# MR. Illuminate: Zero-Shot Low-Light Image Enhancement with Diffusion Prior

## Supplementary Material

## Contents



# A. Preliminaries

## A.1. Diffusion Models

Diffusion probabilistic models [21, 49] define a forward process that gradually corrupts a clean sample $z_0 \sim p_{\text{data}}(z)$ by adding Gaussian noise over $T$ steps. Let $\{\alpha_t\}_{t=1}^{T}$ denote the cumulative noise schedule. The marginal distribution of the noisy variable at timestep $t$ evaluates to

$$\begin{aligned} q(z_t \mid z_0) &:= \int q(z_{1:t} \mid z_0)\, \mathrm{d}z_{1:(t-1)} \\ &= \mathcal{N}(z_t;\, \sqrt{\alpha_t}\, z_0,\, (1-\alpha_t)\mathbf{I}) \end{aligned} \tag{1}$$

and samples from this marginal follow the reparameterization $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\,\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

Accordingly, the generative process reverses this corruption through a noise-prediction network $\epsilon_\theta(z_t)$ trained to predict the added noise at each timestep by minimizing

$$\min_\theta \; \mathbb{E}_{z_0,\epsilon,t}\big[\, \|\epsilon - \epsilon_\theta(z_t)\|_2^2 \,\big]. \tag{2}$$

For deterministic generation, we use DDIM sampling [50], which replaces the stochastic noise sampling used in DDPM [21] with a deterministic denoising at each timestep:

$$\begin{aligned} z_{t-1} = \sqrt{\alpha_{t-1}} \cdot \frac{z_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(z_t)}{\sqrt{\alpha_t}} \\ + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(z_t). \end{aligned} \tag{3}$$

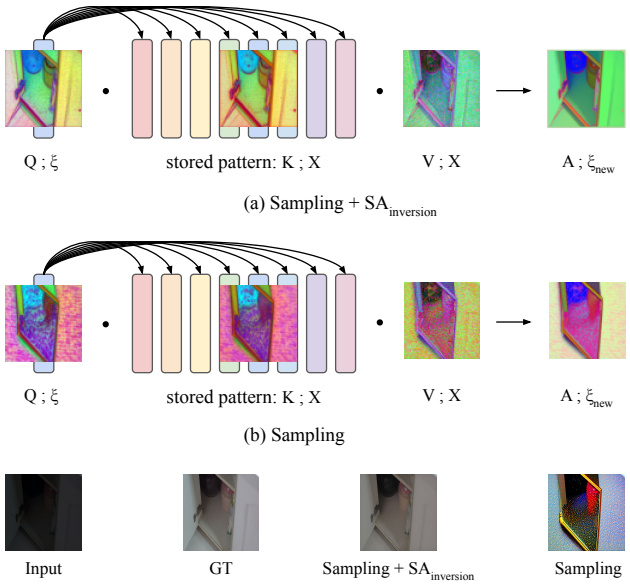


Figure A. *Row 1.* Reusing inversion-time self-attention (SA) preserves the *correct stored patterns*, ensuring that Hopfield updates decrease energy in an input-aligned direction. *Row 2.* Standard sampling generates *drifted stored patterns*, causing Hopfield updates to decrease energy with respect to a misaligned energy landscape and leading to semantic drift. *Row 3.* From left to right: input, ground truth, reconstruction *with* self-attention injection, and reconstruction *without* self-attention injection.

## A.2. DDIM Inversion and Accumulated Error

DDIM inversion provides an image-to-noise mapping by transforming the input $z_0$ to a trajectory $\{z_t\}_0^T$, such that DDIM sampling initialized at $z_T$ yields a close reconstruction of the original $z_0$. Each transition $z_{t-1} \mapsto z_t$ is approximated by

$$\begin{aligned} z_t = \sqrt{\alpha_t} \cdot \frac{z_{t-1} - \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(z_{t-1})}{\sqrt{\alpha_{t-1}}} \\ + \sqrt{1-\alpha_t}\,\epsilon_\theta(z_{t-1}). \end{aligned} \tag{4}$$

The inversion relies on the assumption that the ODE process can be reversed in the limit of small steps. Inversion replaces the actual noise realization $\epsilon_\theta(z_t)$ with $\epsilon_\theta(z_{t-1})$ when computing $z_t$, resulting in an approximate latent whose error accumulates over steps. As a consequence, the terminal latent $z_T$ obtained from inversion differs from the ideal latent that would faithfully reconstruct the input image. Sampling from $z_T$ therefore produces a reconstruction that deviates from the original, as shown in Figure B.

### A.3. Energy-Based Interpretation of Self-Attention

Self-attention is mathematically equivalent to the update step of a modern Hopfield network [43]. Modern Hopfield networks store feature patterns and update a query by moving it toward a similarity-weighted combination of these patterns.

Given stored patterns $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ arranged in matrix form $X = [x_1, \dots, x_N]$ and a query state $\xi \in \mathbb{R}^d$, the model defines an energy function whose gradient indicates how a query should be updated relative to the stored patterns:

$$E(\xi) = -\frac{1}{\beta} \log \Big( \sum_{i=1}^{N} \exp(\beta x_i^\top \xi) \Big) + \tfrac{1}{2}\|\xi\|^2, \qquad (5)$$

where $\beta > 0$ denotes the inverse temperature parameter that determines how strongly the model emphasizes the most similar pattern. Intuitively, the first term encourages the query to align with patterns that have high similarity, while the second term regularizes the magnitude of the query.

The update step in a modern Hopfield network updates the query according to

$$\xi_{t+1} = X \operatorname{softmax}\big(\beta\, X^\top \xi_t\big)\,,$$

which produces a similarity-weighted combination of the stored patterns. This update reduces the Hopfield energy defined in Eq. (5), with the stored patterns determining the energy landscape along which this descent occurs, and its derivation from a convex–concave optimization procedure (CCCP) [65] is presented in the next section.

Transformer self-attention follows the same mathematical form in which keys and values correspond to stored patterns, the query represents the current state, and softmax similarity scores determine how the update combines these patterns:

$$\mathrm{A} = \operatorname{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V, \qquad (6)$$

where $Q, K, V \in \mathbb{R}^{N \times d}$, with $N = H'W'$ denoting the number of tokens and $d$ the token feature dimension. The quantities $H'$ and $W'$ correspond to the height and width of the latent feature map at that attention layer. In this correspondence, the Hopfield parameters $(\beta, X, \xi)$ map directly to the attention parameters $\big(\frac{1}{\sqrt{d}}, K, Q\big)$, with the retrieved state corresponding to $V$:

$$(\xi, X, \beta, \xi_{\text{new}}) \quad \longleftrightarrow \quad (Q, \{K, V\}, 1/\sqrt{d}, \mathrm{A}). \qquad (7)$$

## B. Attention Injection as Hopfield Constraint

The main paper analyzes MR. Illuminate through its two-step structure: a *Modulate* step that adjusts global illumination and color, and a *Refine* step that restores local structure and color through self-attention (SA) injection. Here, we provide a complementary theoretical interpretation of the *Refine* step through the lens of modern Hopfield networks. The purpose of this analysis is not to introduce a new causal mechanism, but to clarify why the Refine step benefits from enforcing the self-attention patterns computed during DDIM inversion.

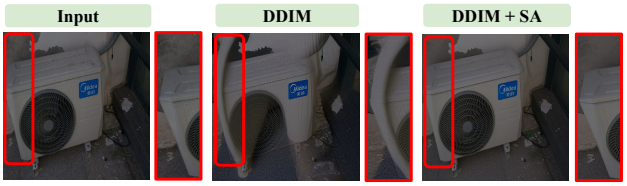


Figure B. **DDIM reconstruction failure.** Left: input image. Middle: reconstruction obtained via DDIM inversion followed by DDIM sampling. Accumulated errors during inversion prevent faithful reconstruction of the input image. Right: reconstruction obtained via DDIM inversion followed by DDIM sampling with self-attention (SA) injection, which mitigates inversion drift.

**Hopfield Perspective.** From Section A.3, attention layers can be viewed as performing a Hopfield-style update: they adjust the current representation so that it becomes more similar to the patterns encoded in the keys and values. Under this viewpoint, the stored patterns define an energy landscape, and the attention update moves the representation downhill on this landscape. This perspective clarifies why self-attention injection is essential because it restores the appropriate stored patterns, ensuring that refinement follows an input-aligned trajectory.

