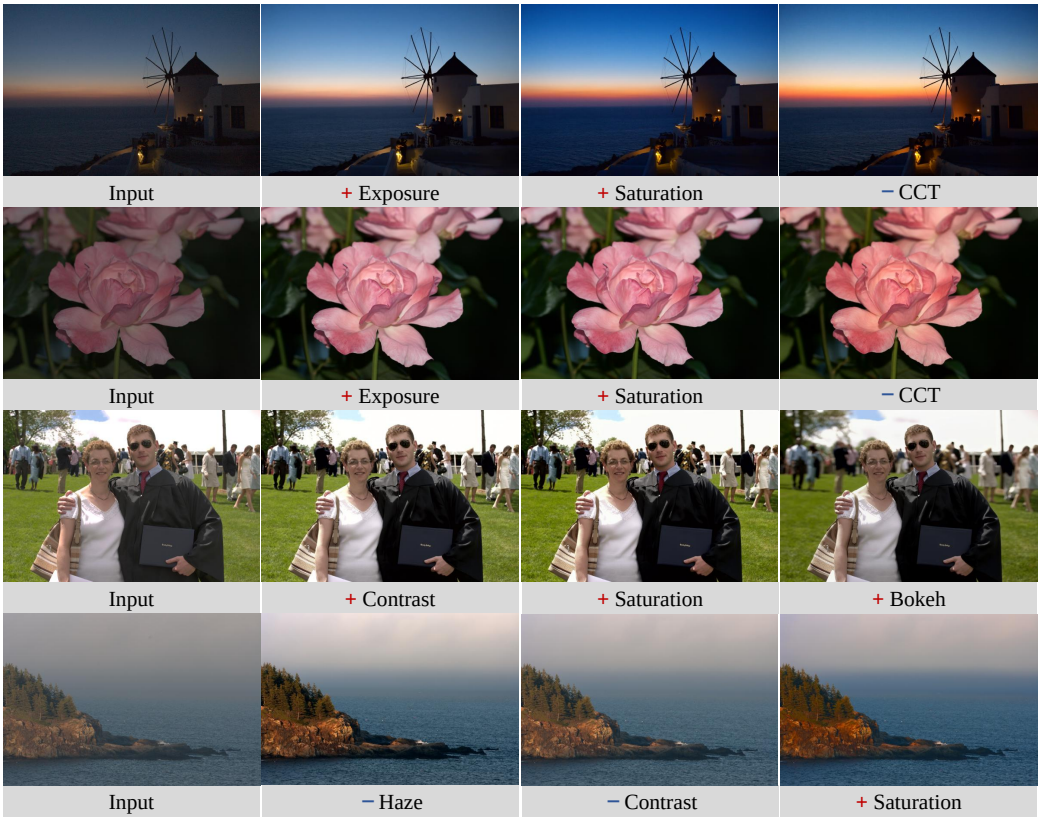


Figure 7: Examples of combined photographic adjustments generated by SmartPhotoCrafter. Each row starts from the input and illustrates the changes associated with the indicated editing operations.

operations. The sequences illustrate how multiple adjustments affect the same photograph: tonal and color changes alter brightness and appearance, while background blur and haze removal change subject separation and visibility. The main objects and spatial arrangement remain recognizable across the illustrated edits.

These examples complement the combined-edit results in Table 3 by showing the effects of prescribed operations; automatic enhancement additionally requires the Critic to choose the adjustments.

H Evaluation Protocol of User Study

We recruit 35 participants to evaluate enhancement results for 20 input photographs from SmartPhotoCrafter, JarvisEvo (Lin et al., 2025b), FLUX.2-Dev (Labs, 2025), and Nano Banana 2 (Google, 2026). Each participant completes 40 questions covering two criteria for each photograph: photorealistic aesthetics and image consistency. Each question presents the four model outputs for independent rating on a scale from 1 (lowest) to 5 (highest), with higher scores indicating better performance. Outputs are presented in random order with model names hidden.

For photorealistic aesthetics evaluation, participants view the four enhanced images without the input photograph and assess whether each result looks both realistic and visually appealing, considering visible degradations, tonal and color quality, and overall naturalness.

