Dotting the Eye: An Intent-Driven Image Retouching Agent for Visual Focus Enhancement

Chujie Qin[1,2*], Zilong Zhang[1,2*], Zewei Chang[1,2], Chunle Guo[1,2], Ruixing Wang[4†], Tao Hu[4], Ming-Ming Cheng[1,2,3], and Chongyi Li[1,2‡]

[1] VCIP,CS,Nankai University
[2] NKIARI, Shenzhen Futian
[3] AAIS, Nankai University
{chujie.qin,zhangzilong}@mail.nankai.edu.cn
{guochunle, cmm, lichongyi,}@nankai.edu.cn
[4] DJI Technology Co., Ltd
ruixingw@hustunique.com
hubert.hu@dji.com

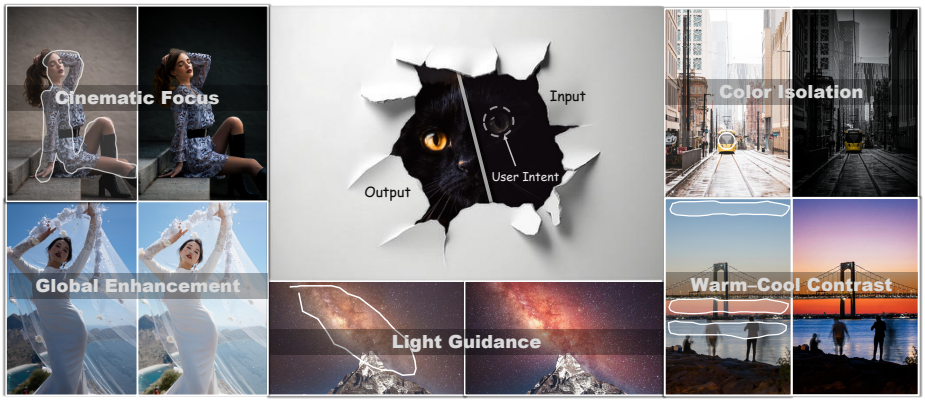

Fig. 1: EyeControl focuses on highlighting the user-intended visual focus through image retouching. With just a few clicks or casual strokes, EyeControl can deliver professional retouching looks—*Cinematic Focus*, *Color Isolation*, *Light Guidance*, *Warm–Cool Contrast*, *Global Enhancement*—and, of course, please dot the eye to make the cat "break through" the background wall and pop forward as the new visual focus.

Abstract. Image retouching is commonly formulated as enhancing overall visual quality through color adjustment, but in practice, it also serves to emphasize visual focus by guiding viewers' attention toward a specific subject or region. Achieving such focus-oriented retouching is inherently challenging, as it requires well-coordinated global and local adjustments to manipulate perceptual saliency while maintaining visual naturalness.

*Equal Contribution. This project is done during Chujie Qin's internship at DJI.
†Corresponding Author.
‡Project Lead.

This intricate process typically demands substantial professional expertise. In this study, we propose **EyeControl**, a multi-modal large language model (MLLM)-driven agent with a diffusion-based retouching executor that enables visual focus enhancement under weak user intent. With only a few clicks or coarse strokes, EyeControl directs visual attention to the intended region, effectively "dotting the eye" of the image. The core idea is to explicitly link the weak user intention with the target editing region and the corresponding tonal adjustment operations during retouching. To achieve this, the system first interprets the intent and image content to infer the visual focus and generate structured intent guidance for the retouching executor. Second, the retouching executor is encouraged to respond more strongly to the target region, explicitly aligning its attention map with a designed pseudo-intent map. We also introduce an operation-consistency constraint to improve coordination between global and local adjustments, achieving more natural and coherent retouching. Additionally, we contribute ControlArt-Bench, a high-quality evaluation dataset for visual focus enhancement. Extensive evaluations demonstrate that EyeControl yields perceptually appealing results with stronger intent alignment. Code will be released at https://github.com/DragonisCV/EyeControl.

Keywords: Image Retouching · Diffusion Models · Saliency · Agent

1 Introduction

Image retouching has evolved from manual workflows to learning-based methods [11, 17, 26, 36] that enhance perceptual quality through tonal and color adjustments. Recent diffusion- and VLM-based models [4,6,7,21] further automate such edits via textual prompts. However, retouching also serves to guide viewers' attention toward a visual focus—a purposeful enhancement that current models fail to explicitly model, even though they support localized edits.

Building upon this observation, we study intent-driven visual focus retouching under weak spatial cues. Instead of relying solely on text prompts, we adopt sparse spatial interactions, such as clicks or coarse strokes, as intention signals. Text prompts often lack precise spatial grounding for subtle focus manipulation, while real-world editing workflows typically rely on lightweight spatial interactions to indicate regions of interest. Given an input image and a weak intention signal, the goal is to apply tonal and color adjustments that make the intended visual focus easier to notice, while preserving content and perceptual naturalness.

This problem introduces several inherent challenges. **First, spatial intention signals are sparse and ambiguous.** Spatial interactions such as clicks or coarse strokes provide only partial observations of user intention and do not directly specify the desired editing operations. Inferring the intended visual focus from such sparse signals, therefore, requires reasoning over both user interactions and image context. **Second, visual focus lacks explicit control mechanisms in existing retouching models.** While existing retouching models can apply local adjustments, they do not provide explicit ways to translate user inten-

tion into visual focus enhancement. **Third, effective visual focus enhancement requires coordinated global and local adjustments.** Enhancing the prominence of a region often requires coordinated adjustments between the focal area and its surrounding regions, rather than isolated local modifications.

