One Object, Different Corrections: Controllable Low-Light Enhancement with Photometric Actions

Anonymous authors
Paper under double-blind review

Abstract

A vision-language model (VLM) brings semantic understanding to low-light image enhancement, but language guidance alone does not specify how much correction to apply. An instruction to brighten an object leaves open how its shadowed and well-lit parts should be adjusted. We introduce an Executable Photometric Action Space (EPAS) that separates image-dependent correction from semantic execution control. A signed log-gain field specifies correction direction and magnitude, while a separate control field determines where and how strongly it is applied, allowing spatially varying corrections within semantic regions. During training, Metropolis–Hastings sampling explores regional actions under an energy defined on the input image's photometric statistics and VLM-derived preferences, with VLM quality assessment guiding candidate selection. This constructs action targets without paired normal-light references; the selected actions are distilled into a deterministic predictor for inference without action search. Applying the controlled gain produces a preview that guides generative restoration for detail recovery and noise suppression. Experiments show that spatially varying gains reduce the shadow undercorrection and highlight overcorrection caused by object-wise uniform gains, especially when correction requirements vary substantially within an object. Outputs approximately follow supplied actions, with more accurate restoration when using the input's own gain field. A single trained model achieves competitive reconstruction quality, strong perceptual performance, and flexible global and local control across diverse benchmarks and challenging illumination conditions without dataset-specific retraining or fine-tuning.

1 INTRODUCTION

Vision-language models (VLMs) bring scene understanding and task reasoning to low-light image enhancement (LLIE) (Chen et al., 2024; Zhu et al., 2025; Lin et al., 2025a), yet these capabilities alone do not determine how much enhancement an image requires. The same instruction, such as "brighten the person," may call for different adjustments depending on the input's brightness and local lighting conditions. Within a single scene, shadows, well-exposed regions, and bright light sources may require corrections that differ in both magnitude and direction. Semantic understanding thus leaves a gap between identifying the intended change and determining the quantitative correction needed for a given input.

The gap persists because advances in restoration modeling and semantic reasoning do not necessarily make the required correction independently controllable. Conventional deep learning methods learn enhancement behavior from normal-light targets or predefined appearance constraints (Cai et al., 2023; Yan et al., 2025; Guo et al., 2020; Jiang et al., 2024). These objectives guide what an enhanced image should look like, while the model determines how much to alter each region. Although illumination maps and adjustment curves can make this process explicit, the resulting corrections are typically coupled to the model's enhancement objective rather than independently controlled through semantic intent. Chain-of-thought and agentic approaches add reasoning, planning, and feedback around restoration (Zhu et al., 2025; Lin et al., 2025a; Zhou et al., 2025), but largely retain this division of responsibility: the reasoning system decides which operation to invoke or whether its result is satisfactory, while the enhancer determines the actual photometric change. Consequently, better task-level decisions do not directly resolve input-dependent correction errors

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.

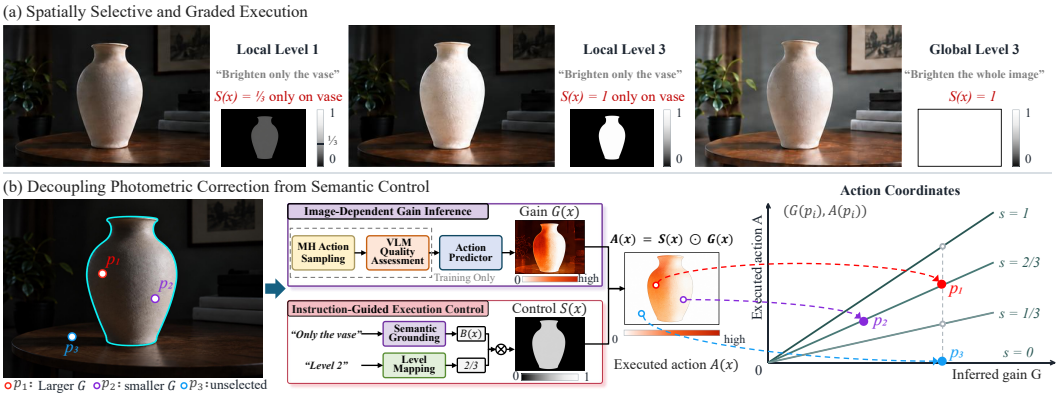


Figure 1: Conceptual illustration of EPAS. (a) A shared gain field supports global and object-specific enhancement at different execution levels. (b) The image-dependent gain field G encodes photometric corrections, while the instruction-derived control map S specifies where and how strongly to execute them. Their composition, $A = S \odot G$, allows different corrections within the same semantic object under a shared control level, with zero action outside the selected region.

under domain shifts and complex illumination. This motivates a representation that separates the correction required by the image from the semantic decision of where and how strongly to apply it.

We introduce an Executable Photometric Action Space (EPAS) that makes the connection between semantic intent and photometric correction explicit. EPAS separates the correction inferred from the image from the choice of where and how strongly to execute it. A signed log-gain field G encodes correction direction and magnitude, while a control field S specifies the selected regions and execution levels. Their product, $A = S \odot G$, defines the photometric action applied to the input. Figure 1 (a) illustrates global and object-specific enhancement at different levels using a shared gain field. Figure 1 (b) shows that points p_1 and p_2 within the same object can receive different corrections under a shared control level, while an unselected point p_3 receives zero photometric action. The regions used to specify an edit therefore need not match those used to infer its photometric correction.

We realize EPAS through a learning framework that connects regional action inference with spatially controlled generative restoration. During training, Metropolis–Hastings (MH) sampling explores a compact regional parameter space under photometric criteria and VLM-derived preferences. VLM quality assessment guides candidate selection, yielding action targets without paired normal-light references. These targets train a deterministic predictor that infers spatially varying gains without test-time search. The inferred gain and independently specified control fields guide a pretrained image editor, adapted on examples with different regional enhancement levels to recover detail and suppress noise. Controlled experiments show that retaining within-object gain variation reduces shadow undercorrection and highlight overcorrection, with larger benefits as correction requirements become more heterogeneous (Figure 5). Gain interventions show that outputs approximately follow supplied actions, while gain shuffling increases reference error despite similar tracking error (Figure 6). These findings distinguish action execution from the suitability of the inferred correction for the input.

Our main contributions are summarized as follows:

- We introduce EPAS, an executable photometric action space that separates input-dependent correction from semantic control through a gain field G and a control field S. Their composition enables spatially selective and graded execution while accommodating different photometric requirements within the same semantic object.
- We develop a unified framework that connects regional action inference with spatially controlled generative restoration. Training-time sampling and VLM assessment are distilled into direct action prediction, while the composed action field provides explicit photometric guidance for detail recovery and noise suppression.
- We validate within-object gain variation and distinguish action execution from input–gain correspondence through controlled experiments. A single trained model achieves competitive reconstruction quality, strong perceptual performance, and global and local control

across diverse benchmarks and challenging illumination, without dataset-specific retraining or fine-tuning.

2 RELATED WORK

2.1 Low-light image enhancement

Paired LLIE methods address illumination, color, noise, and detail through Retinex-based Transformers, photometric spaces, or generative models (Zhang et al., 2021; Wu et al., 2022; Cai et al., 2023; Wang et al., 2023; Yan et al., 2025; Feijoo et al., 2025), but their targets encode exposure, sensor response, white balance, and tone mapping, making learned brightness preferences sensitive to domain shifts. Reference-free methods exploit unpaired images, physical priors, illumination constraints, or test-image statistics (Guo et al., 2020; Jiang et al., 2021; 2024; Wang et al., 2024; Jin et al., 2025; Kong et al., 2025), yet retain normal-light distribution assumptions or exposure and illumination priors that may fail under mixed lighting. Photon-limited shadows require amplification and denoising, bright sources protection or attenuation, and well-exposed regions preservation. These corrections are typically tied to appearance objectives, limiting transfer, independent adjustment, and regional recombination. Controllable methods expose discrete levels, continuous strengths, or semantic masks (Yin et al., 2023; Zhang et al., 2025a; Yang et al., 2026), but these controls need not specify interpretable photometric operations: levels order outputs without quantifying required corrections, strengths modulate learned responses dependent on data and content, and masks select regions while leaving correction direction and magnitude to the enhancer. Spatial selection and photometric adjustment thus generally remain coupled.

2.2 Vision-language agents and reasoning for image restoration

Vision-language models support degradation diagnosis, tool selection, and instruction following (Zuo et al., 2025; Lin et al., 2025a; Li et al., 2026; Lu et al., 2026; Wei et al., 2026). RestoreAgent and AgenticIR sequence restorers for composite degradations (Chen et al., 2024; Zhu et al., 2025), while MAIR coordinates expert planning and execution and TIR-Agent learns tool-calling policies to reduce search (Jiang et al., 2026; Jia et al., 2026). These agents determine which operations to invoke and when, leaving brightness correction to the enhancer. Reasoning-based models add iterative feedback and stepwise degradation or quality analysis (Chen et al., 2026; Li et al., 2025; Wu et al., 2025b), characterizing inputs or ranking outputs without specifying the direction and magnitude of scene-dependent regional corrections. Such advances in orchestration and evaluation generally retain the restorer's learned mapping from image conditions to local corrections, and therefore do not directly resolve its brightness biases or robustness limitations under domain shifts, extreme darkness, and mixed illumination.

3 Methodology

Overview. Given a low-light image $I_{\text{low}} \in [0,1]^{H\times W\times 3}$, EPAS separates photometric correction G from spatial execution control S. Training-time MH search supplies targets for direct action prediction at inference. The controlled action produces a preview that guides an image editor, together with spatial conditions and a restoration prompt, to generate I_{out}.

3.1 Executable Photometric Action Space

Let Θ denote regional action parameters and $\mathcal{R}_\Theta$ the photometric operator. We express its correction relative to the input through the rendered image and corresponding log-gain field:

$$\begin{cases} I_{\text{ph}}(\Theta) = \mathcal{R}_\Theta(I_{\text{low}}), \\ G_\Theta = \log\left(I_{\text{ph}}(\Theta) + \epsilon\right) - \log\left(I_{\text{low}} + \epsilon\right), \end{cases} \tag{1}$$

where $\epsilon > 0$ ensures numerical stability. Positive and negative gains describe amplification and attenuation, respectively. They quantify intensity changes, not calibrated camera exposure.