For image consistency evaluation, Participants are additionally shown the input photograph as a reference and assess how faithfully each output preserves its content, including object identity, shape, spatial layout, scene textures, etc. Changes needed for restoration and retouching are allowed, such as removing blur or haze and adjusting tone or color. The assessment focuses on unnecessary alterations to scene content rather than differences inherent to needed corrections.

I CASE STUDIES

Figures 8–10 present full Critic responses, input photographs, and Artist outputs for images of varying quality. Correct observations are highlighted in blue; inaccurate or suboptimal observations and suggestions are marked in red.



(a) Input Image (b) Enhanced Image

Figure 8: Ground-truth Score: 80.65

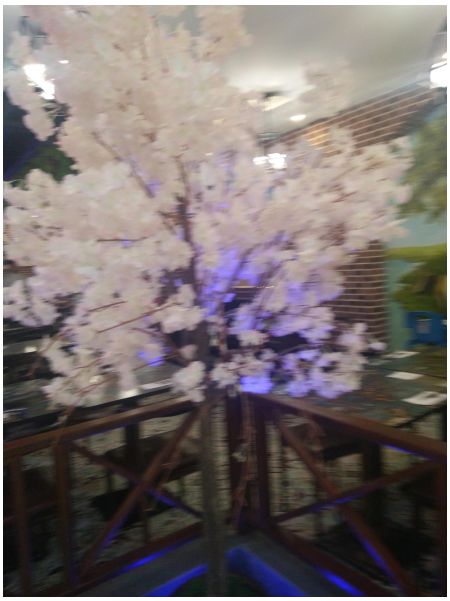
Image Quality&Aesthetic Reasoning
1. Image Quality/Degradations Analysis: The image exhibits no significant motion blur, defocus, or distortion. However, there is noticeable low exposure, particularly in the darker areas surrounding the dog. This creates a lack of detail in the background and could be considered a form of visual interference due to shadow occlusion. There is no apparent digital noise or compression artifacts visible at this resolution. 2. Color Performance&Lighting Analysis: The saturation appears moderate, with natural-looking colors. The white balance seems accurate, as the dog's fur tones look true-to-life. However, the contrast is somewhat subdued due to the low exposure, leading to less distinction between the subject and its surroundings. The lighting is focused on the dog, creating a dramatic effect, but it also results in some underexposed regions. Shadows are intentionally used to create depth, but they obscure details in certain areas. 3. Composition&Layout Analysis: The main subject, the dog, is clearly defined and positioned slightly off-center, adhering to the rule of thirds. The focus is sharp on the dog, drawing attention directly to it. The camera viewpoint is at eye level, providing a natural perspective. The background blur is effective in isolating the subject, though it might be slightly excessive, causing some loss of environmental context. The photography style leans towards a cinematic approach, emphasizing mood over clarity. 4. Aesthetic Impression Analysis: The tone style is somewhat muted due to the low exposure, giving the image a contemplative and serene feel. The creativity lies in the use of lighting and shadow to evoke emotion. The emotional expression is calm and introspective, with the dog appearing relaxed yet alert. The semantic richness is moderate, as the image tells a simple story of a quiet moment. The environment integrates well with the subject, enhancing the overall mood. The context adaptability is good within the intended artistic style. 5. Comprehensive Evaluation: The image has a strong aesthetic appeal due to its effective use of lighting and composition. However, the low exposure affects the visibility of details and could be improved. The color performance is generally good, but increasing the light and exposure would enhance the vibrancy. The composition is strong, with a clear focus on the subject, but reducing the background blur slightly could add more context. The overall tone and emotional expression are compelling, making the image engaging despite the technical issue.
Edit Suggestions
1. moderately increase light&exposure; 2. slightly decrease the depth of field (enhance the bokeh effect)
Image Quality Scoring
Predicted Score: 82/100