**Self-Attention Derived from DDIM Inversion.** DDIM inversion begins at the observed input image and proceeds step by step toward a noisy latent. The self-attention features computed along this trajectory encode the stored patterns that consistently lead *back* to the input. In Hopfield terms, these maps induce an energy landscape aligned with the input. Reusing them during sampling ensures that the Refine step operates within this input-aligned landscape, helping preserve the input's structural and appearance cues.

**Self-Attention Derived from DDIM Sampling.** As discussed in Section A.2, DDIM inversion is approximate. The terminal latent produced by inversion does not lie exactly on the ideal reconstruction trajectory. When sampling begins from this latent, the attention features generated during sampling encode drifted stored patterns. These features induce a different energy landscape than the one defined during inversion. Although each attention update continues to decrease energy, it does so with respect to this *misaligned* landscape, pulling the refinement away from the input and producing semantic drift (Figure A).

**Self-Attention Injection.** To avoid descending along the drifted landscape, we enforce the sampling process to reuse

the attention features recorded during inversion:

$$\{Q_t, K_t, V_t\}_{\text{samp}} \leftarrow \{Q_t, K_t, V_t\}_{\text{inv}}.$$

This restores the correct stored patterns and thus the input-aligned energy landscape during the Refine step. As a result, each Hopfield-style update moves the latent representation in a direction consistent with the input's spatial structure and color appearance. This explains why self-attention injection stabilizes the Refine step and helps maintain both the form and appearance of the input.

**Formal Justification of the Update Rule.** The central claims above rely on one key property: a Hopfield update (and thus self-attention) decreases the energy defined by the $(Q, K, V)$ patterns used at that timestep. The proposition [43] below formalizes this property. The proof is included to clarify that attention updates are principled energy-descent steps, not heuristic transformations. It also establishes the foundation for understanding why sampling-time attention—defined by a drifted trajectory—leads to semantic drift, and why enforcing inversion-time attention corrects this behavior.

**Proposition 1.** *Each Hopfield update defined by*

$$\xi_{t+1} = X \,\mathrm{softmax}(\beta X^\top \xi_t)$$

*monotonically decreases the Hopfield energy function*

$$E(\xi_{t+1}) \le E(\xi_t),$$

*with equality if $\xi_t$ is already a stationary point.*

*Proof.* We consider a single Hopfield update step from $\xi_t$ to $\xi_{t+1}$. The energy can be decomposed into convex and concave parts through the Concave-Convex Procedure (CCCP) [65]:

$$\begin{aligned} E(\xi) &= E_1(\xi) + E_2(\xi), \qquad E_1(\xi) = \tfrac{1}{2}\|\xi\|^2, \\ E_2(\xi) &= -\tfrac{1}{\beta}\log \sum_{i=1}^{N} e^{\beta x_i^\top \xi}. \end{aligned} \tag{8}$$

Since lse (log-sum-exp) is convex, $E_2$ is concave, and its first-order inequality gives

$$E_2(\xi) \le E_2(\xi_t) + (\xi - \xi_t)^\top \nabla E_2(\xi_t).$$

Thus, the surrogate function

$$E_C(\xi, \xi_t) = E_1(\xi) + E_2(\xi_t) + (\xi - \xi_t)^\top \nabla E_2(\xi_t)$$

satisfies

$$E(\xi) \le E_C(\xi, \xi_t), \qquad E(\xi_t) = E_C(\xi_t, \xi_t).$$

Minimizing this surrogate yields the optimality condition

$$\nabla_\xi E_C(\xi, \xi_t)\big|_{\xi=\xi_{t+1}} = \nabla E_1(\xi_{t+1}) + \nabla E_2(\xi_t) = 0.$$

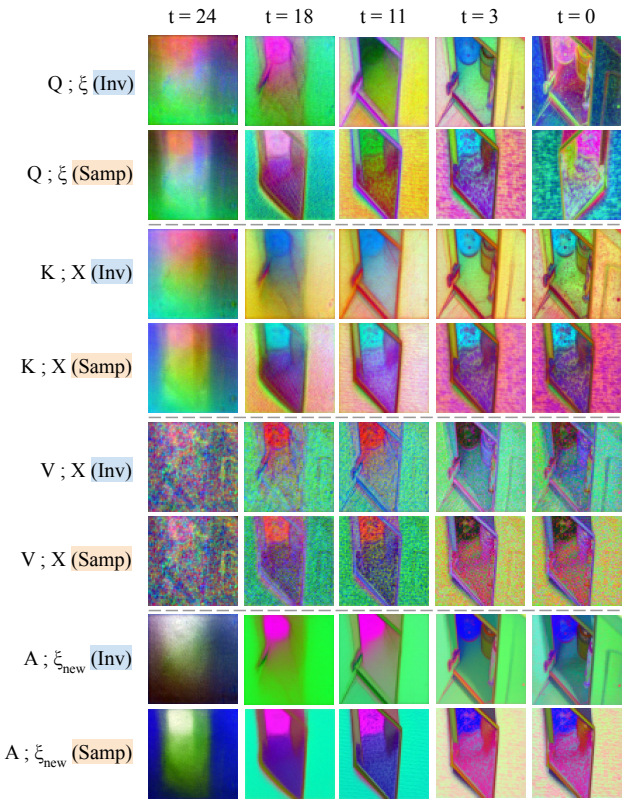


Figure C. PCA visualization of query (Q; $\xi$), key (K; $X$), value (V; $X$), and the attention outputs (A; $\xi_{\text{new}}$) from up-block self-attention layers during DDIM inversion and sampling. Interpreted as Hopfield updates, K and V store the patterns retrieved during each step: *inversion-time stored patterns* remain aligned with the input, whereas *sampling-time stored patterns* drift due to accumulated errors. For the input image and its ground truth, see Figure A.

that is,

$$\xi_{t+1} = X \,\mathrm{softmax}(\beta X^\top \xi_t).$$

Because $E(\xi) \le E_C(\xi, \xi_t)$ for all $\xi$, substituting $\xi = \xi_{t+1}$ gives

$$E(\xi_{t+1}) \le E_C(\xi_{t+1}, \xi_t).$$

Next, since $\xi_{t+1}$ minimizes the surrogate $E_C(\cdot, \xi_t)$, its surrogate value cannot exceed the surrogate at any other point, including $\xi_t$:

$$E_C(\xi_{t+1}, \xi_t) \le E_C(\xi_t, \xi_t).$$

Finally, by construction, the surrogate coincides with the true energy at $\xi_t$,

$$E_C(\xi_t, \xi_t) = E(\xi_t).$$

Combining these three relations yields

$$E(\xi_{t+1}) \;\le\; E_C(\xi_{t+1}, \xi_t) \;\le\; E_C(\xi_t, \xi_t) \;=\; E(\xi_t).$$

This interpretation shows that every attention operation performs an energy-decreasing update step, pulling the state toward configurations defined by the stored keys. By reusing inversion-time triplets $(Q_t, K_t, V_t)_{\text{inv}}$ during sampling, we ensure that each update follows the same input-aligned energy landscape as inversion, thus preserving semantic fidelity and preventing divergence.

**Summary.** By reusing inversion-time attention triplets $(Q_t, K_t, V_t)_{\text{inv}}$, MR. Illuminate retrieves information from a non-drifted, input-aligned state. This preserves the correct spatial and chromatic relationships needed for faithful local reconstruction. Meanwhile, the Modulate step (via AdaIN) adjusts global illumination and color. Together, these components form a coherent mechanism in which AdaIN establishes global consistency, while Hopfield-constrained self-attention enforces locally accurate, drift-free refinement.

# C. Related Works and Comparisons

## C.1. Additional Related Works

**Traditional Methods.** Conventional image enhancement methods, including histogram equalization [11, 41] and gamma correction [42], rely on global adjustments to enhance image contrast. Despite their computational efficiency, these methods are inherently limited by their inability to account for varying scene-specific lighting conditions.

**Supervised Methods.** Supervised methods, trained with paired datasets, learn an explicit image-to-image mapping through direct loss optimization. Among these, Convolutional Neural Networks (CNNs) are adept at learning transformations from under-exposed to well-lit images, effectively capturing local textures and patterns [35, 59, 63, 66]. However, their limitation in capturing long-range dependencies has led to alternative approaches. These include ensemble approaches [3], transformer-based architecture [8, 31, 60, 61], synthetic data augmentation [4, 38], Mixture of Experts-based method [32], frequency-domain methods [14, 28], and event-based illumination estimation [54]. However, performance depends on *dataset diversity and scale*; limited variation in scenes, noise, and lighting reduces generalization.