To address the above challenges, we propose **EyeControl**, an intent-driven image retouching agent for visual focus enhancement. To enable controllable visual focus enhancement, we introduce Pseudo-Intention Attention Alignment (PAA), which explicitly supervises the model's internal attention using pseudo-intention labels derived from saliency difference maps. This encourages the model to align its attention with intention-driven saliency shifts, allowing focus modulation to be learned during generation rather than relying on indirect color perturbations. Finally, we introduce an operation consistency constraint to improve coordination between global and local adjustments. Specifically, the guidance for visual focus enhancement is decomposed into global and local operations, and enforces consistency between the result of applying them jointly and the results obtained by applying them sequentially (e.g., global-then-local or local-then-global). This encourages the executor to produce more natural and coherent visual focus enhancement.

In summary, our contributions are threefold:

- We present a visual-focus-centric perspective on image retouching, enabling visual focus enhancement through retouching and producing more expressive and visually appealing results beyond existing methods, as shown in Fig. 1
- We propose **EyeControl**, an intent-driven image retouching framework that integrates intention reasoning, attention-level focus modulation, and cross-scale coordination, enabling controllable and stable visual focus enhancement while preserving global coherence and perceptual naturalness.
- We introduce **ControlArt-Bench**, the first image retouching benchmark for visual focus enhancement with paired spatial user-intent annotations and focus-oriented evaluation metrics. Extensive experiments show that our method achieves state-of-the-art performance in visual focus enhancement, while remaining competitive on global and local retouching tasks.

2 Related Work

2.1 Global Image Retouching

With the release of large-scale retouching datasets [3, 19], deep learning has become a dominant approach for image retouching. Some methods model the retouching process through interpretable adjustment operators, such as tone curves [15, 16, 24, 26], bilateral grid transformations [6, 8, 17], or lookup tables (LUTs) [28,34,36], aiming to mimic traditional editing pipelines. Another line of work learns image-to-image mappings directly using convolutional networks [5], Transformers [29], or diffusion models [6]. However, most existing approaches treat retouching as a global enhancement problem. Moreover, widely used paired datasets mainly contain globally adjusted targets without explicit modeling of

user intention or region-specific emphasis [3, 19]. As a result, the learned supervision often reflects an averaged color preference rather than an intention-driven objective, causing models to converge toward globally consistent but perceptually generic solutions. In contrast, practical editing is often guided by subjective intention, where adjustments aim to reshape perceptual focus rather than merely improve overall appearance.

2.2 VLM-Driven Image Retouching

Recent advances in vision-language models (VLMs) have enabled language-driven image retouching frameworks [4, 20, 21]. By leveraging multimodal understanding, these systems interpret user instructions, analyze image content, and generate editing plans. Some approaches further employ large language models to decompose complex requests into sequential operations or parameter adjustments [7, 32]. However, these methods primarily focus on semantic reasoning and instruction following. The connection between weak spatial intention signals (e.g., clicks or coarse masks) and controlled perceptual saliency redistribution remains indirect, as spatial cues are typically translated into textual instructions or heuristic editing steps. In contrast, our approach incorporates intention signals directly into the attention mechanism of a diffusion backbone, enabling structured mask–image interaction and explicit supervision of saliency change.

2.3 Saliency Retargeting

Saliency retargeting [1, 23, 30] studies how image appearance can be modified to redirect visual attention. Early methods rely on computational saliency models and apply handcrafted adjustments—such as color, contrast, or luminance manipulation—to enhance target regions or suppress distractions [1, 10]. These works demonstrate that appearance modulation can alter perceptual ordering. However, existing saliency retargeting techniques are largely rule-based and operate outside modern generative editing frameworks. They typically assume explicit target regions and predefined transformations, and are not designed to handle weak or ambiguous intention signals. Moreover, aesthetic plausibility is rarely considered: aggressive color or contrast manipulation may increase saliency but distort object attributes and reduce perceptual realism. In contrast, we integrate saliency redistribution directly into a diffusion-based editing backbone. By modeling mask-guided attention interactions and supervising the induced saliency difference, our approach enables retouching for visual focus enhancement.

3 Method

We begin by presenting the overall workflow of EyeControl (Sec. 3.1). We then describe the proposed data generation pipeline, which constructs a high-quality dataset containing diverse samples of weak, spatially specified user intent(Sec. 3.2). Finally, we introduce the core innovations of EyeControl, including VLM-based ambiguous intention understanding and guidance, as well as the training framework for the retouching executor(Sec. 3.3). Together, these components enable intention-driven image retouching that effectively enhances the user-specified visual center under weak or ambiguous inputs.

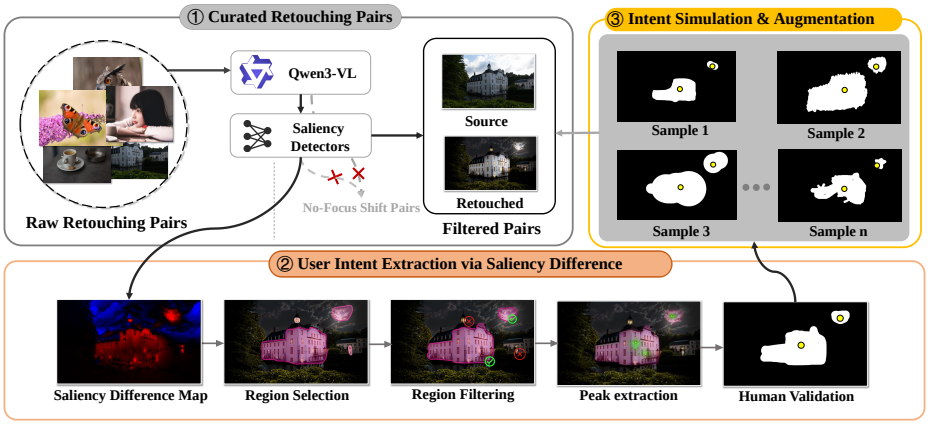

Fig. 2: Pipeline for generating intent-driven retouching supervision: (1) curating retouching pairs, (2) deriving user intent via saliency difference, and (3) simulating and augmenting intent signals.