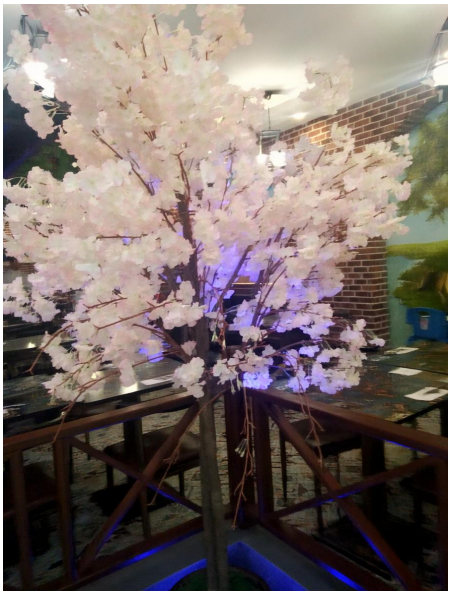
(a) Input Image (b) Enhanced Image

Figure 9: Ground-truth Score: 66.19

Image Quality&Aesthetic Reasoning
1. Image Quality/Degradations Analysis: The image appears relatively free from significant distortions such as motion blur, digital noise, or lens distortion. However, there is a slight overexposure affecting the glass and its contents, causing some loss of detail in brighter areas. There is no noticeable glare, flare, or shadow occlusion impacting the main subject. The distortion severity is moderate due to the overexposure. 2. Color Performance&Lighting Analysis: The saturation seems adequate, providing a natural appearance to the drink and glass. The white balance appears slightly off, leaning towards a cooler tone, which might not accurately represent the warmth of the beverage. The contrast is moderate but could be enhanced for better differentiation between the drink and the background. The exposure issue mentioned earlier affects the brightness levels, making parts of the image appear washed out. 3. Composition&Layout Analysis: The main subject, the glass of drink, is centered and clearly visible, adhering to standard composition rules. The focus is sharp on the glass, ensuring it remains the primary point of interest. The camera viewpoint is straightforward, capturing the subject at eye level. The depth of field is sufficient, keeping both the foreground and background in focus. The photography style leans toward a casual, everyday capture rather than a specific artistic approach. 4. Aesthetic Impression Analysis: The tone style is somewhat neutral, lacking vibrancy or warmth that could enhance the inviting nature of the drink. The creativity is minimal, focusing solely on presenting the drink without additional context or storytelling elements. The environment and background are simple and unobtrusive, integrating well with the subject. The overall impression is functional but lacks emotional engagement. 5. Comprehensive Evaluation: To improve the image's appeal, addressing the overexposure and adjusting the color temperature would significantly enhance the visual experience. Increasing the contrast and slightly warming the tones can make the drink more visually appealing. Adjusting the depth of field to create a subtle bokeh effect in the background could further draw attention to the main subject. Transforming the photography style to something like "Food" could add a professional touch.
Edit Suggestions
1. decrease light&exposure; 2. change color temperature to slightly warmer 3. increase contrast 4. slightly increase the depth of field (reduce the bokeh effect)
Image Quality Scoring
Predicted Score: 72/100



(a) Input Image



(b) Enhanced Image

Figure 10: Ground-truth Score: 37.18

Image Quality&Aesthetic Reasoning
1. Image Quality/Degradations Analysis: The image suffers from significant motion blur, which affects the clarity and sharpness of the entire scene. There is no noticeable noise or compression artifacts, but the blur makes it difficult to discern fine details. No evident distortion or visual interference is present. 2. Color Performance&Lighting Analysis: The colors appear somewhat muted due to the blur, affecting saturation levels. The white balance seems slightly off, leaning towards cooler tones, possibly due to artificial lighting. The contrast is low, contributing to the lack of depth. The exposure is adequate, but the lack of sharpness diminishes the impact of lighting. 3. Composition&Layout Analysis: The main subject appears to be a decorative tree with white blossoms, but the blur obscures its details. The composition lacks a clear focal point due to the lack of sharpness. The camera viewpoint seems to be at eye level, but the blur detracts from any intended artistic effect. The depth of field is shallow, but this is overshadowed by the motion blur. 4. Aesthetic Impression Analysis: The tone style leans towards a cool, subdued palette, but the blur prevents the image from conveying a strong emotional expression or storytelling. The environment and background are not clearly integrated due to the lack of detail. The overall impression is one of unintentional abstraction rather than deliberate artistic choice. 5. Comprehensive Evaluation: The image suffers from severe motion blur, which is the primary issue affecting its quality and aesthetic appeal. The composition lacks a clear focal point, and the intended artistic style is obscured by the blur. To improve the image, addressing the blur and enhancing the color performance would be essential steps.
Edit Suggestions
1. remove blur 2. slightly increase contrast 3. moderately increase the depth of field (reduce the bokeh effect)
Image Quality Scoring
Predicted Score: 37/100