Recently, generative methods have exhibited promising results in low-light image enhancement, with diffusion models [12, 21, 45, 49, 51, 52] demonstrating particular efficacy because of their strong generative ability, being free from the instability and mode-collapse that are prevalent in previous generative models. However, standard Gaussian noise assumptions of diffusion models do not model the complex noise of low-light images. In response, new training strategies have been proposed for raw and RGB low-light image enhancement [22–24, 34, 39, 58, 64]. However, these methods still remain dependent on supervised learning, without leveraging the generative priors of the pre-trained diffusion models and inheriting the same constraints of paired data and limited dataset diversity.

**Unsupervised Methods.** Unsupervised frameworks [15, 20, 25, 27, 29, 30, 30, 36, 44, 47, 48, 56, 62, 67] remove the need for paired supervision and instead learn *domain-level* correspondences between low- and normal-light distributions. EnlightenGAN [25], NeRCo [62], and FoCo [19] perform adversarial learning on unpaired data, CLIP-LIT [30] leverages CLIP prior and learnable prompt embeddings, PairLIE [15] learns priors from paired low-light images, and RoSe [29] adopts a NeRF-based [37] illuminance-transition model for enhancement. Another category of unsupervised methods is the zero-reference approach, wherein a dataset comprising a single class is leveraged for training. This method capitalizes on the intrinsic color properties of natural images, drawing upon established theoretical frameworks such as Retinex theory [26] and the Kubelka-Munk theory [17]. However, the aforementioned principles may not consistently align with the real-world behavior of noise in under-exposed data. Zero-DCE [20] and Zero-DCE++ [27] employ neural networks to estimate the parameters of a predefined curve function, facilitating adaptive image enhancement. Methods such as RUAS [44], SCI [36], and ZeroIG [48] employ Retinex-theory-based decomposition to enhance illumination and contrast, whereas QuadPrior [56], trained on the COCO [33] dataset, relies on the Kubelka-Munk theory. Additionally, Lit-the-Darkness [47] and Semantic-GuidedLLIE [67] incorporate an objective function designed to refine color fidelity, texture details, and semantic integrity, and GEFU [55] introduces an unsupervised diffusion-based framework with semantically consistent fine-tuning. However, because unsupervised methods learn *domain-level* correspondences between low- and normal-light distributions, they are inherently *weakly constrained*, often restoring brightness while producing unnatural color tones or inconsistent local illumination. To overcome these limitations, recent zero-shot approaches extend the unsupervised paradigm by performing *instance-specific* optimization guided by self-consistency and natural-light priors.

## C.2. Further Comparisons

We provide additional quantitative and qualitative comparisons to complement the results presented in the main paper. These experiments evaluate our zero-shot approach alongside recent supervised and unsupervised methods on the MIT-Adobe FiveK (MIT5K) [7] and SID [9] benchmarks. Following the evaluation protocol of Retinexformer [8], we use 500 low/normal-light image pairs from MIT-Adobe FiveK and 598 short/long-exposure pairs from the SID dataset.

**Qualitative Results.** Figures D and E illustrate (1) how


Figure D. **Additional qualitative comparisons on the SID dataset.** Please zoom in without night-light mode to accurately compare colors and observe noise reduction in each method. Green: supervised methods, Red: unsupervised methods, Blue: zero-shot methods. Each output is labeled using the format *method name–training dataset* to indicate the model and its corresponding training data.

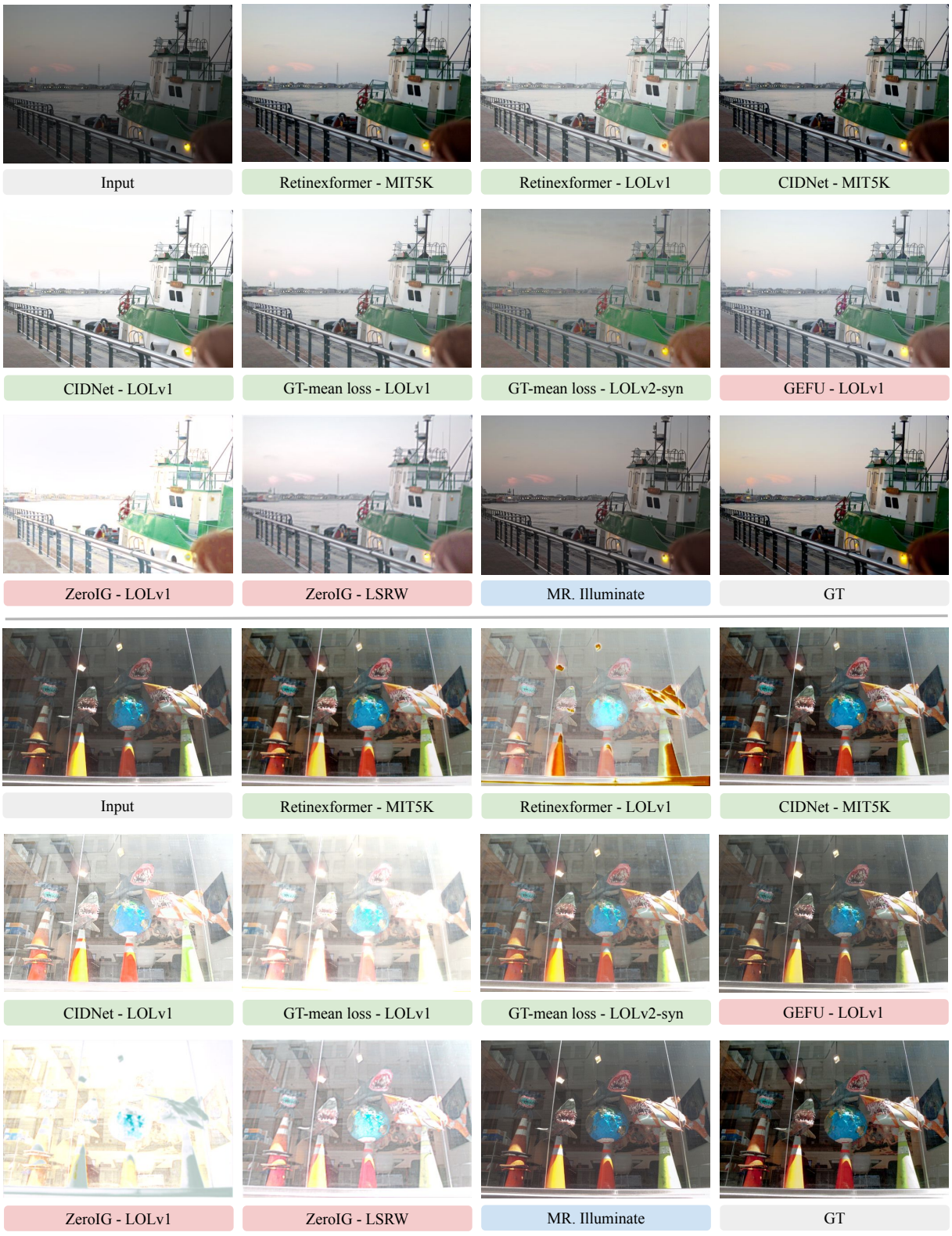


Figure E. **Additional qualitative comparisons on the MIT-Adobe FiveK dataset.** Please zoom in without night-light mode to accurately compare colors and observe noise reduction in each method. Green: supervised methods, Red: unsupervised methods, Blue: zero-shot methods. Each output is labeled using the format *method name–training dataset* to indicate the model and its corresponding training data.

| | Method | Train Data | MIT-Adobe FiveK | | | SID | | |
|---|---|---|---|---|---|---|---|---|
| | | | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| S | Retinexformer [8] | LOLv1 | 13.8688 | 0.6922 | 0.2223 | 14.1451 | 0.2287 | 0.7970 |
| | Retinexformer [8] | MIT5K | 24.4639 | 0.8870 | 0.1109 | 13.2346 | 0.2162 | **0.6789** |
| | CIDNet [61] | LOLv1 | 11.0917 | 0.6895 | 0.2184 | 13.6785 | 0.1754 | 0.9097 |
| | CIDNet [61] | MIT5K | **24.9025** | **0.9058** | **0.0618** | 13.2167 | 0.1812 | 0.8652 |
| | GT-mean loss [31] | LOLv1 | 13.0742 | 0.6940 | 0.2321 | **14.3544** | **0.2507** | 0.7732 |
| | GT-mean loss [31] | LOLv2-syn | **16.2477** | **0.7599** | **0.1668** | 12.3209 | 0.1520 | 0.8497 |
| U | ZeroIG [48] | LOLv1 | 6.2999 | 0.4524 | 0.4912 | 13.0053 | 0.1968 | 0.8288 |
| | ZeroIG [48] | LSRW | 9.8991 | 0.5924 | 0.3233 | 12.0176 | 0.0863 | 0.9566 |
| | GEFU [55] | LOLv1+ | **15.7206** | **0.7598** | **0.1699** | **14.0896** | **0.2974** | **0.7650** |
| Z | MR. Illuminate | n/a | **19.0120** | **0.8011** | **0.1647** | **15.7761** | **0.4014** | **0.5953** |