3.1 Overview

EyeControl is a user-intention-driven image retouching agent that supports not only global enhancement (*Global Mode*) and local adjustment (*Local Mode*), but also visual focus enhancement (*Focus Mode*) based on weak or ambiguous user inputs such as clicks or free-form brush strokes. The overall workflow of EyeControl is illustrated in Fig. 3.

The framework consists of three components: a user interaction interface, a VLM-based Planner $\mathcal{A}$, and a Diffusion Transformer-based retouching Executor $\mathcal{E}$. Through the interaction interface, users indicate the desired visual focus using clicks, strokes, or region markings, optionally accompanied by textual guidance t describing their intention. The Planner $\mathcal{A}$ integrates and interprets these ambiguous signals, performs intention reasoning, determines the appropriate execution mode, and generates refined textual guidance for downstream processing.

Conditioned on the input image I, the inferred intention mask M_u, and the generated text guidance $\mathcal{G}$, the Executor performs the retouching operation and produces the updated result I_r, which can then be fed back into the system for subsequent interaction rounds. Formally, EyeControl implements a function:

$$\mathcal{F}_\theta(I, M_u) \rightarrow I_r. \tag{1}$$

3.2 Data Generation Pipeline

We design a three-stage data generation pipeline to construct training pairs, with a particular focus on extracting user-intended visual focus, as shown in Fig. 2. Further Details of our train set can be found in the supplemental materials.

Stage 1: Retouching Pairs Curation. Following the practice of JarvisArt [20], we curate a large-scale retouching corpus from PPR10K [19], Lightroom Community, and portfolios of professional retouchers. To ensure that the collected edits indeed induce visual focus enhancement, we apply a two-stage filtering procedure. First, Qwen3-VL-32B [33] is used to remove retouching pairs with negligible

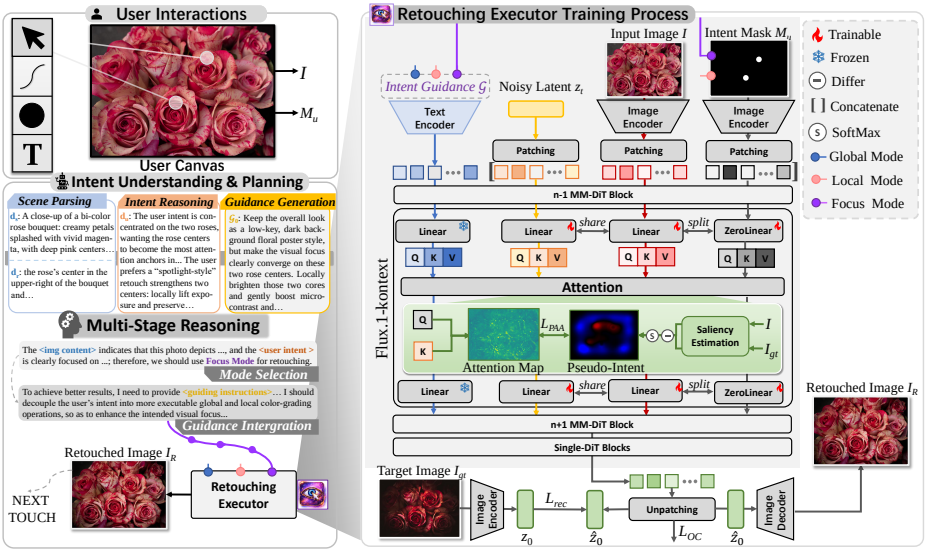

Fig. 3: The overall workflow of **EyeControl**. Given an image to be retouched, the system first collects user interactions. The *Planner* then performs multi-stage reasoning to parse the image, infer user intent, and select an appropriate execution mode, producing decoupled guidance prompts. Subsequently, the *Executor* carries out the actual image retouching conditioned on the input image, the inferred intention representation, and the generated guidance. The result is then returned to support iterative refinement.

perceptual differences or without a distinct visual focus before and after editing. Second, we estimate saliency difference between the pre- and post-retouch images using an ensemble of saliency detectors [9, 22], and discard samples whose saliency change falls below a predefined threshold. This procedure ensures that the retained image pairs present clear visual focus enhancement.

Stage 2: User-Intent Extraction via Saliency Difference. To construct structured supervision for visual focus enhancement, we derive user-intent signals from saliency difference maps computed between reference image pairs. Given a saliency difference map $D \in \mathbb{R}^{H \times W}$, we extract polarity-specific responses and process them through the following steps. 1) Region Selection: Determine threshold τ such that the selected region covers a predefined proportion ρ of total saliency energy, producing an initial binary mask. 2) Region Filtering: Perform connected-component analysis, remove small regions, and retain dominant components based on accumulated energy and size constraints. 3) Peak Extraction: Extract top-K local maxima within the selected region using non-maximum suppression to obtain sparse intention anchors. 4) Human-in-the-loop validation: select valid intent regions and peaks for training.

Stage 3: Intent Simulation and Augmentation. We further introduce an online intent augmentation strategy to better simulate real-world user interaction patterns. Given the intent representation extracted in the previous stage—including coarse regions and salient peaks, we randomly generate diverse interaction masks

to simulate typical user inputs, such as point clicks, free-form strokes, and region-based painting. While preserving the original intent, this augmentation produces varied interaction patterns and spatial layouts, improving the model's robustness to different user inputs in retouching.

3.3 EyeControl Framework

Intent Understanding and Planning User interactions in our setting are sparse, ambiguous, and often expressed in non-professional language. To bridge this gap, we introduce a collaborative multi-VLM agent $\mathcal{A}$, designed as a *Multi-Stage Reasoning* planner that progressively reconstructs the desired visual focus from incomplete observations, as shown in Fig. 3.