To construct $\mathcal{R}_\Theta$, a network combines image features and frozen vision-language features to predict K soft photometric memberships M_k, normalized by a softmax so that $\sum_k M_k(p) = 1$. Each region

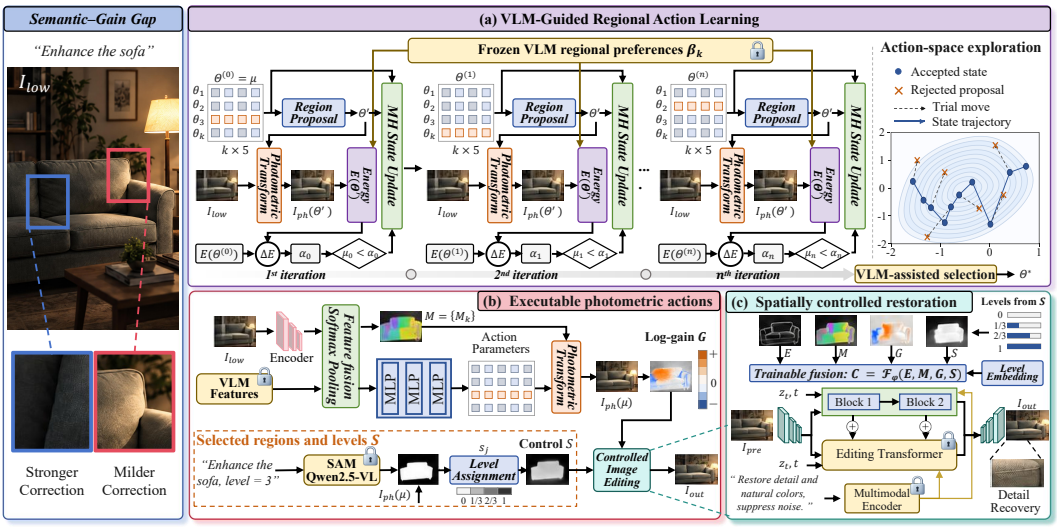


Figure 2: Overview of EPAS. (a) Regional action estimation through MH sampling and VLM assessment. (b) Image enhancement by composing photometric gains with semantic control and applying an image editor. (c) Editor architecture with spatial condition fusion for controlled restoration.

has five action parameters $\vartheta_k \in \mathbb{R}^5$, collected in $\Theta \in \mathbb{R}^{K\times 5}$. The dense parameter field is

$$\begin{cases} P_\Theta(p) = \sum_{k=1}^{K} M_k(p)\vartheta_k, \\ \boldsymbol{a}_\Theta(p) = \boldsymbol{a}_{\min} + (\boldsymbol{a}_{\max} - \boldsymbol{a}_{\min}) \odot \sigma(P_\Theta(p)), \end{cases} \tag{2}$$

where p indexes pixels, σ is the sigmoid function, and the bounds enforce valid ranges. The channels $\boldsymbol{a}_\Theta = (\gamma, w, g, b, h)$ control the base tone response, illumination dependence, gain, shadow adjustment, and highlight suppression. A smooth illumination estimate determines the tone response:

$$\begin{cases} L = \mathrm{Avg}_c\left(\mathcal{B}(I_{\mathrm{low}})\right), \\ \kappa = (\gamma + wL)^{2L-1}, \end{cases} \tag{3}$$

where $\mathcal{B}$ averages Gaussian-filtered images at two scales and Avg_c denotes a channel mean. The rendering applies tone correction followed by shadow and highlight adjustments:

$$\begin{cases} U_\Theta = (I_{\mathrm{low}} + \epsilon_r)^\kappa, \\ V_\Theta = U_\Theta + (1-L)bU_\Theta(1-U_\Theta), \\ I_{\mathrm{ph}}(\Theta) = \mathrm{clip}_{[0,1]}\left(g\left[V_\Theta - LhV_\Theta^2\right]\right), \end{cases} \tag{4}$$

where $\epsilon_r > 0$ stabilizes exponentiation. Operations are elementwise, with spatial parameters shared across color channels. This parameterization supports spatially varying corrections while keeping action search to $5K$ variables.

The photometric memberships M_k need not coincide with user-selected semantic regions. A separate control map $S \in [0,1]^{H\times W}$, broadcast across channels, executes the inferred gain:

$$\begin{cases} A_\Theta = S \odot G_\Theta, \\ I_{\mathrm{pre}} = \mathrm{clip}_{[0,1]}\left(I_{\mathrm{low}} \odot \exp(A_\Theta)\right). \end{cases} \tag{5}$$

Increasing S strengthens amplification or attenuation according to the sign of G_Θ. Numerical stabilization and gain clipping mean that $S = 1$ need not reproduce I_{ph}. Zero control preserves the input in I_{pre}, but imposes no hard constraint on the final output.

3.2 Learning Regional Actions via Sampling

An MLP predicts initial regional parameters μ and log-variances ℓ from pooled image features and frozen vision-language features. Training-time MH search supplies targets for direct prediction.

Candidate states are evaluated through $I_{\mathrm{ph}}(\Theta)$. Let $b_k(\Theta)$ denote its channel-mean intensity averaged within region M_k. A region-assessment prompt asks the frozen vision-language model to inspect the initial rendering and its mask overlay, providing regional brightness preferences β_k. These preferences guide evaluation rather than serve as calibrated exposure measurements. We define

$$E(\Theta) = |\langle I_{\mathrm{ph}}(\Theta)\rangle - b_0| + \lambda_{\mathrm{tv}}\,\mathrm{TV}\left(I_{\mathrm{ph}}(\Theta)\right) + \frac{\lambda_{\mathrm{pref}}}{K}\sum_{k=1}^{K}\left(b_k(\Theta) - \beta_k\right)^2, \tag{6}$$

where $\langle\cdot\rangle$ is the image mean, b_0 is a global brightness prior, and TV penalizes neighboring intensity differences. Starting from $\Theta^{(0)} = \mu$, each transition perturbs one region, cycling through the K regions while leaving the others unchanged:

$$\begin{cases} \vartheta_k' = \vartheta_k^{(n)} + s_k \odot \xi, \qquad \xi \sim \mathcal{N}(0, I_5), \\ \alpha_n = \min\left\{1,\, \exp\left(-\dfrac{E(\Theta') - E(\Theta^{(n)})}{\tau}\right)\right\}, \end{cases} \tag{7}$$

where $s_k = \mathrm{clip}(\eta\exp(\ell_k/2), s_{\min}, s_{\max})$ uses a fixed multiplier $\eta > 0$ and positive bounds, and $\tau > 0$ is the sampling temperature. Memberships, preferences, and scales remain fixed within each chain, with symmetric Gaussian proposals. Rejected proposals leave the current state unchanged.

After burn-in, retained states and the initial prediction form a candidate pool $\mathcal{C}$. The frozen vision-language model evaluates their rendered quality using an assessment prompt. With scores $\widehat{q}$ standardized within the pool, the training target is

$$\Theta^\star = \arg\min_{\Theta\in\mathcal{C}} \left[E(\Theta) - \lambda_q \widehat{q}(\Theta)\right]. \tag{8}$$

This assessment ranks completed candidates without altering the sampling rule. The predictor learns from the selected state through parameter and rendering consistency:

$$\mathcal{L}_{\mathrm{teach}} = \lambda_\Theta\,\mathrm{Huber}\left(\mu, \mathrm{sg}(\Theta^\star)\right) + \frac{\lambda_I}{3HW}\left\|I_{\mathrm{ph}}(\mu) - \mathrm{sg}\left(I_{\mathrm{ph}}(\Theta^\star)\right)\right\|_1. \tag{9}$$

where sg stops gradients through sampled targets. An auxiliary KL penalty regularizes the predicted Gaussian $\mathcal{N}(\mu, \mathrm{diag}(e^\ell))$ toward $\mathcal{N}(0, I)$. These targets are constructed from the low-light input rather than paired normal-light references. At inference, the MLP directly predicts μ without sampling. Eq. (1) then gives the gain field, clipped to $[-g_{\max}, g_{\max}]$ for stable execution.

3.3 Action-Anchored Restoration with Spatial Control

We adapt Qwen-Image-Edit (Wu et al., 2025a) to refine I_{pre} under textual and spatial conditioning. The restoration prompt directs the editor to follow the specified regions and enhancement levels while preserving detail and color and suppressing noise. Regional control is learned from Light100K brightness variants (Yang et al., 2026), without paired references from the evaluation benchmarks.

Each training scene provides a low-light image $I^{[0]} = I_{\mathrm{low}}$ and three brightness variants $\{I^{[1]}, I^{[2]}, I^{[3]}\}$. Using a grounding prompt, Qwen2.5-VL locates semantic regions in $I_{\mathrm{ph}}(\mu)$, and SAM converts their bounding boxes into masks $\{B_j\}_{j=1}^{J}$. These semantic masks are distinct from the photometric memberships M_k. Assigning each region a level $s_j \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ gives

$$\begin{cases} S_0(p) = \dfrac{\sum_j s_j B_j(p)}{\max\left(1, \sum_j B_j(p)\right)}, \\ S = \mathcal{H}_{I_{\mathrm{low}}}(S_0), \end{cases} \tag{10}$$

where overlap normalization prevents accumulation and $\mathcal{H}_{I_{\mathrm{low}}}$ smooths boundaries using image edges. Random regional assignments provide local-control examples, while constant maps $S = s$ provide global-control examples.

Crucially, the same S determines both the preview in Eq. (5) and the training target I_{tar}^S. Discrete values select the corresponding brightness variants, while intermediate values interpolate adjacent

variants. Let $l = \lfloor 3S(p) \rfloor$, $u = \min(l+1, 3)$, and $a = 3S(p) - l$. Defining luminance $Y_r = \max(\mathrm{Lum}(I^{[r]}), \epsilon)$ and chromaticity $\chi_r = I^{[r]}/Y_r$, we construct

$$I^S_{\mathrm{tar}}(p) = \mathrm{clip}_{[0,1]} \Big([(1-a)\chi_l(p) + a\chi_u(p)] \cdot \exp\left[(1-a)\log Y_l(p) + a \log Y_u(p)\right] \Big). \tag{11}$$

This yields regional targets without separately captured references, preserving the input at zero control and interpolating between brightness variants at intermediate levels.