Table A. **Quantitative comparisons on MIT-Adobe FiveK (MIT5K) [7] and SID [9] datasets across supervised (S), unsupervised (U), and our zero-shot (Z) methods.** The best overall scores are highlighted in **red bold**. For supervised methods evaluated on datasets different from their training sets, we additionally highlight the best cross-dataset performance in **green bold** to emphasize generalization.

| Variant | QP VAE | SA inj. | Res inj. | Input avg. | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|---|---|---|
| Ours w/ SD Decoder | – | ✓ | – | 30 | 19.927 | 0.600 | 0.248 |
| Ours w/o SA | ✓ | – | – | 30 | 13.173 | 0.437 | 0.506 |
| Ours w/ Res | ✓ | – | ✓ | 30 | 19.363 | 0.748 | 0.240 |
| Ours w/ $SA_{avg=input}$ | ✓ | ✓ | – | input | 21.238 | **0.822** | 0.187 |
| Ours w/ $SA_{avg=60}$ | ✓ | ✓ | – | 60 | 20.798 | 0.785 | 0.194 |
| Ours (final) | ✓ | ✓ | – | 30 | **21.739** | 0.815 | **0.177** |

Table B. **Quantitative ablation study.** Results are reported on the LOL dataset. Columns explicitly indicate which components are active in each variant: *QP VAE* uses the QuadPrior VAE decoder (otherwise the standard Stable Diffusion decoder is used); *SA inj.* denotes self-attention injection in the U-Net up-blocks; *Res inj.* denotes residual feature injection in the up-blocks instead of self-attention; *Input avg.* denotes the target mean intensity of the input image, either using the original image or the same image rescaled to an average intensity of 30 or 60.

models with identical architectures can produce noticeably different outputs depending on the dataset on which they were trained, and (2) how supervised methods often exhibit limitations when applied outside their training domain. These qualitative observations reinforce the robustness and adaptability of our proposed zero-shot method.

**Quantitative Results.** Table A reinforces the two observations highlighted above. In addition, our zero-shot approach achieves stable and competitive performance across both benchmarks without relying on paired supervision.

## D. Further Ablation Study

In this section, we conduct an ablation study to analyze the contribution of each component and justify the effectiveness of our system design.

**Quantitative Ablation Study.** A quantitative ablation study on the LOL dataset (Table B) analyzes the contribution of each component in our framework, which operates *without* text prompts. First, we examine a preprocessing step that rescales very dark inputs so that the self-attention features are extracted from the input with sufficient contrast and signal information. Second, we compare which diffusion internal feature is most effective for guiding reconstruction by evaluating self-attention injection compared with residual feature injection in the U-Net up-blocks quantitatively. For qualitative analysis, please refer to the ablation study presented in the main paper. Third, we assess the influence of the VAE decoder by comparing the standard Stable Diffusion decoder with the QuadPrior decoder.

**Preprocessing.** As shown in Figure G, we apply a linear gain adjustment in pixel space when its overall intensity is very low before an image is passed to the VAE encoder. If the mean pixel value is below 30 (in the [0, 255] range), the image is linearly rescaled so that its average reaches 30; otherwise, it is left unchanged. This adjustment is necessary because very dark inputs contain *limited contrast and signal information*, which leads to self-attention features with weak or uninformative signals. Increasing the target to higher values, such as 60, however, (1) *amplifies noise* and (2) causes *irreversible information loss*, particularly through over-brightening in white or near-white regions. Empirically, a target value of 30 offers an effective balance—preserving input fidelity while providing enough signal during the Refine step of our framework.

**Adopted Configuration.** Guided by the results in Table B and the qualitative ablation in the main paper, we adopt the QuadPrior decoder with self-attention injection computed from inputs rescaled to an average intensity of 30, as this setup offers an effective balance between restoration (PSNR/SSIM) and perceptual quality (LPIPS).

**Sensitivity of sampling styles & timesteps.** Our method requires an approximately *reversible* scheduler and must

| Scheduler | Time steps | Time | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|---|
| DPM-Solver++ (order: 2) | 5 | 0.03 | 21.0578 | 0.7896 | 0.2526 |
| | 10 | 0.05 | 21.6014 | **0.8105** | 0.1968 |
| | 15 | 0.07 | 21.2309 | 0.7879 | 0.2063 |
| | 25 | 0.11 | **21.6420** | 0.8009 | **0.1911** |
| DDIM | 5 | 0.03 | 20.4389 | 0.7971 | 0.2327 |
| | 15 | 0.07 | 21.5841 | 0.8133 | 0.1805 |
| | 25 | 0.12 | **21.7393** | **0.8152** | **0.1771** |
| | 150 | 0.64 | 21.6241 | 0.8114 | 0.1810 |

Table C. **Ablation study on schedulers & timesteps.** End-to-end runtime and metrics are measured on the LOL test set (A10 GPU).

satisfy: (1) support both sampling and inversion, unlike most schedulers which only support sampling; (2) ensure trajectory consistency, i.e., $z_t^{\text{inv}} \approx z_t^{\text{samp}}$ for all $t$, because we record self-attention features during inversion and re-inject them *at the same timesteps* during sampling to re-anchor the denoising trajectory, rather than only requiring accurate final reconstruction $z_0^{\text{inv}} \approx z_0^{\text{samp}}$; and (3) perform inversion using model-predicted noise (e.g., unlike DDPM), since attention features are produced by the U-Net. Under these 3 constraints, we evaluate valid schedulers (DDIM and DPM-Solver++) and analyze sensitivity over scheduler choices and timesteps. As shown in Tab. C & Fig. F, performance remains stable across different schedulers and timestep configurations, with only minor metric variations under the required reversibility and trajectory-consistency constraints. Beyond computational efficiency, the number of DDIM steps also affects reconstruction characteristics. As illustrated in Fig. F, using too few steps (e.g., 5) results in insufficient refinement, leading to detail loss, while an excessive number of steps (e.g., 150) amplifies noise due to repeated injection of dark and noisy input information during denoising. In contrast, 25 steps provide the best qualitative balance, preserving structural details while avoiding noise amplification. Combined with its stable quantitative performance and lower runtime, this observation justifies our choice of deterministic DDIM with 25 steps as the default configuration.

**Determinism.** As shown in Table D, using different random seeds for $z_T^s$ leads to marginal variance, indicating that our method is robust to the sampled style. This design ensures stable and reproducible enhancement results across runs, a desirable property for restoration tasks where consistency is essential.

**Sensitivity to Diffusion Inversion Quality.** Our approach relies on DDIM inversion to obtain latent representations and attention maps. To examine robustness under challenging conditions, we include extremely dark and noisy examples in Figures O and J, where substantial scene information is degraded in the input (see the *Scaled* column). The results demonstrate that the proposed method is able to recover coherent structures and produce visually faithful enhancements even in such demanding scenarios.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| Main paper | 17.6634 | 0.5185 | 0.2829 |
| 10 runs | $17.404 \pm 0.270$ | $0.519 \pm 0.006$ | $0.292 \pm 0.009$ |

Table D. **Determinism.** $\mu \pm \sigma$ over 10 runs on LSRW using different *random* seeds for the Gaussian style latent $z_T^s \sim \mathcal{N}(0, I)$.