Formally, the Planner is modeled as a structured reasoning operator:

$$\mathcal{G} = \mathcal{A}(I_r, M_u), \tag{2}$$

which internally decomposes intent understanding into scene parsing, intent reasoning, and guidance generation before execution planning.

Intent Understanding The agent first decomposes the scene into complementary semantic channels: a scene-level description $d_s = \phi_s(I_r)$ capturing global stylistic attributes, and a region-aware description $d_r = \phi_r(I_r, M_u)$ focusing on the potential focal area. The user intent description d_u is inferred as:

$$d_u = \phi(I, d_s, d_r), \tag{3}$$

Finally, the planner $\mathcal{A}$ generates an initial unified guidance $\mathcal{G}_0$ based on the scene description d_s and the inferred user intention d_u:

$$\mathcal{G}_0 = \psi(d_s, d_u), \tag{4}$$

where $\psi(\cdot)$ denotes the guidance generation function.

Planning After completing intention understanding, the Planner determines the appropriate adjustment mode based on its joint interpretation of the inferred $\hat{d}_u$ and d_s. Specifically, it selects among *Global Mode* for overall tonal and color adjustments, *Local Mode* for region-specific refinement, or **Focus Mode**, often treated as the default setting, which coordinates both global and local adjustments to enhance the intended visual focus.

To implement this decision, the Planner first produces a unified guidance representation $\mathcal{G}_0$ that captures scene context and inferred intention. This representation is then decoupled into executor-compatible instructions, consisting of a global edit instruction $\mathcal{G}_{\text{global}}$ and a mask-aware local edit instruction $\mathcal{G}_{\text{local}}$.

The final executor-compatible instruction is formulated as:

$$\mathcal{G} = m_1 \cdot \{\mathcal{G}_{\text{global}}\} \cup m_2 \cdot \{\mathcal{G}_{\text{local}}\}, \tag{5}$$

where $\mathbf{m} = (m_1, m_2) \in \{(1,0),(0,1),(1,1)\}$ is a binary mode indicator controlling the activation of global and local instructions. Specifically,

$$\mathbf{m} = \begin{cases} (1,0), & \text{Global Mode}, \\ (0,1), & \text{Local Mode}, \\ (1,1), & \text{Focus Mode}. \end{cases} \tag{6}$$

Under *Focus Mode*, both global and local instructions are activated to collaboratively enhance the intended visual focus. The assembled instruction $\mathcal{G}$ is then fed into the diffusion executor together with the input image I and the intent mask M_u. In Global Mode, the mask is replaced by an all-zero mask, so that the retouching is driven purely by global guidance.

Retouching Executor Next, we describe the architecture and training process of the retouching executor.

Architecture The overall architecture is built upon Flux.1-Kontext [14] and augmented with intention-aware conditioning. Given an input image I, a user-provided intent mask M_u, and structured intent guidance text $\mathcal{G}$, the image is first encoded into a latent representation via a VAE encoder. During training, the latent is perturbed according to the diffusion process to obtain a noisy latent z_t. I, M_u, z_t, and $\mathcal{G}$ are patchified into token sequences and jointly processed by stacked MM-DiT blocks. Within each block, latent tokens interact with image-condition, mask, and text tokens through multi-modal attention. To mitigate feature interference across modalities, the mask branch is equipped with dedicated projection parameters, enabling mask-aware modulation of latent representations. After passing through multiple MM-DiT layers followed by Single-DiT refinement blocks, the network predicts the denoised latent. Training is supervised in latent space using a standard Flow-Matching reconstruction loss:

$$L_{rec} = \|\hat{z}_0 - z_0\|_2^2, \tag{7}$$

where $\hat{z}_0$ is the predicted clean latent and z_0 denotes the encoded latent of the ground-truth retouched image I_{gt}. The final retouched output I_r is obtained by decoding the predicted latent through the VAE decoder.

Pseudo-Intent Guided Attention Alignment(PAA) Although mask tokens participate in multi-modal conditioning, standard attention mechanisms do not explicitly regulate how spatial intention influences latent interactions. As a result, mask concatenation alone cannot guarantee that user interaction leads to consistent emphasis on the intended regions during denoising.

In the DiT, attention at timestep t follows the standard formulation:

$$\mathbf{A}_t = \text{Softmax}\left(\frac{\mathbf{Q}_t \mathbf{K}_t^\top}{\sqrt{d}}\right), \tag{8}$$

where $\mathbf{Q}_t = W_Q \mathbf{X}_t$ and $\mathbf{K}_t = W_K \mathbf{X}_t$ are the query and key projections of the latent features $\mathbf{X}_t$, and d denote the channel dimension.

When mask conditioning is incorporated, mask tokens derived from M_u are appended to the token sequence. Let $\mathbf{Q}_t^{(M)}$ denote queries originating from mask tokens, and $\mathbf{K}_t^{(z)}$ denote keys from noisy latent tokens z_t. The corresponding mask-to-latent attention map is:

$$\mathbf{A}_t^{(M\to z)} = \text{Softmax}\left(\frac{\mathbf{Q}_t^{(M)}\left(\mathbf{K}_t^{(z)}\right)^\top}{\sqrt{d}}\right). \tag{9}$$

This sub-attention determines which spatial regions in the latent representation are influenced by the mask guidance. Supervising $\mathbf{A}_t^{(M\to z)}$ therefore directly constrains the effective region of mask-guided modulation.

We construct a pseudo-intent map $\tilde{A}$ from saliency difference between pre- and post-adjustment signals. Unlike a binary mask, $\tilde{A}$ encodes signed prominence variations, indicating both regions to be enhanced and regions to be suppressed.