The condition-fusion module encodes an edge map E, memberships M, gain G, and control map S, with level embeddings included in the control branch. The edge map is extracted from $I_{\mathrm{ph}}(\mu)$ and shared across levels. Spatial and textual conditions enter the editor as

$$\begin{cases} C = \mathcal{F}_\phi(E, M, G, S), \\ e_{\mathrm{txt}} = \mathcal{T}(p_{\mathrm{fix}}, I_{\mathrm{pre}}), \\ R_\omega = \mathcal{C}_\omega(z_t, t, C, e_{\mathrm{txt}}), \\ v_{\omega,\phi} = \mathcal{D}_{\mathrm{frozen}}\left(z_t, t; \mathcal{E}(I_{\mathrm{pre}}), e_{\mathrm{txt}}, R_\omega\right), \end{cases} \tag{12}$$

where $\mathcal{T}$ is the frozen multimodal prompt encoder, $\mathcal{E}$ is the VAE encoder, and $\mathcal{C}_\omega$ is a two-layer ControlNet supplying residuals to the frozen editing transformer $\mathcal{D}_{\mathrm{frozen}}$. The fixed prompt p_{fix} requests natural detail, faithful colors, noise removal, and adherence to spatial illumination control. Its embedding conditions both the ControlNet and transformer, while G and S provide the image-dependent numerical and spatial control.

For training, let $z_0 = \mathcal{E}(I^S_{\mathrm{tar}})$ and $z_t = (1-t)z_0 + t\xi$, with $t \sim \mathcal{U}(0,1)$ and $\xi \sim \mathcal{N}(0, I)$. We optimize a spatially weighted flow-matching objective:

$$\mathcal{L}_{\mathrm{edit}} = \mathbb{E}\left[\frac{1}{D} \sum_i W_i \left\| v_{\omega,\phi}(z_t, t \mid I_{\mathrm{pre}}, C, e_{\mathrm{txt}})_i - (\xi - z_0)_i \right\|_2^2 \right]. \tag{13}$$

Here, i indexes latent tokens and D is their dimension. The weights combine edge agreement, control boundaries, regions outside the semantic masks, enhanced regions, and zero-level regions through $W_i = \sum_r \lambda_r w_{r,i} / \max(1, \sum_j w_{r,j})$. Only the ControlNet and fusion parameters (ω, ϕ) are optimized, while the transformer, VAE, and prompt encoder remain frozen. At inference, user-selected regions and levels replace the random training assignments, and the editor generates I_{out}.

4 EXPERIMENTS

4.1 Experimental Setup

We evaluate a single trained EPAS model on standard low-light enhancement, challenging illumination, and global and local brightness control, without dataset-specific retraining or fine-tuning. See Appendices B and D for implementation and evaluation details.

Mechanistic analysis protocol. We examine the need for within-object gain variation and distinguish action execution from image–gain correspondence through two controlled analyses. Using paired evaluation images, the first analysis compares Full with Object-uniform, which replaces the predicted log-gain field within each selected object by its spatial mean in each color channel. The input, object mask, control map S, prompt, editor weights, and random seed remain fixed. We update both the preview and the gain-conditioning branch consistently after each intervention. Object log-luminance MAE measures agreement with the reference, while the within-object standard deviation of the reference-required log-luminance correction measures correction heterogeneity. We also examine signed errors in fixed shadow, midtone, and highlight groups shared by both settings. The second analysis holds S fixed and scales G by $\alpha \in \{0, 0.5, 1, 1.5\}$, comparing prescribed and realized regional log-luminance changes relative to the same input. The prescribed change is measured from the recomputed preview, and the realized change from the final output. Cross-image gain permutation provides a correspondence control: tracking error is measured against the action actually supplied to each image, whereas reference error is evaluated against that image's reference at $\alpha = 1$. This distinguishes following an action from selecting a suitable action.

Table 1: Quantitative comparison on paired and unpaired low-light benchmarks.

Method	LOLv1	LOLv2-Real	LOLv2-Syn	LSRW	Paired	Unpaired
	PSNR↑/SSIM↑/LPIPS↓	PSNR↑/SSIM↑/LPIPS↓	PSNR↑/SSIM↑/LPIPS↓	PSNR↑/SSIM↑/LPIPS↓	NIQE↓/MUSIQ↑/CLIP-IQA↑	NIQE↓/MUSIQ↑/CLIP-IQA↑
RetinexFormer^{P}	25.160/0.845/0.158	22.820/0.849/0.170	25.670/0.930/0.059	17.984/0.610/0.322	3.985/61.835/0.376	3.382/59.750/0.435
PyDiff^{P}	23.275/0.859/0.108	27.545/0.880/0.138	16.846/0.791/0.260	17.264/0.510/0.335	4.996/67.714/0.437	3.682/65.641/0.467
KinD++^{P}	17.752/0.758/0.198	17.660/0.761/0.216	17.477/0.786/0.232	16.085/0.394/0.366	4.425/67.185/0.471	3.415/65.954/0.503
GSAD^{P}	27.839/0.877/0.091	28.818/0.895/0.095	28.670/0.944/0.047	19.418/0.542/0.386	4.815/68.027/0.508	3.560/62.318/0.529
HVI-CIDNet^{P}	23.593/0.840/0.094	23.760/0.857/0.132	24.638/0.912/0.054	17.668/0.454/0.344	3.971/64.667/0.514	3.177/62.807/0.539
ReDDiT^{P}	21.586/0.824/0.120	27.676/0.879/0.103	17.666/0.775/0.229	18.428/0.466/0.313	4.316/65.003/0.517	3.393/62.470/0.548
CWNet^{P}	23.601/0.848/0.123	21.523/0.612/0.163	24.912/0.910/0.054	19.708/0.582/0.362	4.380/65.377/0.514	3.331/62.832/0.483
PairLIE^{U}	19.514/0.731/0.254	19.885/0.778/0.317	19.074/0.794/0.230	17.602/0.501/0.323	4.470/59.589/0.349	3.874/60.918/0.465
LightenDiffusion^{U}	20.453/0.803/0.192	21.256/0.827/0.173	17.886/0.756/0.248	18.555/0.539/0.311	4.015/59.834/0.335	3.264/60.663/0.492
QuadPrior^{U}	18.777/0.779/0.208	20.471/0.810/0.200	16.103/0.750/0.253	16.947/0.549/0.382	4.958/58.794/0.380	4.276/62.661/0.475
AGLLDiff^{U}	19.836/0.806/0.193	19.990/0.835/0.192	18.435/0.791/0.261	17.359/0.544/0.364	6.000/67.110/0.333	4.490/65.098/0.445
DI-Retinex^{U}	20.505/0.771/0.246	21.368/0.799/0.203	17.527/0.727/0.293	16.190/0.514/0.411	4.420/52.572/0.425	4.683/51.307/0.384
LASQ^{U}	18.609/0.775/0.215	21.339/0.839/0.182	18.370/0.791/0.240	18.915/0.531/0.348	3.986/58.757/0.348	3.202/61.932/0.494
S^{2}UNet^{U}	20.427/0.783/0.229	18.648/0.769/0.284	18.754/0.794/0.243	17.589/0.517/0.391	4.306/55.090/0.355	3.410/60.662/0.488
JarvisIR^{S}	17.661/0.760/0.256	18.987/0.792/0.226	22.409/0.896/0.073	17.118/0.473/0.338	4.142/56.643/0.480	3.305/62.132/0.545
AgenticIR^{P}	18.452/0.791/0.217	19.634/0.798/0.215	20.543/0.792/0.376	17.819/0.520/0.484	4.431/58.759/0.486	3.937/54.582/0.506
RAR^{P}	19.631/0.774/0.254	23.865/0.749/0.304	12.894/0.537/0.309	16.729/0.526/0.631	6.039/55.037/0.531	5.092/52.670/0.602
Ours^{S}	20.516/0.824/0.161	21.389/0.820/0.175	21.232/0.801/0.215	18.706/0.540/0.303	3.982/68.076/0.541	3.071/67.427/0.554

^{P} denotes methods trained with paired target-domain data. ^{U} denotes methods trained without paired target-domain data, including unpaired, zero-reference, and training-free approaches. ^{S} denotes training solely on synthetic paired data without using the training splits of the evaluation benchmarks. Paired averages metrics across LOLv1, LOLv2-Real, LOLv2-Syn, and LSRW. Unpaired reports pooled results on VV, DICM, NPE, LIME, and MEF. Red and blue indicate the best and second-best results for each metric.

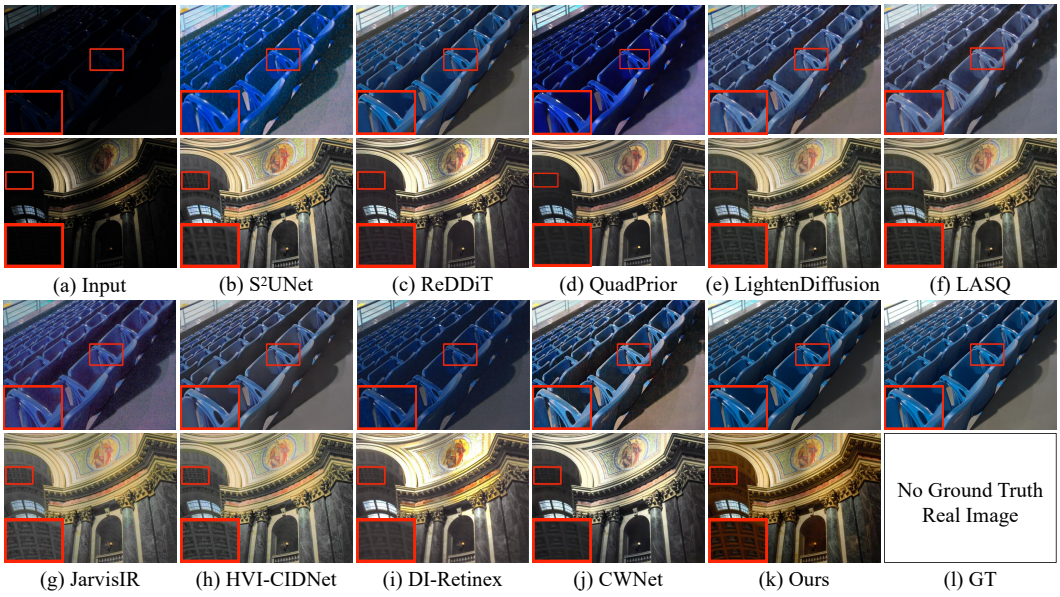


Figure 3: Qualitative comparison of EPAS and competitive methods.

4.2 Performance Evaluation

We summarize the main results below and provide detailed analysis in Appendix E.