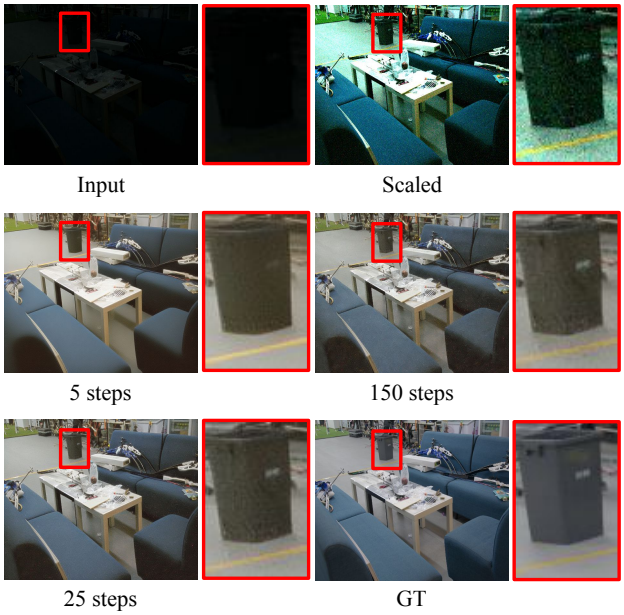


Figure F. **Effect of DDIM steps.** With self-attention guidance, 5 steps lead to loss of details (white marks); 150 steps result in increased noise levels *due to repeated dark noisy input information*; 25 steps achieve the best trade-off between quality and runtime.

## E. Limitations and Hallucination Comparisons

Figures I and J present failure cases of our method alongside the hallucination behavior exhibited by our zero-shot diffusion baseline. As shown in Figure I, our approach tends to maintain chromatic brightness levels near the mid-intensity range. While this may occasionally introduce deviations from the ground truth, it is generally advantageous—particularly for low-light images where color information is severely degraded or when the input contains large brightness variations. In the same figure, we observe that recent SOTA supervised methods, CIDNet and GT-mean loss trained on LOLv1 [59], produce overexposed outputs due to their limited exposure to semi-dark inputs. In contrast, our method exhibits strong robustness across diverse lighting conditions.

Furthermore, our method preserves structural and semantic integrity, whereas our diffusion-based baseline GDP [13] exhibits hallucinations on very dark or noisy inputs (Figure J). Our framework avoids such behavior by

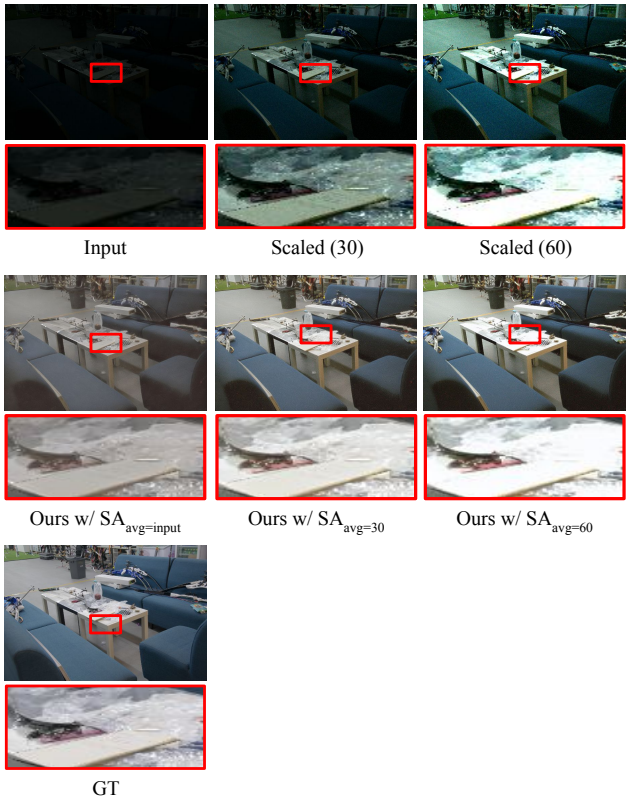


Figure G. **Preprocessing Step.** "Scaled (60/30)" applies linear rescaling to the target mean. "Ours w/ SA$_{avg=*}$" uses attention statistics from the corresponding scaled (or original) input. GT denotes the ground truth.

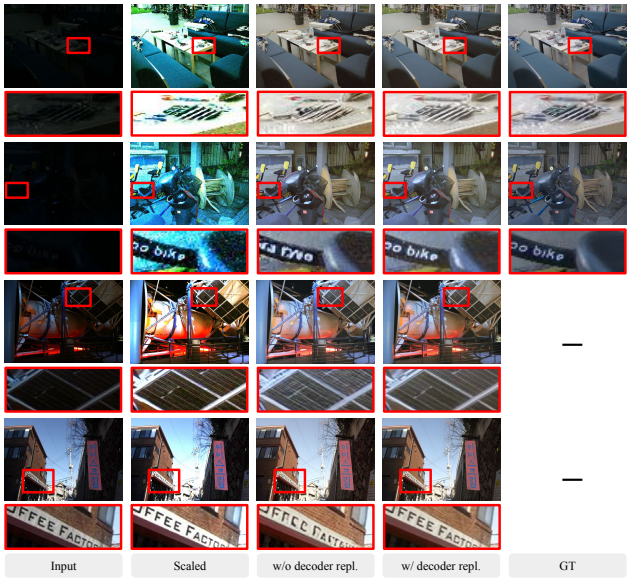


Figure H. **Decoder replacement.** Comparison of outputs using the original Stable Diffusion decoder [53] and the fine-tuned decoder from QuadPrior [56]. The first two rows show examples from the LOL dataset, and the last two from the Unpaired dataset. "Scaled" denotes images normalized to an average intensity of 120. As the encoder remains fixed, all latent-space operations of our method are unchanged; thus, results are nearly identical between "w/o decoder repl." and "w/ decoder repl." However, using the fine-tuned decoder better preserves fine-scale details such as edges and text, improving perceptual fidelity without altering global appearance.

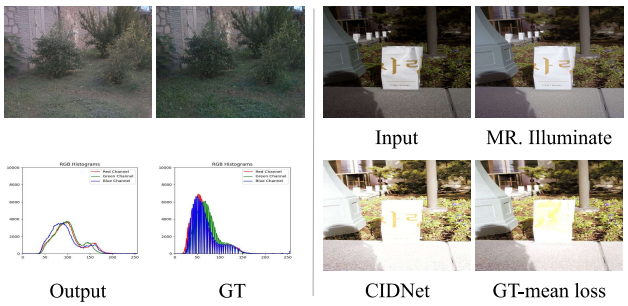


Figure I. **Failure cases.** Our method maintains mid-range chromatic brightness, which can cause slight deviations from ground truth. However, this property becomes beneficial for inputs with varying brightness, as illustrated on the right.

constraining the diffusion trajectory through two complementary mechanisms. First, the *Modulate* step injects structural information at initialization via AdaIN, aligning the noise distribution to the input and preventing early-stage generative drift. Second, the *Refine* step enforces inversion-time self-attention through attention injection, ensuring that sampling remains consistent with the original image. In addition, replacing the default Stable Diffusion decoder with the fine-tuned QuadPrior decoder (Figure H) further reduces semantic distortion by preserving fine-grained structural details during latent-to-image decoding. This is because the fine-tuned decoder leverages intermediate features from the encoder to reinject local structural details that may be lost in the latent representation during encoding. Together, these components maintain semantic fidelity and suppress hallucination while enabling consistent low-light enhancement.

## F. AWB Quantitative and Qualitative Analysis

**Quantitative Results.** As shown in Table E, without any modifications to our LLIE framework, our method achieves competitive performance with supervised methods [1, 2, 6] while surpassing the general image restoration methods TAO [18] and DDNM [57] with a degradation function aligning each channel mean. Additionally, our approach achieves competitive results compared to supervised methods trained specifically for this task.

**Qualitative Results.** As illustrated in Figure L, although our method is not explicitly trained for the AWB task, its results closely align with the ground truth and are on par


Figure J. **Hallucination.** While diffusion prior is effective for image restoration, improper application can lead to unintended hallucinations, where the model generates nonexistent structures or alters scene semantics. For example, GDP [13], a robust and versatile image restoration method, often hallucinates in the presence of substantial noise and darkness in input images. As shown in row 1, a blue-colored cabinet is inaccurately reconstructed as a sky, a pink cabinet as a building, and the entire scene resembles a battle.