Let $S(\cdot) : \mathbb{R}^{H\times W\times 3} \to [-1,1]^{H\times W}$ denote a saliency estimation function that produces a normalized saliency map. The pseudo-intent map is defined as:

$$\tilde{A} = SoftMax(S(I_{gt}) - S(I)), \tag{10}$$

where I and I_{gt} denote the input image and the ground-truth retouched image, respectively. Unlike a binary mask, $\tilde{A}$ encodes signed prominence variations, indicating both regions to be enhanced and regions to be suppressed. We align the normalized mask-to-latent attention with the pseudo-intent map:

$$\mathcal{L}_{\text{PAA}} = \mathbb{E}_t\left[\frac{1}{L}\sum_{\ell=1}^{L}\left\|\text{Norm}\left(\mathbf{A}_t^{(M\to z),\ell}\right) - \text{Norm}(\tilde{A})\right\|_2^2\right], \tag{11}$$

where t denotes the diffusion timestep, L is the number of attention layers, and $\mathbf{A}_t^{(M\to z),\ell}$ represents the mask-to-latent attention map at layer ℓ and timestep t. Norm$(\cdot)$ denotes spatial normalization applied to both maps to ensure comparable scales. Through this alignment, the model learns to associate spatial masks with consistent attention allocation, enabling intent-aware retouching.

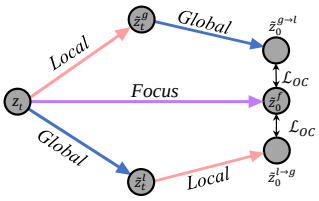

Fig. 4: Illustration of Operation Consistency Loss.

Operation-Consistensy Loss Global and local retouching operations form compositional transformations within a unified editing space. In particular, samples annotated under the *focus* mode simultaneously contain global and local editing instructions. This allows us to decompose a *focus* sample into two independent conditioning signals corresponding to the Global and Local modes.

Ideally, applying these operations sequentially should yield results consistent with applying them

jointly under the *focus* condition. However, diffusion models trained under mixed modes often exhibit order-sensitive behavior: dominant global adjustments may suppress local guidance, while local edits may disrupt overall visual coherence.

To address this issue, we introduce an **operation-consistency** objective using *focus* mode samples as supervision. For each *focus* sample, we construct one joint denoising path and two sequential paths with randomly sampled execution orders (Fig. 4).

For the focus path, the model performs a standard forward pass under the focus mode, producing the latent prediction $\hat{z}_t^f$ at timestep t, which serves as the anchor target.

Global-to-Local Path. We first denoise under the global-only condition to obtain $\hat{z}_0^g$, then inject noise:

$$\tilde{z}_t^g = (1 - \sigma_t)\hat{z}_0^g + \sigma_t \epsilon, \tag{12}$$

where ϵ is Gaussian noise and σ_t controls the noise injection strength. The renoised latent $\tilde{z}_t^g$ is refined under the local condition, producing $\hat{z}_0^{g\to l}$.

Local-to-Global Path. Similarly, swapping the execution order yields $\hat{z}_0^{l\to g}$.

Finally, we enforce operation-consistency by aligning the sequential prediction with the Focus anchor:

$$\mathcal{L}_{\mathrm{OC}} = \lambda \left\| (\hat{z}_0^o) - \mathrm{stopgrad}(\hat{z}_0^f) \right\|_2^2, \tag{13}$$

where $o \in \{g\to l,\ l\to g\}$ denotes the execution order sampled during training.

In summary, the overall training objective is:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{PAA} + \lambda_2 \mathcal{L}_{OC}. \tag{14}$$

4 Experiments

4.1 Implementation details

Training setting We utilize Flux-1.0-Kontext [14] as the base model. For fine-tuning, we apply LoRA (rank $r = 128$) and optimize the network using a learning rate of 2×10^{-5}. The training is conducted across 4×H20 for 10,000 steps with a batch size of 1 per GPU.

ControlArt-Bench To the best of our knowledge, there is currently no publicly available retouching benchmark designed to evaluate image editing quality under ambiguous user intentions, particularly for the assessment of visual focus enhancement. Therefore, we introduce **ControlArt-Bench**, a high-quality evaluation dataset specifically constructed for focus enhancement. ControlArt-Bench contains 200 groups of real-user retouching samples, each consisting of an *IntentMask–Image* pair. The dataset spans diverse categories, including portraits, natural landscapes, architecture, food, and animals. To further improve its usability and reproducibility, we additionally provide, for each sample, an editing instruction aligned with the corresponding retouching intention.

Table 1: Quantitative evaluation of retouching performance on ControlArt-Bench. The best and second-best results are highlighted. FA, PQ, and O denote the metrics evaluated by Qwen2.5-VL-72B [2].

Method	PSNR↑	SSIM↑	FA↑	PQ↑	O↑	KL↓	CC↑	SIM↑
Advanced Retouching Agents								
JarvisArt [20]	18.8169	0.7463	7.8350	8.7050	8.2087	0.2229	0.9548	0.8785
JarvisEvo [21]	20.9291	0.8293	7.9700	9.2000	8.4895	0.2928	0.9644	0.8913
PerTouch [4]	15.9370	0.5946	7.4650	8.4100	7.8240	0.2700	0.9621	0.8828
Open-Source Editing Diffusion Models								
UniWorld-v2 [18]	18.0759	0.6879	8.2850	8.5850	8.3832	0.7567	0.9363	0.8494
Step1X-Edit [35]	18.1433	0.7533	8.1950	8.8350	8.4632	0.2508	0.9564	0.8789
FLUX.1-Kontext [14]	18.0646	0.7471	8.2200	7.2150	7.4000	0.8231	0.9243	0.8362
Qwen-Image-Edit-2511 [31]	18.7092	0.7029	8.2550	8.6200	8.3530	0.2610	0.9577	0.8783
Commercial Closed-Source Models								
GPT-Image-1.5 [25]	16.7133	0.5659	8.3291	8.8101	8.5196	0.6651	0.9141	0.8269
Nano-Banana-2 [27]	21.1127	0.7633	8.1800	9.2900	8.6785	0.2389	0.9735	0.9085
EyeControl(Ours)	21.8845	0.8517	8.2100	9.3050	8.7054	0.2215	0.9743	0.9068

Fig. 5: Qualitative comparisons with other methods on ControlArt-Bench.