Benchmark Evaluation. Without training on the evaluation benchmarks, EPAS leads reference-free baselines on eight of twelve reference-based metrics and achieves the lowest LSRW LPIPS across all methods (Table 1). It ranks first on four of six aggregated reference-free metrics and second on the remaining two, although paired target-domain methods generally achieve higher PSNR and SSIM. Figure 3 shows improved detail visibility with less noise and better-preserved colors. Additional qualitative comparisons are provided in Appendix E.1.

Challenging Illumination. EPAS ranks first on five of nine metrics and second on three (Table 2). Although FLUX-Kontext achieves higher MUSIQ on Light-effect and Extreme DR, Figure 10 in Appendix E.2 shows altered object shapes and background content. EPAS better preserves these

Table 2: Comparison under challenging illumination.

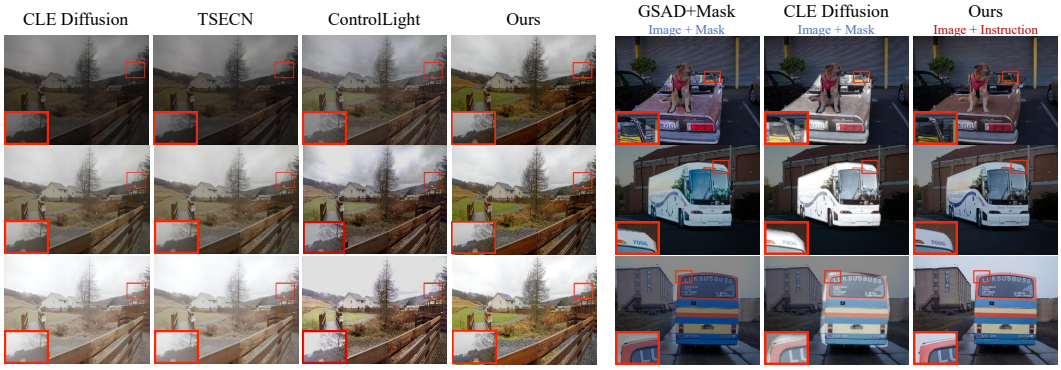
Method	Light-effect	LED	Extreme DR
RetinexFormer	2.78/57.53/0.27	4.60/47.09/0.24	6.96/38.10/0.24
GSAD	2.85/61.51/0.37	4.78/48.15/0.26	5.99/63.35/0.58
RAR	4.18/48.37/0.44	7.88/41.52/0.46	7.13/57.82/0.50
AgenticIR	3.66/65.02/0.21	4.79/51.49/0.20	7.68/28.21/0.29
JarvisIR	2.74/61.43/0.26	4.73/47.62/0.27	5.92/22.46/0.23
FLUX-Kontext	2.90/71.33/0.58	5.57/56.88/0.51	4.95/69.50/0.56
Ours	2.10/67.48/0.43	4.70/60.88/0.52	4.25/65.95/0.61

Entries report NIQE↓/MUSIQ↑/CLIP-IQA↑. Red and blue denote the best and second-best results.

Table 3: Brightness control experiments.

Method	Global		Local	
	B-MAE↓	MUSIQ↑	MUSIQ↑	NIQE↓
GSAD+Mask^{M}	N/A	69.710	71.271	3.507
CLE Diffusion^{M}	0.133	45.342	64.258	3.511
ControlLight	0.183	68.411	N/A	N/A
TSECN	0.177	67.183	N/A	N/A
Ours^{T}	0.101	74.034	75.430	3.503

M denotes user-provided masks, and T denotes text instructions without user-provided masks. Red and blue denote the best and second-best results.



(a) Three-level global enhancement (b) Local enhancement with mask or language input

Figure 4: Comparison of three-level global enhancement and region-selective local enhancement.

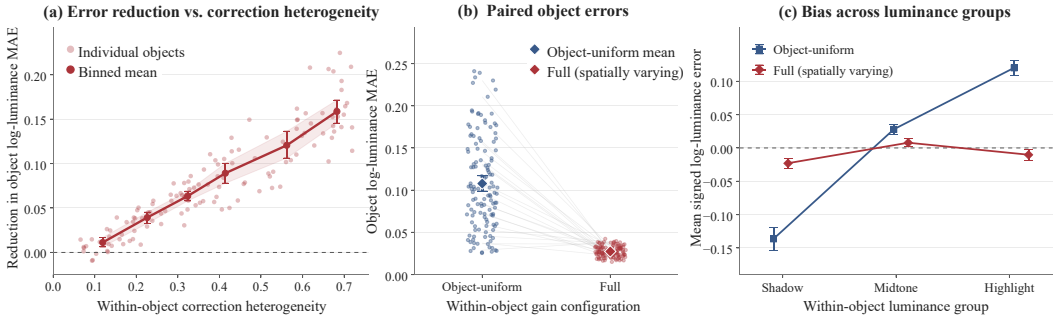


Figure 5: Within-object gain variation. (a) MAE reduction over Object-uniform versus correction heterogeneity. (b) Paired object log-luminance errors under Object-uniform and Full. (c) Signed log-luminance errors in shadow, midtone, and highlight regions, where negative and positive values indicate outputs darker and brighter than the reference, respectively.

structures while revealing dark regions, illustrating that higher perceptual scores do not necessarily imply greater content fidelity.

Brightness Control. EPAS achieves the lowest global B-MAE and the highest global and local MUSIQ among the compared methods (Table 3). For local enhancement, it also obtains a slightly lower NIQE while requiring no user-provided masks. Figure 4 illustrates gradual brightness changes with retained texture and smooth local transitions.

4.3 Mechanistic Analysis

Within-object heterogeneity. Object-uniform preserves each object's mean gain but removes its spatial variation. Figure 5 (a,b) shows that Full reduces object errors, with larger benefits at higher correction heterogeneity. In (c), Object-uniform undercorrects shadows and overcorrects highlights, whereas Full reduces both biases. These trends link the improvement to spatial allocation of correction within objects, whose semantic identity does not imply uniform photometric requirements.

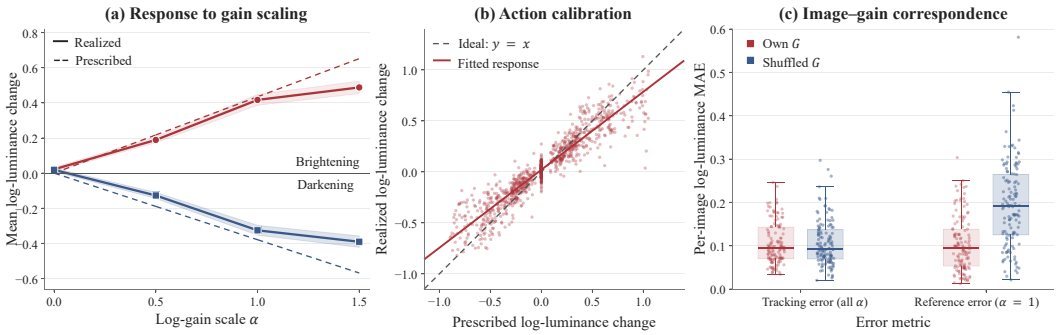


Figure 6: Action execution and image–gain correspondence with fixed S. (a) Prescribed (preview) and realized (output) regional log-luminance changes under αG, relative to the input in brightening and darkening regions. (b) Realized versus prescribed changes. (c) Tracking errors against supplied actions across all α and reference errors at $\alpha = 1$.

Table 4: Quantitative results of ablation studies.

Variant	LOLv2-Real			LED	COCO val2017 (Local)	
	PSNR↑	LPIPS↓	MUSIQ↑	MUSIQ↑	MUSIQ↑	$\mathrm{MAE}_{\mathrm{out}}$ ↓
Qwen-Image-Edit-2511 only	10.707	0.749	42.307	53.247	65.382	0.098
Without action-derived inputs	20.124	0.287	66.215	56.862	71.751	**0.067**
Without direct gain conditioning	21.339	0.179	69.848	60.773	74.608	0.082
Greedy search instead of MH	20.956	0.181	68.950	59.242	74.194	0.077
Full EPAS	**21.389**	**0.175**	**69.982**	**60.884**	**75.430**	0.073

Action execution and gain correspondence. With S fixed, Figure 6 (a,b) shows that outputs respond to the direction and magnitude of G, although stronger actions are attenuated. In (c), gain shuffling leaves tracking errors similar but increases reference errors, indicating that the pipeline continues to follow actions that are less suitable for the recipient image. Thus, accurate restoration requires both an appropriate gain field and its execution through the preview-conditioned pipeline. Detailed analysis and metric descriptions are provided in Appendix E.4.

4.4 Ablation Studies

We evaluate restoration on LOLv2-Real and LED, and local control on COCO val2017. The action-input ablation replaces the preview with the low-light image and removes G and action-derived auxiliary conditions while retaining spatial control. The direct-gain ablation removes only G from condition fusion, retaining the preview and all other inputs. Both match Full's training data, targets, and budget. Greedy search replaces MH under the same candidate-evaluation budget and quality reranking. Qwen-only is unadapted and receives text instructions without spatial control maps.

Removing action-derived inputs reduces LOLv2-Real PSNR by 1.27 dB and LED MUSIQ by 4.02 (Table 4), supporting their value beyond editor adaptation. The preview retains most performance, while direct gain conditioning improves local MUSIQ. MH outperforms greedy search under the matched budget. Full improves local MUSIQ and reduces region-exterior error over Qwen-only, whereas removing action-derived inputs yields lower exterior error at the cost of output quality.

5 Conclusion

We presented EPAS, an executable photometric action space linking semantic intent to input-dependent correction for LLIE. Separating inferred gains from spatial control enables selective and graded enhancement. Training-time action search is distilled into direct prediction for generative restoration. Controlled experiments support spatially varying corrections, indicating that semantic selection and photometric correction are complementary rather than interchangeable coordinates for restoration. A single model achieves strong perceptual quality and global/local control across benchmarks and challenging illumination without dataset-specific retraining or fine-tuning.

AI use statement

We used GPT-6 (OpenAI) to assist with language editing and manuscript organization. The model also provided feedback on methodological descriptions, assisted with checking their consistency with implementation code, and supported the interpretation and presentation of experimental results. The authors retain responsibility for assessing these suggestions and verifying the accuracy of the manuscript. We take full responsibility for the final content, including the methodology, experimental results, conclusions, and any AI-assisted text or artifacts.

Ethics statement

This work investigates controllable low-light image enhancement, with the aim of improving image visibility and giving users greater flexibility in adjusting global and local brightness. The research focuses on methodological development, using existing image datasets and training examples derived through synthetic degradation and brightness adjustment. It does not involve studies with human participants or the collection of new personal data. Within the scope of the experiments and intended use, we do not anticipate substantial ethical concerns beyond those generally associated with image enhancement research. We encourage responsible use of the proposed method and adherence to the usage terms of the underlying datasets and pretrained models.