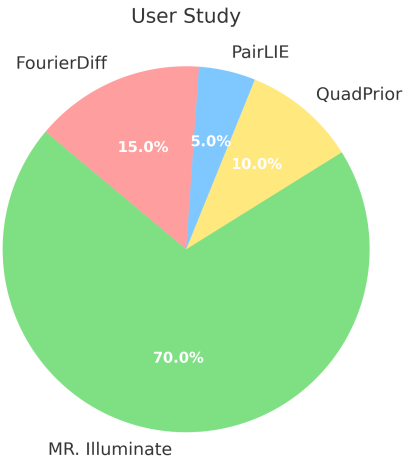


Figure K. **User Study.** In a four-alternative forced choice user study, 30 participants compared outputs from 4 methods across 20 random images from LOL, LSRW, and Unpaired. All questions were mandatory. While low-light image enhancement is often judged quantitatively, we also evaluate perceptual quality through a user study, providing a complementary perspective that standard metrics may not capture.

with supervised methods [1, 2, 6]. In contrast, TAO [18] introduces residual noise, DDNM [57] exhibits color shifts, and Quasi-CC leads to over-exposure.

| | Method | Train Data | $\Delta E \downarrow$ | MAE $\downarrow$ | MSE $\downarrow$ |
|---|---|---|---|---|---|
| S | mixedillWB [2] | RenderedWB [1] | **7.24** | 4.16° | **115.25** |
| | WB_sRGB [1] | NUS [10], Gehler [16] | 7.81 | **4.04°** | 284.66 |
| | Quasi-CC [6] | Flickr100k [40] | 24.26 | 6.24° | 3185.06 |
| | Quasi-CC [6] | ilsvrc12 [46] | 24.12 | 6.13° | 3179.82 |
| | Quasi-CC [6] | Places365 [68] | 24.26 | 6.21° | 3194.27 |
| Z | DDNM [57] | n/a | 45.73 | 18.97° | 6146.39 |
| | TAO [18] | n/a | 16.84 | 7.48° | 1133.40 |
| | **MR. Illuminate** | n/a | **13.41** | **5.75°** | **593.46** |

Table E. Quantitative comparison on the CUBE+ dataset [1, 5] with existing methods whose code and pretrained models are both *publicly available* and *runnable*. We use the evaluation code from WB_sRGB [1]. Metrics include average $\Delta E$ (CIE76), MAE (deg), and MSE.

## G. Additional LLIE Qualitative Comparisons

In the first six figures, we present same-scene qualitative comparisons on the LOL (paired) and the unpaired DICM, LIME, MEF, NPE, and VV datasets, which contain multiple captures of identical scenes under varying illumination. Our method maintains color constancy across these variations, while other approaches often show inconsistencies. Baselines use their official checkpoints (ZeroIG trained on LSRW; Retinexformer and CIDNet on MIT-Adobe FiveK (5K)), and results are labeled as *method–training dataset*.

The next six figures provide additional qualitative examples from LOL (paired), LSRW (paired), and the unpaired datasets, further demonstrating the robustness of our method across diverse scenes, using official baseline checkpoints.

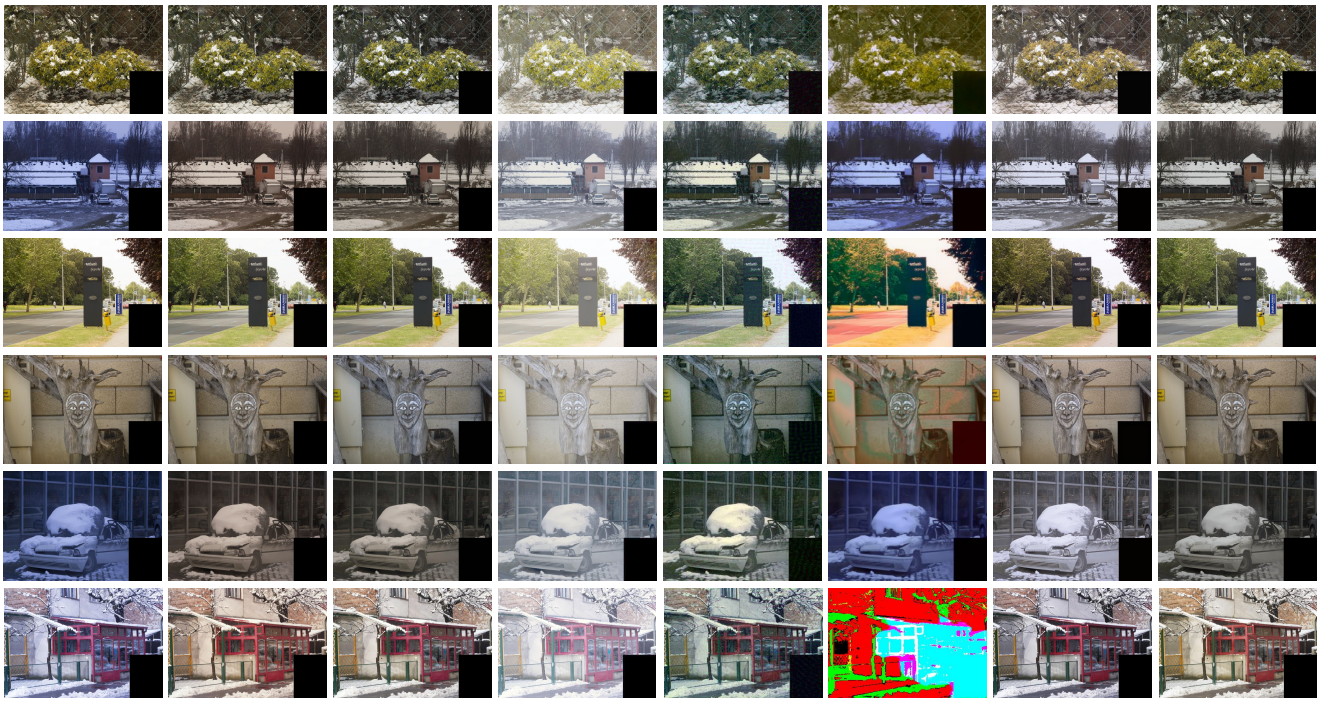


Figure L. Qualitative evaluation of our method against existing supervised and general image restoration methods on auto white balance task on CUBE+ dataset [1, 5]. In this dataset, the calibration object is masked out using a black box. Please *zoom in without night-light mode* to accurately compare colors and observe noise reduction in each method. The column labeled as Quasi-CC represents the Quasi-CC method trained on the Places365 [68].