Metrics We report eight metrics: PSNR, SSIM, FA, PQ, O, KL, CC, and SIM. PSNR and SSIM measure the overall fidelity and structural similarity to the reference image. We introduce Focus Alignment (FA) to evaluate how well the visual focus in the retouched image aligns with the regions specified by the user intent (0–10 scale). PQ measures contextual coherence and artifacts (0–10 scale) [12]. The overall score $O = \sqrt{FA \times PQ}$. Furthermore, KL, CC, and SIM quantify visual-focus (saliency) discrepancies between the retouched and reference images [13]. See the supplementary material for details.

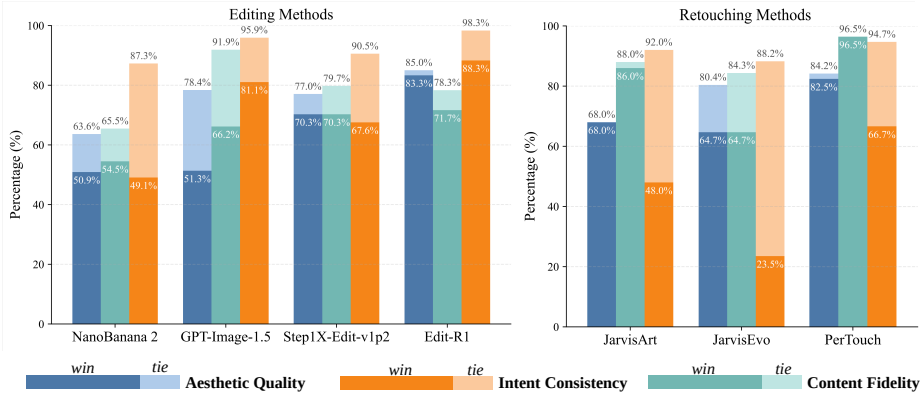

Fig. 6: User Preference Study. Comparison across three criteria: aesthetic quality, intent consistency, and content fidelity. We report the percentages of win and tie outcomes of EyeControl against each competing method. Results are shown separately for image editing models (left) and agent-based retouching models (right).

4.2 Compared with other Methods

We compare EyeControl with open-source agent-based retouching models (JarvisArt [20], JarvisEvo [21], PerTouch [4]), generative editing models (UniWorld-v2 [18], Step1X-Edit [35][1], FLUX.1-Kontext [14], Qwen-Image-Edit-2511 [31]), and commercial systems (GPT-Image-1.5 [25][2], Nano-Banana 2 [27][3]). To ensure a fair comparison, we provide detailed long-form editing instructions that are carefully aligned with the underlying user intentions. In addition, based on the Intent Mask, we generate corresponding bounding-box location descriptions to accommodate the input formats required by different models. Further implementation details, along with additional qualitative results on global and local image retouching, are provided in the supplementary material.

Comparison on ControlArt-Bench As shown in Tab. 1, EyeControl outperforms existing image retouching models, open-source diffusion-based editing models, and GPT-Image-1.5 on the majority of evaluation metrics. Moreover, our approach achieves performance comparable to NanoBanana 2, demonstrating its strong competitiveness against advanced commercial systems. We observe in-Fig. 5 that some editing models score well on the Focus Alignment metric mainly because they introduce large structural changes, rather than focus guidance via subtle retouching. In contrast, our method demonstrates clear advantages in visual focus enhancement. It combines the high content fidelity typically observed in retouching models with the broad editing flexibility characteristic of diffusion-based editing approaches. More importantly, EyeControl substantially surpasses competing methods in intention alignment, achieving significantly stronger consistency with the specified user intent.

[1] the latest version of *Step1X-Edit-v1p2*, released on Nov 26, 2025
[2] Generative model from OpenAI, released in December, 2025
[3] the latest model from Google, released in March, 2026

Table 2: Quantitative Ablation of EyeControl.

Variants	Methods	Focus				Global		Local		Multi-Round		Average	
		PSNR	SSIM	O	KL	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
A	baseline	21.29	0.8502	8.64	0.2006	18.54	0.7655	27.63	0.9428	19.03	0.8165	19.62	0.8438
B	A+Planner	21.71	0.8565	8.64	0.2199	18.07	0.7596	27.75	0.9433	19.48	0.8254	21.76	0.8462
C	B+PAA	21.92	0.8616	8.66	0.1655	17.75	0.7657	28.55	0.9569	20.33	0.8340	22.14	0.8546
D	C+OCLoss	21.88	0.8517	8.71	0.2215	23.34	0.800	28.37	0.9408	20.81	0.8353	24.17	0.8587

Fig. 7: Effect of Pseudo-Intent Guided Attention Alignment.

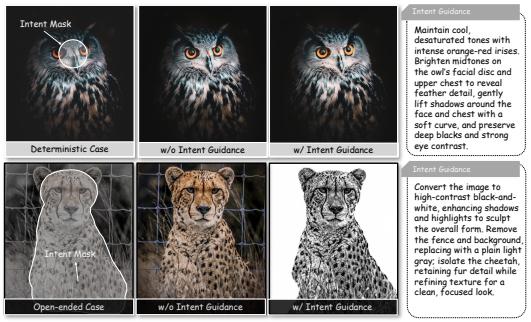

Fig. 8: Effect of intent guidance under different levels of user specificity.

User Preference Study Evaluating intent-driven image retouching is subjective, as aesthetic preference varies across individuals. To quantitatively assess this subjectivity, we conduct a large-scale user study on ControlArt-Bench. We recruit 50 participants to compare our method against seven state-of-the-art approaches, including four image editing methods and three image retouching methods.