Reproducibility statement

To support reproducibility, we describe the model architecture, training procedures, and evaluation protocols in the relevant sections of the paper, with additional details in the Appendix. Upon acceptance, we will make the complete source code, pretrained model weights, and configuration files publicly available so that others can reproduce our results, examine the implementation, and build on this work.

References

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 1122–1131. IEEE, 2017.

Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE transactions on image processing*, 27(4):2049–2062, 2018.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 12470–12479. IEEE, 2023.

Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *Advances in Neural Information Processing Systems*, 37:110643–110666, 2024.

I-Hsiang Chen, Isma Hadji, Enrique Sanchez, Adrian Bulat, Sy-Yen Kuo, Radu Timofte, Georgios Tzimiropoulos, and Brais Martinez. Restore, assess, repeat: A unified framework for iterative image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15506–15515, June 2026.

Daniel Feijoo, Juan C Benito, Alvaro Garcia, and Marcos V Conde. Darkir: Robust low-light image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10879–10889. IEEE, 2025.

Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22252–22261. IEEE, 2023.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1777–1786. IEEE, 2020.

Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation*, 90:103712, 2023.

Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36:79734–79747, 2023.

Guoli Jia, Yisheng Zhang, Haote Hu, Shanxu Zhao, Kaikai Zhao, Long Sun, Xinwei Long, Kai Tian, Che Jiang, Zhaoxiang Liu, et al. Tir-agent: Training an explorative and efficient agent for image restoration. *arXiv preprint arXiv:2603.27742*, 2026.

Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *European conference on computer vision*, pp. 161–179. Springer, 2024.

Xu Jiang, Gehui Li, Bin Chen, and Jian Zhang. Multi-agent image restoration. *International Journal of Computer Vision*, 134(5):205, 2026.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

Shuai Jin, Yuhua Qian, Feijiang Li, Guoqing Liu, and Xinyan Liang. Pasd: A pixel-adaptive swarm dynamics approach for unsupervised low-light image enhancement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9070–9079. IEEE, 2025.

Yeying Jin, Wenhan Yang, and Robby T Tan. Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *European conference on computer vision*, pp. 404–421. Springer, 2022.

Derong Kong, Zhixiong Yang, Shengxi Li, Shuaifeng Zhi, Li Liu, Zhen Liu, and Jingyuan Xia. Luminance-aware statistical quantization: Unsupervised hierarchical learning for illumination enhancement. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 116286–116313. Curran Associates, Inc., 2025. doi: 10.52202/085713-3876.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Guanzhou Lan, Qianli Ma, Yuqi Yang, Zhigang Wang, Dong Wang, Xuelong Li, and Bin Zhao. Efficient diffusion as low light enhancer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21277–21286. IEEE, 2025.

Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing*, 22(12):5372–5384, 2013.

Bingchen Li, Xin Li, Yiting Lu, and Zhibo Chen. Hybrid agents for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22636–22647, 2026.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 36802–36827. Curran Associates, Inc., 2025. doi: 10.52202/085713-1237.

Lei Liang, Zhihua Chen, Lei Dai, Ruoxin Chen, Yunyi Zhang, and Xufeng He. A two-stage exposure controllable network for image enhancement with low-light degradation. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22369–22380. IEEE, 2025a.

Yunlong Lin, Tian Ye, Sixiang Chen, Zhenqi Fu, Yingying Wang, Wenhao Chai, Zhaohu Xing, Wenxue Li, Lei Zhu, and Xinghao Ding. Aglldiff: Guiding diffusion models towards unsupervised training-free real-world low-light image enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 5307–5315, 2025b.

Jia Liu, Yu Luo, Guanghui Yue, Jie Ling, Liang Liao, Chia-Wen Lin, Guangtao Zhai, and Wei Zhou. Self-supervised unfolding network with shared reflectance learning for low-light image enhancement. *IEEE Transactions on Image Processing*, 35:800–815, 2026.

Jianglin Lu, Yuanwei Wu, Ziyi Zhao, Hongcheng Wang, Felix Jimenez, Abrar Majeedi, and Yun Fu. Restore-r1: Efficient image restoration agents via reinforcement learning with multimodal llm perceptual feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8629–8639, June 2026.

Aashish Sharma and Robby T Tan. Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11972–11981. IEEE, 2021.

Shangquan Sun, Wenqi Ren, Jingyang Peng, Fenglong Song, and Xiaochun Cao. Di-retinex: Digital-imaging retinex model for low-light image enhancement. *International Journal of Computer Vision*, 133(12):8293–8314, 2025.

Vassilios Vonikakis, Rigas Kouskouridas, and Antonios Gasteratos. On the evaluation of illumination compensation algorithms. *Multimedia Tools and Applications*, 77(8):9211–9231, 2018.

Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013.

Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26057–26066. IEEE, 2024.

Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Exposurediffusion: Learning to expose for low-light image enhancement. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12404–12414. IEEE, 2023.

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

Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2755–2764. IEEE, 2020.

Yanyan Wei, Yilin Zhang, Huan Zheng, Jiahuan Ren, Xiaogang Xu, Zenglin Shi, Zhao Zhang, and Meng Wang. Iamagent: Towards an interactive and adaptive multi-agent system for image restoration. *IEEE Transactions on Image Processing*, 2026.

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.

Tianhe Wu, Jian Zou, Jie Liang, Lei Zhang, and Kede Ma. Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 88167–88190. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2947.

Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5891–5900. IEEE, 2022.

Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5678–5687. IEEE, 2025.

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.

Yufeng Yang, Jianzhuang Liu, Jisheng Chu, Yuqi Peng, Xianfang Zeng, Jiancheng Huang, and Shifeng Chen. Controllight: Towards controllable, consistent, and generalizable low-light enhancement. *arXiv preprint arXiv:2605.25569*, 2026.

Yuyang Yin, Dejia Xu, Chuangchuang Tan, Ping Liu, Yao Zhao, and Yunchao Wei. Cle diffusion: Controllable light enhancement diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 8145–8156, 2023.

Miao Zhang, Jun Yin, Pengyu Zeng, Yiqing Shen, Shuai Lu, and Xueqian Wang. Tscnet: A text-driven semantic-level controllable framework for customized low-light image enhancement. *Neurocomputing*, 625:129509, 2025a.

Tongshun Zhang, Pingping Liu, Yubing Lu, Mengen Cai, Zijian Zhang, Zhe Zhang, and Qiuzhan Zhou. Cwnet: Causal wavelet network for low-light image enhancement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8789–8799. IEEE, 2025b.

Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International journal of computer vision*, 129(4):1013–1037, 2021.

Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023.

Yingjie Zhou, Jiezhang Cao, Farong Wen, Zicheng Zhang, Yu Zhou, Yue Shi, Xiaohong Liu, Radu Timofte, Luc Van Gool, and Guangtao Zhai. Q-agent: Quality-driven chain-of-thought image restoration agent through robust multimodal large language model. *arXiv preprint arXiv:2504.07148*, 2025.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57985–58013, 2025.

Yushen Zuo, Qi Zheng, Mingyang Wu, Xinrui Jiang, Renjie Li, Jian Wang, Yide Zhang, Gengchen Mai, Lihong Wang, James Zou, Xiaoyu Wang, Ming-Hsuan Yang, and Zhengzhong Tu. 4kagent: Agentic any image to 4k super-resolution. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 164106–164175. Curran Associates, Inc., 2025. doi: 10.52202/085713-5476.

APPENDIX

A LIMITS OF LANGUAGE-ONLY PHOTOMETRIC CONTROL

Experimental setup. We examine whether language-guided editing produces input-dependent corrections aligned with a reference-defined brightness target. The language-only evaluation uses 30 normally exposed reference images X_i, each scaled to five intensity levels:

$$X_i^{(c)} = \text{clip}(cX_i, 0, 1), \qquad c \in \mathcal{C} = \{0.5, 0.7, 1, 1.4, 2\}. \tag{A.1}$$

Here, $c < 1$ and $c > 1$ produce darker and brighter inputs, respectively, while $c = 1$ preserves the reference. This controlled perturbation probes photometric response rather than performance on realistic low-light degradation.

We evaluate Qwen-Image-Edit-2511 with two instructions. **Prompt A**, "Brighten this photo, recover shadow details, reduce noise, keep colors natural.", requests directional brightening. **Prompt B**, "Adjust this photo to normal exposure.", specifies a nominal target without prescribing the correction magnitude. The Qwen runs use 40 inference steps, a guidance scale of 4.0, and seed 42. Figure 7 contrasts these responses with EPAS, which supplies an explicit gain field and its photometric preview.

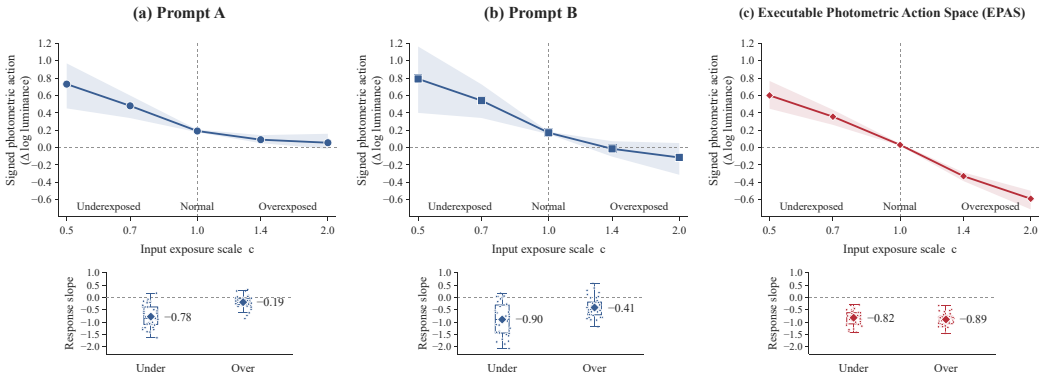


Figure 7: Photometric responses to input intensity scaling. (a,b) Qwen-Image-Edit-2511 under Prompt A and Prompt B. (c) EPAS with explicit photometric actions. Top panels show mean signed log-luminance changes. Bottom panels show per-image response slopes against $\log c$.