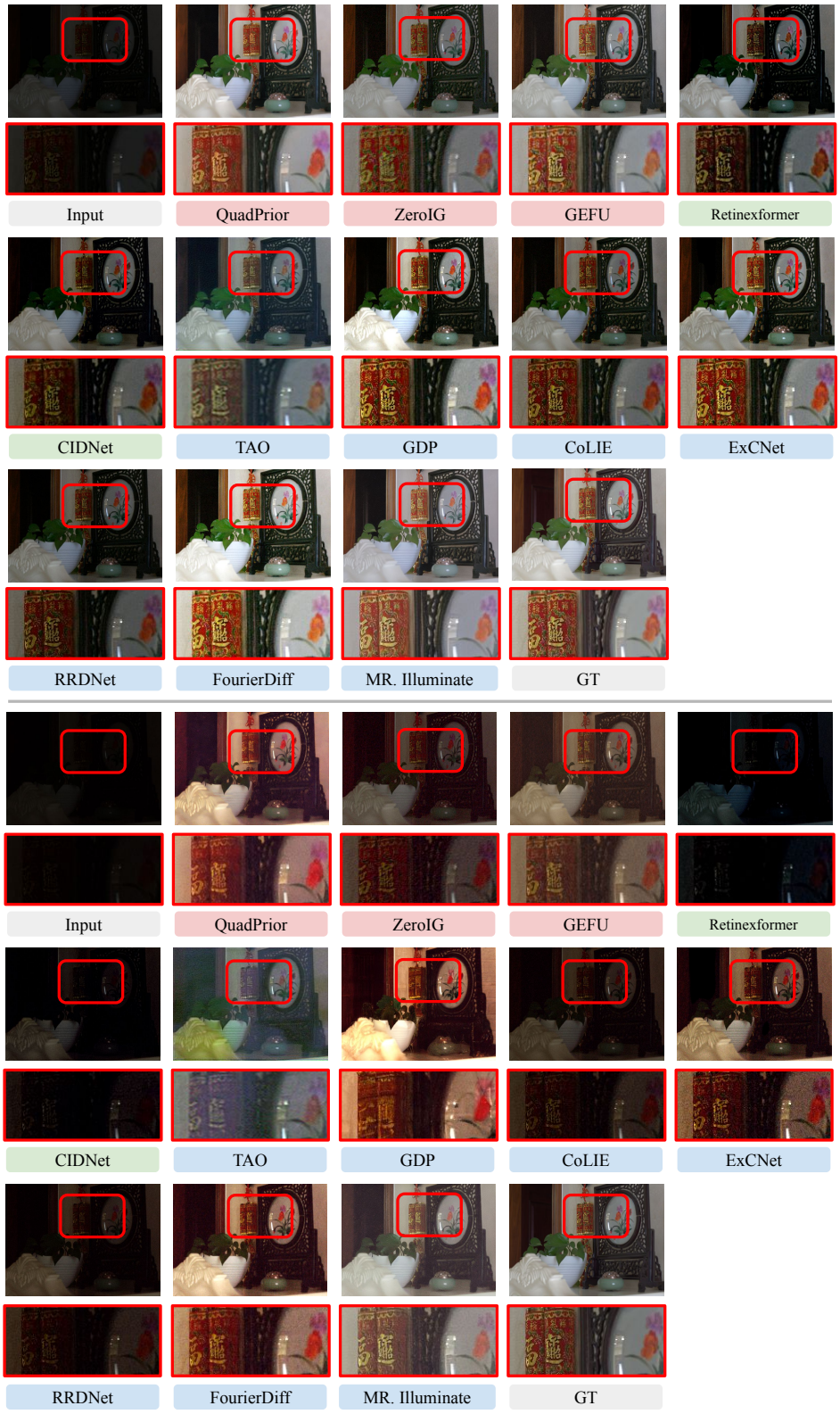


Figure M. **Same-scene qualitative comparisons on the LOL dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.


Figure N. **Same-scene qualitative comparisons on the LOL dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

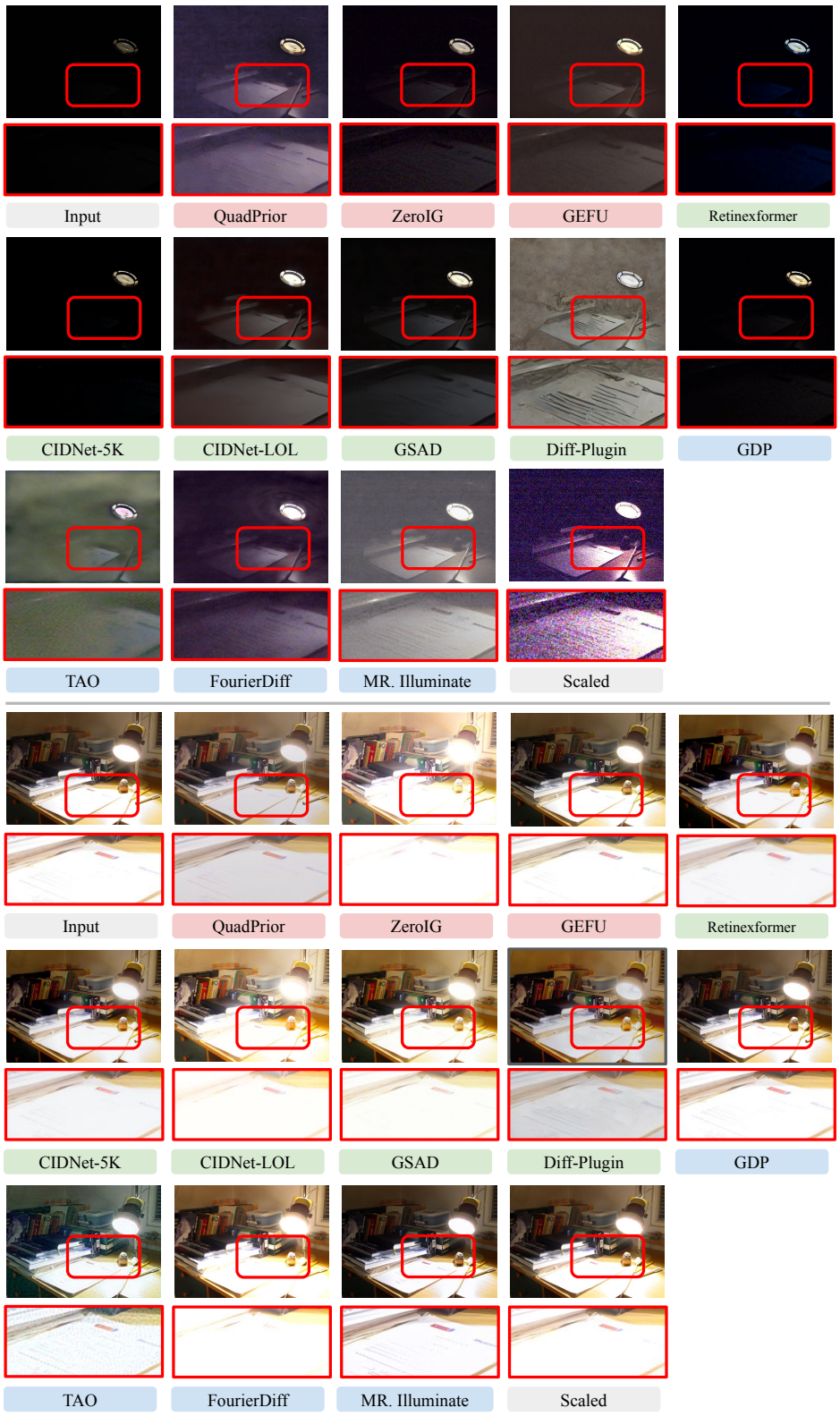


Figure O. **Same-scene qualitative comparisons on the Unpaired dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

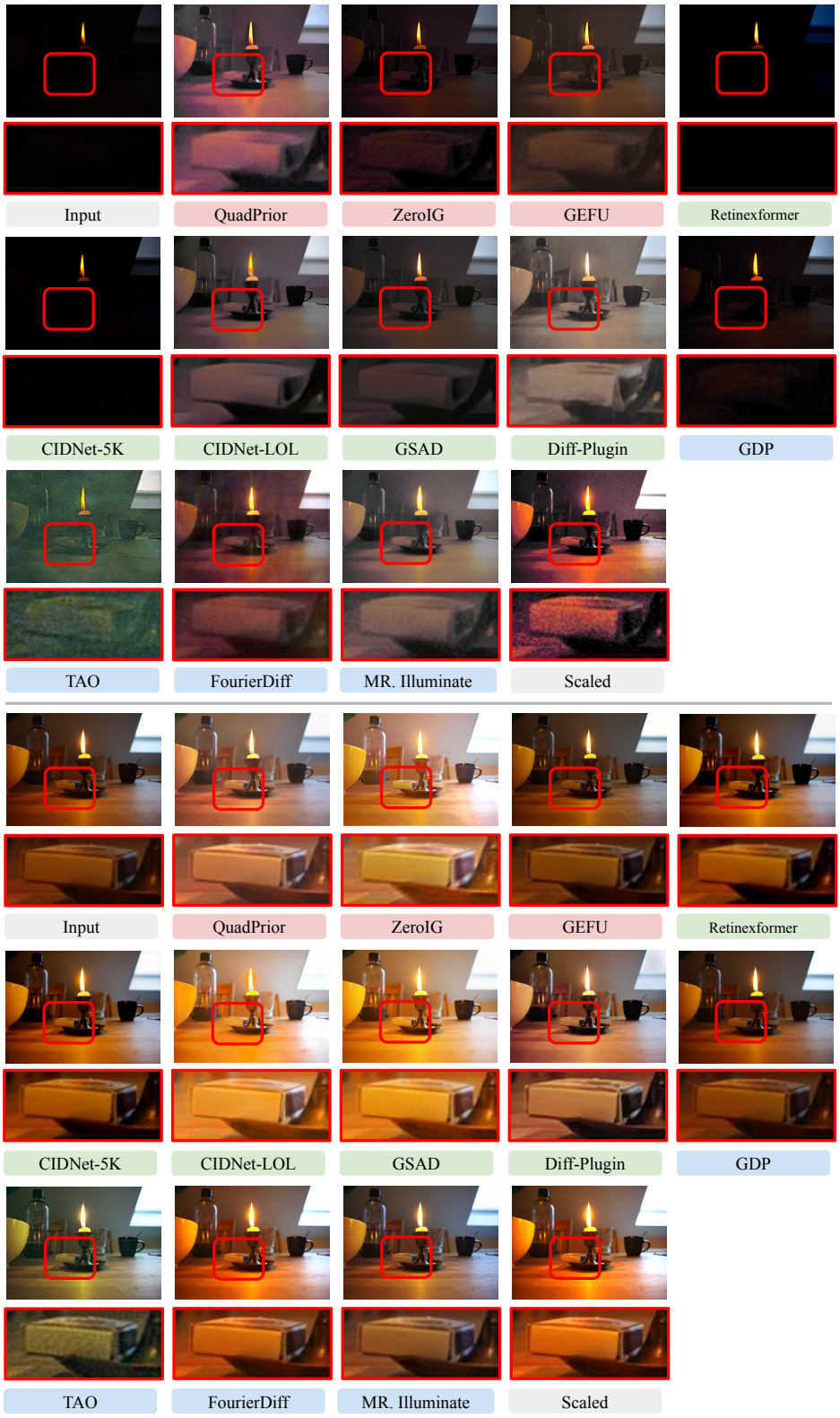


Figure P. **Same-scene qualitative comparisons on the Unpaired dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.