The evaluation is conducted from three perspectives: (1) **Aesthetic Quality**, measuring whether the image is visually pleasing; (2) **Intention Alignment**, assessing whether the highlighted region aligns with the user's intended focus; and (3) **Image Fidelity**, evaluating whether the edited result preserves the original content structure.

For each comparison, we randomly select one competing method and perform a blind pairwise comparison against our approach. Participants are asked to choose the better result under each metric (or indicate a tie). As shown in Fig. 6, EyeControl consistently outperforms all competing approaches across all three evaluation criteria.

4.3 Ablation Study and Discussion

We quantitatively ablate the major components of EyeControl in Tab. 2. The evaluation is conducted on the full test sets without sampling, covering four representative user scenarios: focus enhancement and multi-round retouching on ControlArt-Bench, global enhancement on MIT-FiveK [3], and local enhancement on MMArt-Bench [20]. Starting from a fine-tuned Flux.1-Kontext baseline, adding the planner improves the average performance, suggesting that explicit planning helps constrain the retouching direction. Introducing PAA further improves focus and local editing performance, achieving the best Focus

PSNR/SSIM and Local PSNR/SSIM among all variants. This indicates that attention-level supervision is particularly important when the edit needs to redistribute visual prominence within a localized or semantically intended region. Finally, adding operation-consistency loss leads to the best overall average performance across the four scenarios. Although it does not uniformly improve every individual metric, it substantially improves the Global setting, increasing PSNR/SSIM from 17.75/0.7657 to 23.34/0.800, and also yields the best Multi-turn performance.

We further analyze the contribution of each component in EyeControl through focused discussions. We focus our discussion on three primary aspects:

Are spatial masks sufficient for visual focus enhancement without explicit attention alignment? Fig. 7 suggests that spatial masks alone are insufficient to induce structured focus enhancement. In conventional conditioning designs, mask tokens share QKV projections with image conditions and noisy latent tokens. However, masks encode intention cues rather than visual appearance. When forced to share projections, attention responses remain diffuse and concentrate along mask boundaries, indicating that the model treats the mask mainly as a spatial constraint rather than a driver of prominence redistribution. Consequently, the edits rely primarily on photometric adjustments, resulting in weak perceptual separation between focal and non-focal regions.

With dedicated mask projections and pseudo-intention attention alignment, the attention distribution becomes more concentrated within the intended region and less boundary-driven. The outputs show clearer structural separation between the preserved red telephone box and the desaturated background. These observations suggest that architectural decoupling and explicit attention supervision are necessary to transform spatial masks into effective mechanisms for visual focus modulation.

What role does VLM-driven visual guidance play in focus enhancement? Fig. 8 further reveals that the impact of VLM-driven guidance depends on the ambiguity of the editing intention. When the intended adjustment is explicit and the image offers limited stylistic variation, the difference between using VLM guidance and omitting it is marginal. In such cases, the spatial mask alone provides sufficient structural constraint for reasonable focus enhancement.

However, when multiple plausible adjustment directions exist—such as atmospheric tone balancing, selective desaturation, or contrast redistribution—the absence of VLM guidance leads to inconsistent or unstable edits. Without high-level reasoning, the model may over-enhance local regions or fail to harmonize the global atmosphere with focal emphasis.

These results suggest that VLM-driven guidance mainly acts as a semantic disambiguation mechanism. Rather than amplifying enhancement strength, it constrains the space of plausible editing trajectories, enabling intention-consistent and scene-aware focus modulation in complex retouching scenarios.

How Does Operation-Consistency Loss Affect Retouching? Fig. 9 illustrates the effect of operation-consistency regularization on multi-round re-

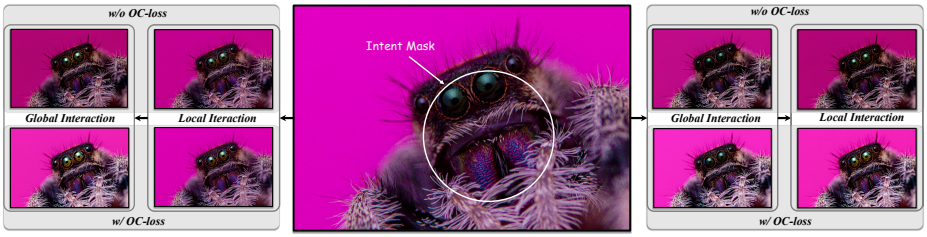

Fig. 9: Effect of operation-consistency loss (OC-loss) on multi-round retouching.

touching. Without OC-loss, the interaction between global and local adjustments becomes path-dependent: different execution orders (Global→Local vs. Local→Global) produce noticeably different results. In particular, repeated local refinement may override the global atmosphere, while dominant global adjustments can suppress focal emphasis. This suggests that, without structural constraints, heterogeneous editing conditions interfere during diffusion.

With OC-loss enabled, editing trajectories across different execution orders become more consistent. The focal region maintains stable prominence, and the global color tone remains coherent across rounds. This indicates that operation-consistency regularization enforces path-invariant editing dynamics and stabilizes the interaction between global and local directives. Rather than merely aligning pixel outputs, OC-loss reshapes the diffusion process to guide multi-scale adjustments toward a coherent retouching trajectory.

5 Conclusion

In this work, we revisit image retouching from the perspective of visual focus enhancement, where the goal is not only to improve overall image quality but also to intentionally guide viewers' attention toward desired regions. To address this problem under weak user intent signals, we propose EyeControl, an intent-driven retouching agent that explicitly links user intention, visual focus, and tonal adjustment operations during the retouching process.

The proposed framework combines intention understanding, attention-level focus modulation, and coordinated global–local adjustments to enable controllable visual focus enhancement while preserving perceptual naturalness. In addition, we introduce ControlArt-Bench, a dedicated benchmark with paired spatial intent annotations and focus-oriented evaluation metrics for systematic assessment of visual focus enhancement.