Photometric response. Let $Z_{i,m}^{(c)}$ denote the output for image i under setting $m \in \{\text{A}, \text{B}, \text{EPAS}\}$. Using luminance $Y = 0.2126R + 0.7152G + 0.0722B$, we define the signed and absolute log-luminance changes as

$$\begin{aligned} g_{i,m}(c) &= \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \log \frac{Y(Z_{i,m}^{(c)})(p) + \epsilon}{Y(X_i^{(c)})(p) + \epsilon}, \\ a_{i,m}(c) &= \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \left| \log \frac{Y(Z_{i,m}^{(c)})(p) + \epsilon}{Y(X_i^{(c)})(p) + \epsilon} \right|, \end{aligned} \tag{A.2}$$

where Ω_i is the image domain and $\epsilon > 0$ ensures numerical stability. Positive and negative $g_{i,m}$ indicate net brightening and darkening, respectively. The absolute measure $a_{i,m}$ captures changes that may cancel in the signed average.

For each image, we fit separate responses over $\mathcal{C}_- = \{0.5, 0.7, 1\}$ and $\mathcal{C}_+ = \{1, 1.4, 2\}$:

$$(\hat{k}_{i,m}^s, \hat{b}_{i,m}^s) = \arg\min_{k,b} \sum_{c \in \mathcal{C}_s} \left[g_{i,m}(c) - k \log c - b \right]^2, \qquad s \in \{-, +\}. \tag{A.3}$$

Without clipping and with negligible stabilization effects, exact recovery of X_i gives $g^*(c) = -\log c$, hence $k^* = -1$ and $g^*(1) = 0$. These characterize source recovery, rather than a universal criterion

for following language instructions. In particular, Prompt A does not explicitly request recovery of the reference, and clipping limits the applicability of $k^* = -1$ on the brighter-input side.

Results and interpretation. Figure 7 (a,b) shows that both prompts adapt their actions to the input state. Prompt A responds strongly to darker inputs but provides little darkening feedback, consistent with its directional wording. Prompt B elicits a stronger bidirectional response, yet its near-ideal slope on darker inputs coexists with a positive action at $c = 1$. Thus, responding to input brightness does not by itself ensure alignment with the reference-defined target.

In Figure 7 (c), EPAS exhibits a more balanced response across darker and brighter inputs, with a signed action closer to zero at $c = 1$. The improvement is therefore reflected in both the brighter-input response and target alignment, rather than simply stronger enhancement.

This behavior is consistent with EPAS separating semantic intent from quantitative photometric correction. The inferred gain field specifies correction direction and magnitude before generative editing, and the preview conveys the resulting adjustment to the editor. The comparison supports explicit photometric actions as a means of improving correction control beyond the tested language instructions.

B Implementation Details

B.1 Training and Inference Settings

All experiments are implemented in PyTorch and conducted on four NVIDIA A800 GPUs. We use frozen Qwen2.5-VL-32B-Instruct and SAM 2.1 Hiera-B+ for semantic guidance and mask generation. Qwen-Image-Edit-2511 (Wu et al., 2025a) is adapted using a two-layer ControlNet and the proposed condition-fusion module, while its transformer, VAE, and text encoder remain frozen.

The action predictor is trained on low-light inputs synthesized from DIV2K (Agustsson & Timofte, 2017) for 300 epochs using Adam, with a learning rate of 5×10^{-5} and a batch size of 8 per GPU. For editor adaptation, we construct global and regional training examples from Light100K brightness variants (Yang et al., 2026). The ControlNet and condition-fusion module are trained for 200 epochs using AdamW, both with a learning rate of 2×10^{-4} and a weight decay of 10^{-2}. The per-GPU batch size is 1, and bfloat16 precision is used. Inference uses 30 denoising steps and a guidance scale of 1.0. The same trained model is used across evaluation datasets without retraining or fine-tuning.

B.2 Photometric and Sampling Hyperparameters

We use $K = 5$ photometric regions. For (γ, w, g, b, h), the parameter bounds are $\boldsymbol{a}_{\min} = (0.5, 0, 0.5, 0, 0)$ and $\boldsymbol{a}_{\max} = (3.5, 1.5, 2, 1, 1)$. Illumination estimation averages Gaussian-filtered images with kernel sizes 15×15 and 31×31 and standard deviations 5 and 15, respectively. The rendering and log-gain stabilization constants are 10^{-6} and 10^{-3}, and gains are clipped to $[-4, 4]$. Control maps undergo four iterations of input-guided, four-neighbor smoothing with an intensity scale of 0.08. The luminance floor for target interpolation is 10^{-4}.

Action sampling uses $b_0 = 0.55$, $\tau = 0.1$, and $\lambda_{\text{tv}} = \lambda_{\text{pref}} = \lambda_q = 0.5$. TV sums the mean absolute horizontal and vertical image differences. Proposal scales are computed as $s_k = \text{clip}(0.35 \exp(\ell_k/2), 0.02, 0.75)$ and remain fixed within each chain. Starting from μ, the sampler cycles through individual regions, discards five transitions, and retains four consecutive states. The initial prediction is also included in candidate selection. Teacher weights $\lambda_\Theta = 0.02$ and $\lambda_I = 0.05$ are linearly increased over the first five epochs. The Gaussian KL regularizer uses an effective coefficient of 10^{-6} on the summed divergence. For editor adaptation, the base flow-matching term has weight 1, and the boundary, region-exterior, controlled-region, and identity terms each have weight 0.2.

C Training and Inference Procedures

Tables 5 and 6 summarize action learning, editor adaptation, and inference. Let f_ψ denote the action predictor and V the frozen vision-language model.

Table 5: Training-time regional action search and distillation. B and N denote the burn-in length and number of retained states.

Input: low-light image I_{low}; predictor f_ψ; frozen V.
Output: updated predictor parameters ψ.

Initialization and regional assessment

Extract frozen vision-language features and predict (M, μ, ℓ) using f_ψ.
Render $I_0 = I_{\text{ph}}(\mu)$ using Eq. (4).
Query V with **P1**, I_0, and a colored membership overlay; obtain brightness preferences β and regional repair goals r.
Set $s = \text{clip}(\eta \exp(\ell/2), s_{\min}, s_{\max})$, $\Theta \leftarrow \text{sg}(\mu)$, and $\mathcal{C} \leftarrow [\Theta]$. Hold M, β, and s fixed during sampling.

Regional MH search without gradient tracking

For $n = 1, \ldots, B + N$ **do**
  Set $k = 1 + ((n-1) \bmod K)$ and copy $\Theta' \leftarrow \Theta$.
  Draw $\xi \sim \mathcal{N}(0, I_5)$ and set $\vartheta'_k \leftarrow \vartheta_k + s_k \odot \xi$; leave other regions unchanged.
  Compute $\Delta E = E(\Theta') - E(\Theta)$ using Eq. (6).
  Draw $u \sim \mathcal{U}(0,1)$. If $\log u < \min(0, -\Delta E/\tau)$, set $\Theta \leftarrow \Theta'$; otherwise retain Θ.
  If $n > B$, append a copy of the current Θ to $\mathcal{C}$.
End for

Candidate selection and predictor update

Query V with **P2**, repair goals r, and the rendered candidates in $\mathcal{C}$ to obtain quality scores q.
Standardize scores within the pool as $\widehat{q} = (q - \overline{q}) / \max(\text{std}(q), 10^{-6})$.
Select $\Theta^\star = \arg\min_{\Theta \in \mathcal{C}} [E(\Theta) - \lambda_q \widehat{q}(\Theta)]$, preferring the initial state in a tie.
Update ψ using the teacher loss in Eq. (9) and the Gaussian KL regularizer; stop gradients through the selected targets.

Action learning. Regional preferences are obtained from the initial rendering and held fixed during each MH chain, together with the memberships and proposal scales. Quality assessment is applied after sampling and does not enter the acceptance test. Retained states include repeated states following rejected proposals; the initial prediction is also eligible for selection. The selected parameters and rendering are detached when supervising the predictor.

Semantic control and editor adaptation. Semantic masks are obtained by grounding objects in $I_{\text{ph}}(\mu)$ with **P3** and passing the returned boxes to SAM 2.1. They can be cached and are distinct from the photometric memberships M. The same control map determines the preview and the brightness-variant training target. The edge condition is extracted from $I_{\text{ph}}(\mu)$ and remains unchanged across control levels. Only the ControlNet and condition-fusion parameters are updated during editor adaptation.

D Performance Evaluation Protocol

D.1 Datasets and Baselines

Standard enhancement. We evaluate on LOLv1, LOLv2-Real, LOLv2-Syn, and LSRW (Wei et al., 2018; Yang et al., 2021; Hai et al., 2023), together with VV, DICM, NPE, LIME, and MEF as an unpaired test set (Vonikakis et al., 2018; Lee et al., 2013; Wang et al., 2013; Guo et al., 2016). Comparisons include representative LLIE methods (Cai et al., 2023; Zhou et al., 2023; Zhang et al., 2021; Hou et al., 2023; Yan et al., 2025; Lan et al., 2025; Zhang et al., 2025b; Fu et al., 2023; Jiang et al., 2024; Wang et al., 2024; Lin et al., 2025b; Sun et al., 2025; Kong et al., 2025; Liu et al., 2026) and restoration agents JarvisIR, AgenticIR, and RAR (Lin et al., 2025a; Zhu et al., 2025; Chen et al., 2026).

Challenging illumination. We use Light-effect, LED, and Extreme Dynamic Range (Jin et al., 2022; Sharma & Tan, 2021; Wei et al., 2020), comparing representative LLIE methods, restoration agents, and FLUX.1 Kontext-dev (Labs et al., 2025). FLUX-Kontext receives the fixed prompt "Brighten

Table 6: Editor adaptation and inference with spatial brightness control. The trained action predictor and pretrained backbone components remain frozen.

Preparation for each input image	
1	Predict M and μ directly and render $I_{\mathrm{ph}}(\mu)$; no MH search is used.
2	Compute G using Eq. (1), clip it to $[-g_{\max}, g_{\max}]$, and extract edges E from $I_{\mathrm{ph}}(\mu)$.
3	For regional control, obtain semantic labels and masks $\{B_j\}$ from cached results or **P3** followed by SAM 2.1.
Editor adaptation on constructed brightness variants	
4	For a global example, choose a level s and set $S = s$. For a regional example, assign $s_j \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ to each semantic mask, ensuring at least one nonzero assignment.
5	For regional examples, normalize mask overlaps and smooth the control boundaries using Eq. (10).
6	Form $I_{\mathrm{pre}} = \mathrm{clip}_{[0,1]}(I_{\mathrm{low}} \odot \exp(S \odot G))$. Construct I^S_{tar} from the same S using Eq. (11).
7	Compute $C = \mathcal{F}_\phi(E, M, G, S)$ and $e_{\mathrm{txt}} = \mathcal{T}(p_{\mathrm{fix}}, I_{\mathrm{pre}})$, where p_{fix} is **P4**.
8	Encode $z_0 = \mathcal{E}(I^S_{\mathrm{tar}})$. Draw $t \sim \mathcal{U}(0,1)$ and $\xi \sim \mathcal{N}(0, I)$; set $z_t = (1-t)z_0 + t\xi$.
9	Predict the flow using Eq. (12) and update only (ω, ϕ) with Eq. (13).
Inference with user-selected regions and levels	
10	Obtain global or regional levels from the user request; map level IDs $\{0,1,2,3\}$ to strengths $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$.
11	Construct S as in steps 4–5, replacing random assignments with the requested levels.
12	Recompute I_{pre}, C, and e_{txt} using the requested S and the fixed restoration prompt **P4**.
13	Run the conditioned editor with 30 denoising steps and guidance scale 1.0; decode and return I_{out}.

this photo to a normal lighting level." Unadapted Qwen-Image-Edit-2511 is additionally evaluated in the ablation study.

Brightness control. Three-level global control is evaluated on SICE Part 2 (Cai et al., 2018), with CLE Diffusion, ControlLight, and TSECN as baselines (Yin et al., 2023; Yang et al., 2026; Liang et al., 2025). Local enhancement is evaluated on COCO val2017 using its instance masks (Lin et al., 2014), with GSAD+Mask and CLE Diffusion as mask-based baselines. EPAS uses text instructions without user-provided masks. Unsupported control settings are marked N/A in Table 3. TSCNet (Zhang et al., 2025a) is not included in the quantitative comparison because we could not locate a publicly available implementation at the time of writing.

D.2 Evaluation Metrics

Restoration quality. We report PSNR, SSIM, and LPIPS on paired benchmarks, and NIQE, MUSIQ, and CLIP-IQA for reference-free quality assessment. Higher PSNR, SSIM, MUSIQ, and CLIP-IQA and lower LPIPS and NIQE indicate better performance. In Table 1, the Paired column averages reference-free scores across LOLv1, LOLv2-Real, LOLv2-Syn, and LSRW; the Unpaired column reports pooled results on VV, DICM, NPE, LIME, and MEF. Challenging-illumination results use NIQE, MUSIQ, and CLIP-IQA for each dataset separately.

Global brightness accuracy. B-MAE measures the absolute difference between output and reference mean luminance at each requested level. For an RGB image I normalized to $[0,1]$, mean luminance is

$$\overline{Y}(I) = \frac{1}{HW} \sum_p \left(0.299 I_R(p) + 0.587 I_G(p) + 0.114 I_B(p)\right). \tag{A.4}$$

For N scenes and three brightness levels,

$$\text{B-MAE} = \frac{1}{3N} \sum_{n=1}^{N} \sum_{l=1}^{3} \left| \overline{Y}\left(I^{[l]}_{\mathrm{out},n}\right) - \overline{Y}\left(I^{[l]}_{\mathrm{ref},n}\right) \right|. \tag{A.5}$$

Luminance is computed from decoded RGB values without linearization. Lower B-MAE indicates closer agreement with the reference brightness levels; MUSIQ measures perceptual quality.

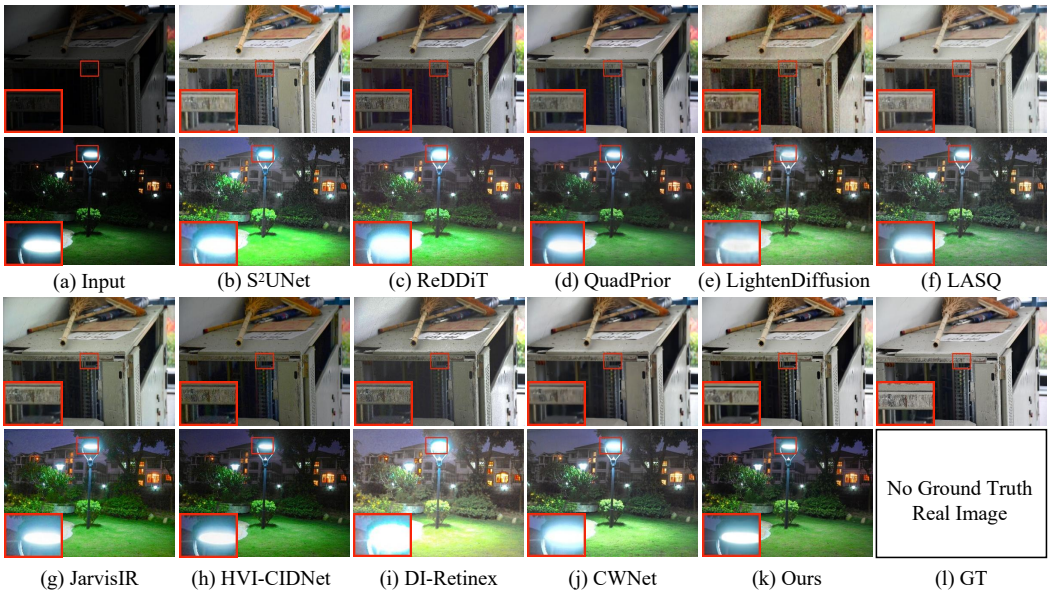


Figure 8: Additional qualitative comparisons of EPAS and competing methods on standard low-light enhancement benchmarks (I).

Region-exterior preservation. Let $B_n(p) \in \{0, 1\}$ be the evaluation mask for image n, with 1 indicating the selected region. Changes outside this region are measured relative to the input:

$$\mathrm{MAE}_{\mathrm{out}} = \frac{1}{N} \sum_{n=1}^{N} \frac{\sum_p \sum_{c=1}^{3} (1 - B_n(p)) \, |I_{\mathrm{out},n,c}(p) - I_{\mathrm{low},n,c}(p)|}{3 \sum_p (1 - B_n(p))}. \tag{A.6}$$

RGB values are normalized to $[0, 1]$. Errors are averaged over region-exterior pixels and channels within each image, then equally across images. Lower values indicate better preservation outside the selected region. MUSIQ assesses the perceptual quality of locally enhanced outputs.

E Detailed Performance Analysis

E.1 Benchmark Evaluation

As shown in Table 1, EPAS achieves competitive restoration quality without using the training splits of the evaluation benchmarks. Supervised baselines trained with paired target-domain data generally achieve higher PSNR and SSIM. Compared with unsupervised and other reference-free baselines, EPAS leads on eight of the twelve reference-based metrics and achieves the lowest LPIPS on LSRW across all methods. Its advantage is more consistent in perceptual quality, ranking first on four of the six aggregated reference-free metrics and second on the remaining two. Figure 3 further illustrates these results. In the seating scene, EPAS preserves clear seat boundaries and closely matches the ground-truth colors and brightness, with less visible noise than several competing methods. In the interior scene, it reveals shadowed architectural patterns and richer warm tones while maintaining scene structure. Figures 8 and 9 provide additional qualitative comparisons on the standard enhancement benchmarks.

E.2 Challenging Illumination

Table 2 shows that EPAS ranks first on five of the nine metrics and second on three. FLUX-Kontext is evaluated with the fixed prompt "Brighten this photo to a normal lighting level." Although it obtains higher MUSIQ on Light-effect and Extreme DR, these scores do not necessarily reflect content fidelity. As illustrated in Figure 10, FLUX-Kontext introduces hallucinated content in the still-life scene, altering fruit shapes and replacing the dark background with a bright, smooth

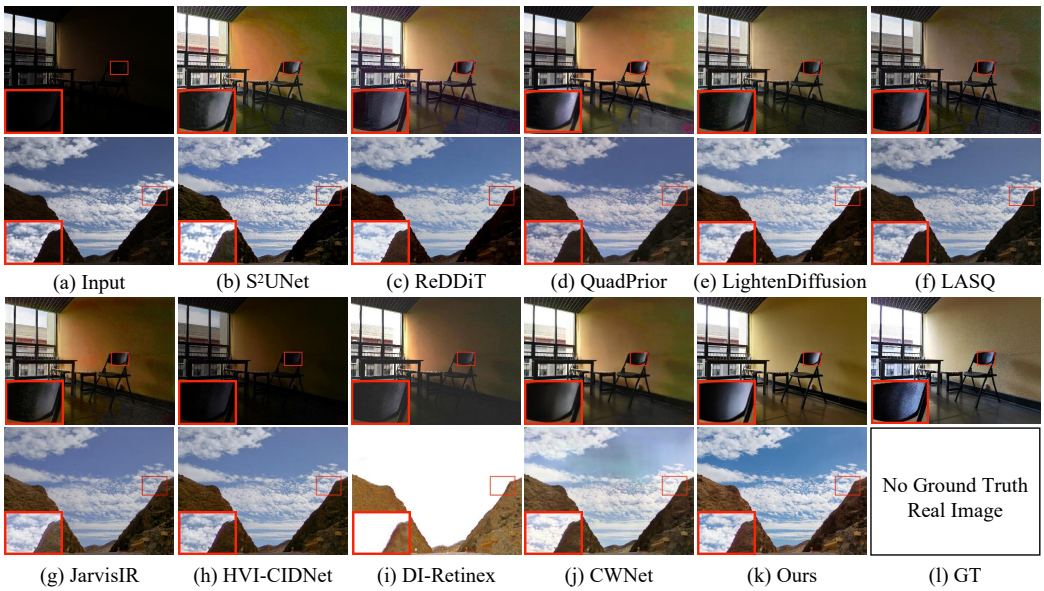


Figure 9: Additional qualitative comparisons of EPAS and competing methods on standard low-light enhancement benchmarks (II).

backdrop. The result appears visually plausible but deviates substantially from the reference scene. In contrast, EPAS reveals the objects while more faithfully preserving their contours, colors, and background. In the nighttime street scene, it improves shadow visibility without the excessive brightening around light sources observed in RetinexFormer, whereas JarvisIR leaves much of the still-life scene underexposed.

In the LED scene, EPAS improves visibility across the image while avoiding visible overexposure around the LED sources. The compared LLIE and agentic baselines overexpose these regions, whereas FLUX-Kontext also avoids visible overexposure but leaves the background insufficiently enhanced. EPAS thus improves background visibility while preserving the local light sources. Together, these comparisons suggest that conditioning restoration on image-dependent photometric actions helps balance visibility enhancement with content preservation under challenging illumination.