6 Acknowledgement

This work is supported in part by the Tianjin Natural Science Foundation Project (25ZXRGGX00290, 24JCJQJC00020, 25JCQNJC01390), the National Natural Science Foundation of China (62306153, 62225604), the Young Elite Scientists Sponsorship Program by CAST (YESS20240686), Shenzhen Science and Technology Program (JCYJ20240813114237048) and the Fundamental Research Funds for the Central Universities (Nankai University, 63253223, 63253219).

References

1. Aberman, K., He, J., Gandelsman, Y., Mosseri, I., Jacobs, D.E., Kohlhoff, K., Pritch, Y., Rubinstein, M.: Deep saliency prior for reducing visual distraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19851–19860 (2022)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
3. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 97–104. IEEE (2011)
4. Chang, Z., Duan, Z.P., Zhang, J., et al.: Pertouch: A unified diffusion-based image retouching framework with vlm-driven agent. In: The 40th Annual AAAI Conference on Artificial Intelligence (2026)
5. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6306–6314 (2018)
6. Duan, Z.P., Zhang, J., Lin, Z., Jin, X., Wang, X., Zou, D., Guo, C.L., Li, C.: Diffretouch: Using diffusion to retouch on the shoulder of experts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2825–2833 (2025)
7. Dutt, N.S., Ceylan, D., Mitra, N.J.: Monetgpt: Solving puzzles enhances mllms' image retouching skills. ACM Transactions on Graphics (TOG) **44**(4), 1–12 (2025)
8. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
10. Jiang, L., Xu, M., Wang, X., Sigal, L.: Saliency-guided image translation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16504–16513 (2021)
11. Kim, H.U., Koh, Y.J., Kim, C.S.: Pienet: Personalized image enhancement network. In: European Conference on Computer Vision. pp. 374–390. Springer (2020)
12. Ku, M., Li, T., Zhang, K., Lu, Y., Fu, X., Zhuang, W., Chen, W.: Imagenhub: Standardizing the evaluation of conditional image generation models. In: The Twelfth International Conference on Learning Representations (2024)
13. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Saliency benchmarking made easy: Separating models, maps and metrics. In: European Conference on Computer Vision. pp. 798–814 (2018)
14. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
15. Li, C., Guo, C., Feng, R., Zhou, S., Loy, C.C.: Cudi: Curve distillation for efficient and controllable exposure adjustment (2022), https://arxiv.org/abs/2207.14273

16. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE transactions on pattern analysis and machine intelligence **44**(8), 4225–4238 (2021)
17. Li, J., Fang, P.: Hdrnet: Single-image-based hdr reconstruction using channel attention cnn. In: Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing. pp. 119–124 (2019)
18. Li, Z., Liu, Z., Zhang, Q., Lin, B., Wu, F., Yuan, S., Yan, Z., Ye, Y., Yu, W., Niu, Y., Wang, S., Cheng, X., Yuan, L.: Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback (2025), https://arxiv.org/abs/2510.16888
19. Liang, J., Zeng, H., Cui, M., Xie, X., Zhang, L.: Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 653–661 (2021)
20. Lin, Y., Lin, Z., Lin, K., Bai, J., Pan, P., Li, C., Chen, H., Wang, Z., Ding, X., Li, W., YAN, S.: Jarvisart: Liberating human artistic creativity via an intelligent photo retouching agent. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
21. Lin, Y., Wang, L., Lin, K., Lin, Z., Gong, K., Li, W., Lin, B., Li, Z., Zhang, S., Peng, Y., Dai, W., Ding, X., Wang, C., Lu, Q.: Jarvisevo: Towards a self-evolving photo editing agent with synergistic editor-evaluator optimization (2025), https://arxiv.org/abs/2511.23002
22. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4722–4732 (October 2021)
23. Miangoleh, S.M.H., Bylinskii, Z., Kee, E., Shechtman, E., Aksoy, Y.: Realistic saliency guided image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 186–194 (2023)
24. Moran, S., McDonagh, S., Slabaugh, G.: Curl: Neural curve layers for global image enhancement. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9796–9803. IEEE (2021)
25. OpenAI: Gpt-image-1.5. https://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here/ (2025), accessed: 2026
26. Song, Y., Qian, H., Du, X.: Starenhancer: Learning real-time and style-aware image enhancement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4126–4135 (2021)
27. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., Johnson, M., et al.: Gemini: A family of highly capable multimodal models (2025), https://arxiv.org/abs/2312.11805
28. Wang, T., Li, Y., Peng, J., Ma, Y., Wang, X., Song, F., Yan, Y.: Real-time image enhancer via learnable spatial-aware 3d lookup tables. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2471–2480 (2021)
29. Wen, X., Xie, L., Jiang, L., Chen, T., Wu, S., Liu, C., Wong, H.S.: Retouchformer: Semi-supervised high-quality face retouching transformer with prior-based selective self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5903–5911 (2024)
30. Wong, L.K., Low, K.L.: Saliency retargeting: An approach to enhance image aesthetics. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV). pp. 73–80. IEEE (2011)

31. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
32. Wu, Q., Shi, J., Jenni, S., Kafle, K., Wang, T., Chang, S., Zhao, H.: Retouchiq: Mllm agents for instruction-based image retouching with generalist reward (2026), https://arxiv.org/abs/2602.17558
33. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
34. Yang, C., Jin, M., Jia, X., Xu, Y., Chen, Y.: Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17522–17531 (2022)
35. Yin, F., Liu, S., Han, Y., Wang, Z., Xing, P., Wang, R., Cheng, W., Wang, Y., Li, A., Yin, Z., Chen, P., Zhang, X., Jiang, D., Zeng, X., Yu, G.: Reasonedit: Towards reasoning-enhanced image editing models (2025), https://arxiv.org/abs/2511.22625
36. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 2058–2073 (2020)