E.3 Brightness Control

Table 3 evaluates global and local brightness control under different input requirements. For global enhancement, B-MAE measures the absolute difference between the mean luminance of each output and its corresponding reference level. EPAS achieves the lowest B-MAE and the highest MUSIQ, indicating closer agreement with the reference brightness levels alongside higher perceptual quality. Figure 4 shows that EPAS retains texture and structure across levels with gradual brightness transitions, whereas competing methods exhibit insufficient enhancement, texture loss, or overexposure at the strongest setting. For local enhancement, EPAS achieves the highest MUSIQ of 75.430 using text instructions without user-provided masks. Its NIQE is 3.503, compared with 3.507 for GSAD+Mask and 3.511 for CLE Diffusion. The NIQE scores are close across the three methods, whereas EPAS obtains a higher MUSIQ score. Figure 4 provides qualitative examples of region-selective enhancement and transitions around the selected objects. Region-exterior preservation is evaluated separately in the ablation study using $\mathrm{MAE}_{\mathrm{out}}$ (Table 4).

Figure 11 shows global enhancement at three brightness levels. Increasing the level makes the building facades, parked cars, and pavement more visible, while retaining the overall scene layout. Figure 12 shows complementary regional settings. Assigning level 3 to the non-ground regions brightens the surrounding facades while leaving the ground comparatively dark; reversing the assignments concentrates the enhancement on the pavement. These examples illustrate how the same

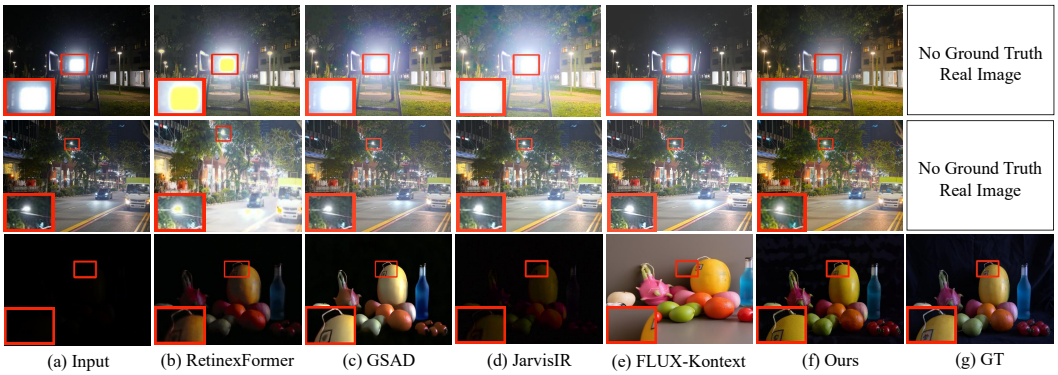


Figure 10: Qualitative comparison under challenging illumination.



Figure 11: Global brightness control with EPAS. Each row shows an input image and its enhanced outputs at levels 1, 2, and 3.

scene can be rendered with different overall brightness levels or different spatial distributions of enhancement.

E.4 Mechanistic Analysis

Within-object correction. Let $L(I,p) = \log(Y(I,p) + \epsilon)$ denote stabilized log-luminance, where Y is the luminance conversion used for evaluation. For a selected object Ω_j and setting $v \in \{\text{Full}, \text{Uniform}\}$, define

$$e_j^v = \frac{1}{|\Omega_j|} \sum_{p \in \Omega_j} |L(I_{\text{out}}^v, p) - L(I_{\text{ref}}, p)|\,, \qquad h_j = \text{Std}_{p \in \Omega_j}\left[L(I_{\text{ref}}, p) - L(I_{\text{low}}, p)\right]. \tag{A.7}$$

Figure 5 (a) plots $\Delta e_j = e_j^{\text{Uniform}} - e_j^{\text{Full}}$ against h_j; positive values indicate lower error with Full. Object MAE is averaged over selected objects. For each fixed luminance group



(a) Input (b) Segmentation (c) Ground: Level 0 Others: Level 3 (d) Ground: Level 3 Others: Level 0

Figure 12: Regional brightness control with EPAS. Each row shows (a) the input, (b) the segmentation overlay, (c) level 0 assigned to the ground and level 3 to the remaining regions, and (d) the reversed assignment. Level 0 specifies no requested photometric enhancement in the corresponding region.

$t \in \{\text{shadow}, \text{midtone}, \text{highlight}\}$, the signed bias is

$$b^v_{j,t} = \frac{1}{|\Omega_{j,t}|} \sum_{p \in \Omega_{j,t}} \left[L(I^v_{\text{out}}, p) - L(I_{\text{ref}}, p) \right]. \tag{A.8}$$

We average these biases over objects with nonempty groups, using identical group assignments across settings. Negative and positive values indicate undercorrection and overcorrection, respectively.

Regional action execution. For image i and a fixed evaluation region R_{ir}, let

$$q_{ir}(I) = \frac{1}{|R_{ir}|} \sum_{p \in R_{ir}} L(I, p). \tag{A.9}$$

The prescribed and realized changes at gain scale α are

$$a_{ir\alpha} = q_{ir}(I^\alpha_{\text{pre}}) - q_{ir}(I_{\text{low}}), \qquad d_{ir\alpha} = q_{ir}(I^\alpha_{\text{out}}) - q_{ir}(I_{\text{low}}). \tag{A.10}$$

For $\mathcal{A} = \{0, 0.5, 1, 1.5\}$ and m_i evaluation regions, per-image tracking and reference errors are

$$T_i = \frac{1}{|\mathcal{A}| m_i} \sum_{\alpha \in \mathcal{A}} \sum_{r=1}^{m_i} |d_{ir\alpha} - a_{ir\alpha}|, \tag{A.11}$$

$$R_i = \frac{1}{m_i} \sum_{r=1}^{m_i} \left| q_{ir}(I^{\alpha=1}_{\text{out}}) - q_{ir}(I_{\text{ref}}) \right|. \tag{A.12}$$

Both metrics are averaged over images. Under gain shuffling, the preview and prescribed action are recomputed using the supplied gain, while the reference remains that of the recipient image. Calibration is summarized by a least-squares fit $d_{ir\alpha} \approx k a_{ir\alpha} + c$, with ideal values $k = 1$ and $c = 0$. Unlike object MAE, these regional metrics take the absolute difference after spatial averaging and are not directly comparable to the object-level values.

Table 7: Mechanistic evaluation summaries.

A. Within-object heterogeneity

Setting	Object MAE↓	Shadow bias	Midtone bias	Highlight bias
Object-uniform	0.108	−0.136	+0.028	+0.120
Full	0.027	−0.023	+0.008	−0.010

B. Action execution and gain correspondence

Gain source	Slope (ideal: 1)	Intercept (ideal: 0)	Tracking MAE↓ (all α)	Reference MAE↓ ($\alpha = 1$)
Own G	0.765	+0.020	0.106	0.103
Shuffled G	0.771	+0.021	0.106	0.194

Table 8: Model complexity, inference cost, and perceptual quality on LSRW.

Method	Params (M)	Latency (s/image)↓	Peak Memory (GiB)↓	LPIPS↓	MUSIQ↑
GSAD	17.434	5.154	6.328	0.386	61.539
LightenDiffusion	27.833	0.679	0.917	0.311	57.427
LASQ	27.833	0.440	0.851	0.348	57.640
RAR	9327.00	16.728	18.936	0.631	55.189
FLUX.1 Kontext-dev	16870.60	31.152	43.829	0.658	62.557
Qwen-Image-Edit-2511	28849.50	38.909	56.720	0.780	45.561
EPAS (ours)	39569.82	45.688	69.467	0.303	68.601

Numerical analysis. Table 7 A reports a decrease in mean object MAE from approximately 0.108 to 0.027. The directional biases help explain this reduction: Object-uniform undercorrects shadows (-0.136) and overcorrects highlights ($+0.120$), whereas Full reduces these biases to -0.023 and -0.010. Together with the increasing benefit at higher heterogeneity, this supports retaining within-object gain variation. The accompanying MAE reduction also shows that the smaller signed biases are not solely due to error cancellation.

In Table 7 B, the Own-G calibration slope of approximately 0.765 indicates attenuated responses to prescribed changes. Own and shuffled gains yield similar mean tracking errors of approximately 0.106, but shuffling increases reference error from 0.103 to 0.194 at $\alpha = 1$. Thus, comparable action tracking can accompany substantially different reference agreement, supporting the importance of matching the gain field to the input. This comparison evaluates correspondence within the complete pipeline and does not isolate the direct gain-conditioning branch.

F Computational Cost and Practical Tradeoffs

Table 8 reports model complexity, inference cost, and perceptual quality on LSRW. EPAS uses approximately 39.57 billion parameters, with a peak memory footprint of 69.467 GiB and a latency of 45.688 seconds per image.

This computational cost accompanies improvements in perceptual quality. EPAS achieves the lowest LPIPS of 0.303 and the highest MUSIQ of 68.601 among the compared methods. Relative to LightenDiffusion, the LPIPS improvement is modest, while MUSIQ increases from 57.427 to 68.601. These metrics assess complementary aspects of reference similarity and perceptual quality. Training-time action search is distilled into direct prediction, eliminating MH sampling and candidate quality assessment at inference.

The reported cost reflects the use of large pretrained components, including a VLM, SAM, and Qwen-Image-Edit, whose semantic grounding and generative restoration capabilities form the basis of the EPAS framework. Our objective is therefore not to develop an efficient LLIE architecture, but to study whether semantic intent and photometric correction can be represented as distinct yet composable factors in restoration. In EPAS, semantic regions and control levels determine where a correction is applied and how strongly it is executed, while the image-dependent gain field determines the photometric correction required by the input. These two factors are complementary rather than interchangeable. Figure 5 shows that regions sharing the same semantic identity can still require substantially different photometric corrections, while Figure 6 shows that the restoration pipeline

can faithfully respond to a supplied action even when that action is poorly matched to the input. Together, these results separate action execution from action suitability and motivate an explicit interface that composes semantic control with image-dependent photometric correction. The results on standard benchmarks, challenging illumination, and brightness control further show that this formulation can be realized as a practical system that supports perceptual restoration, cross-dataset transfer, and global and regional control with a single trained model. Its computational cost is largely associated with the pretrained foundation models used for semantic understanding, region grounding, and generative restoration, while EPAS builds on these capabilities to study explicit semantic and photometric composition rather than lightweight LLIE.