Figure Q. **Same-scene qualitative comparisons on the Unpaired dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.


Figure R. **Same-scene qualitative comparisons on the LOL dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

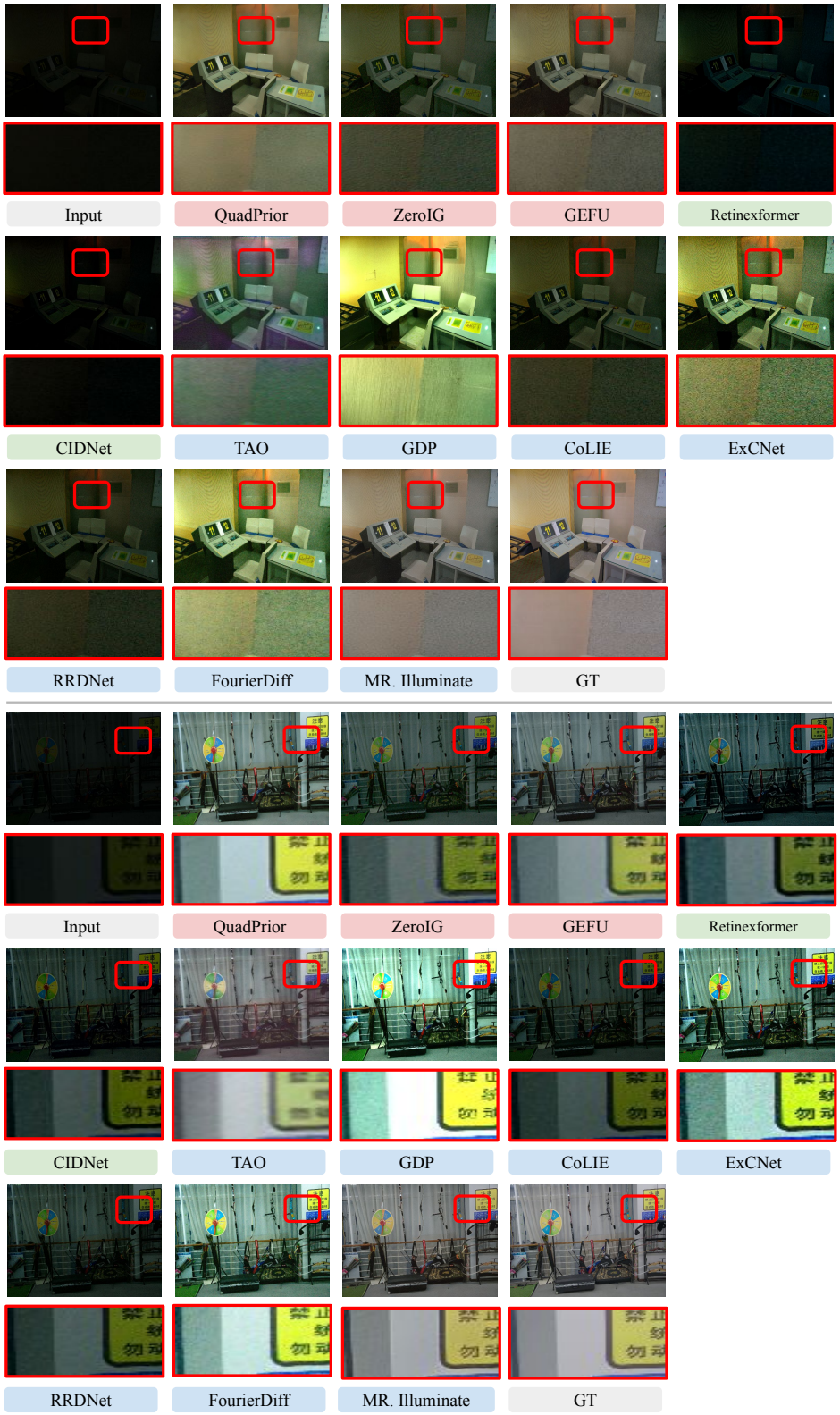


Figure S. **Additional qualitative comparisons on the LOL dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

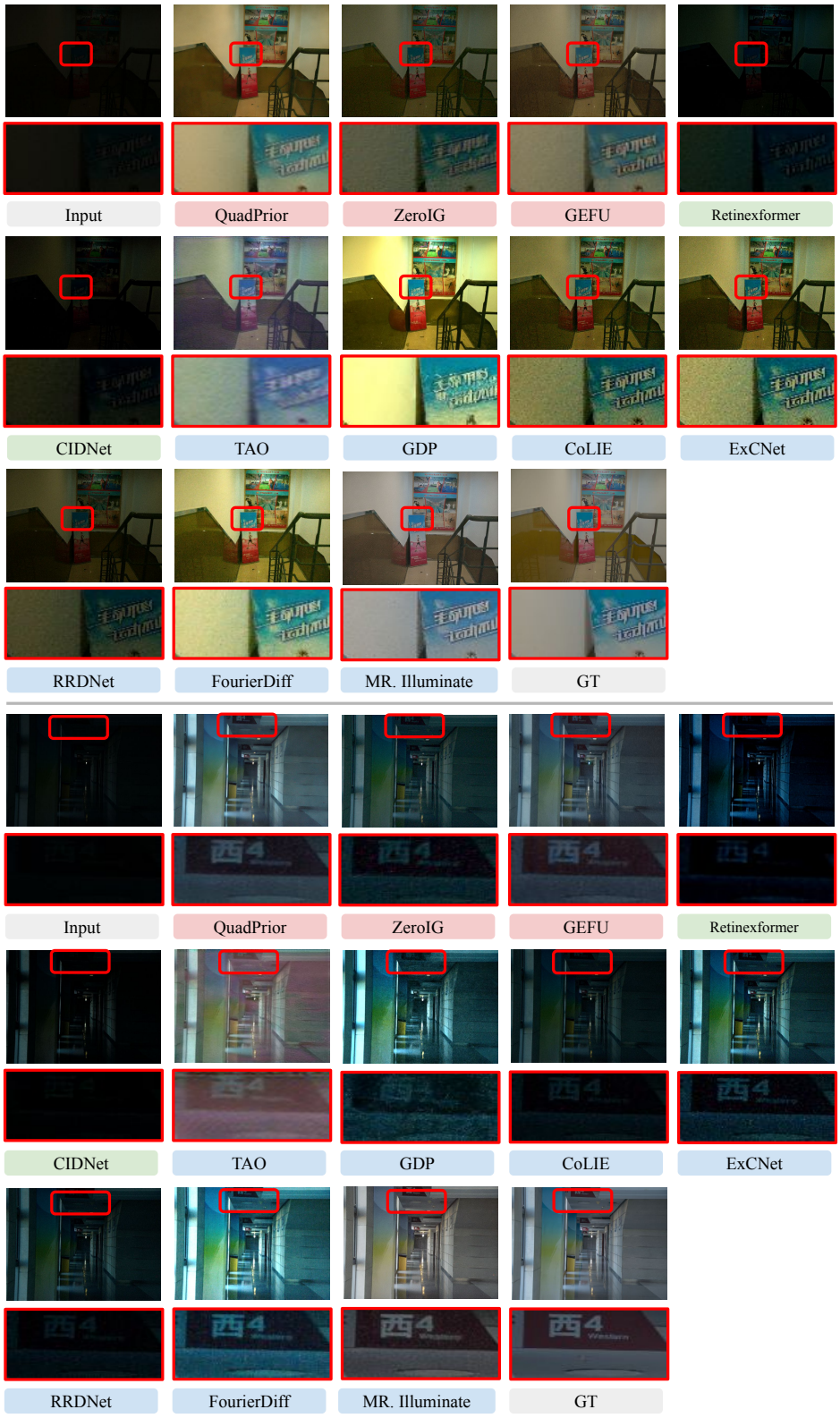


Figure T. **Additional qualitative comparisons on the LOL dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

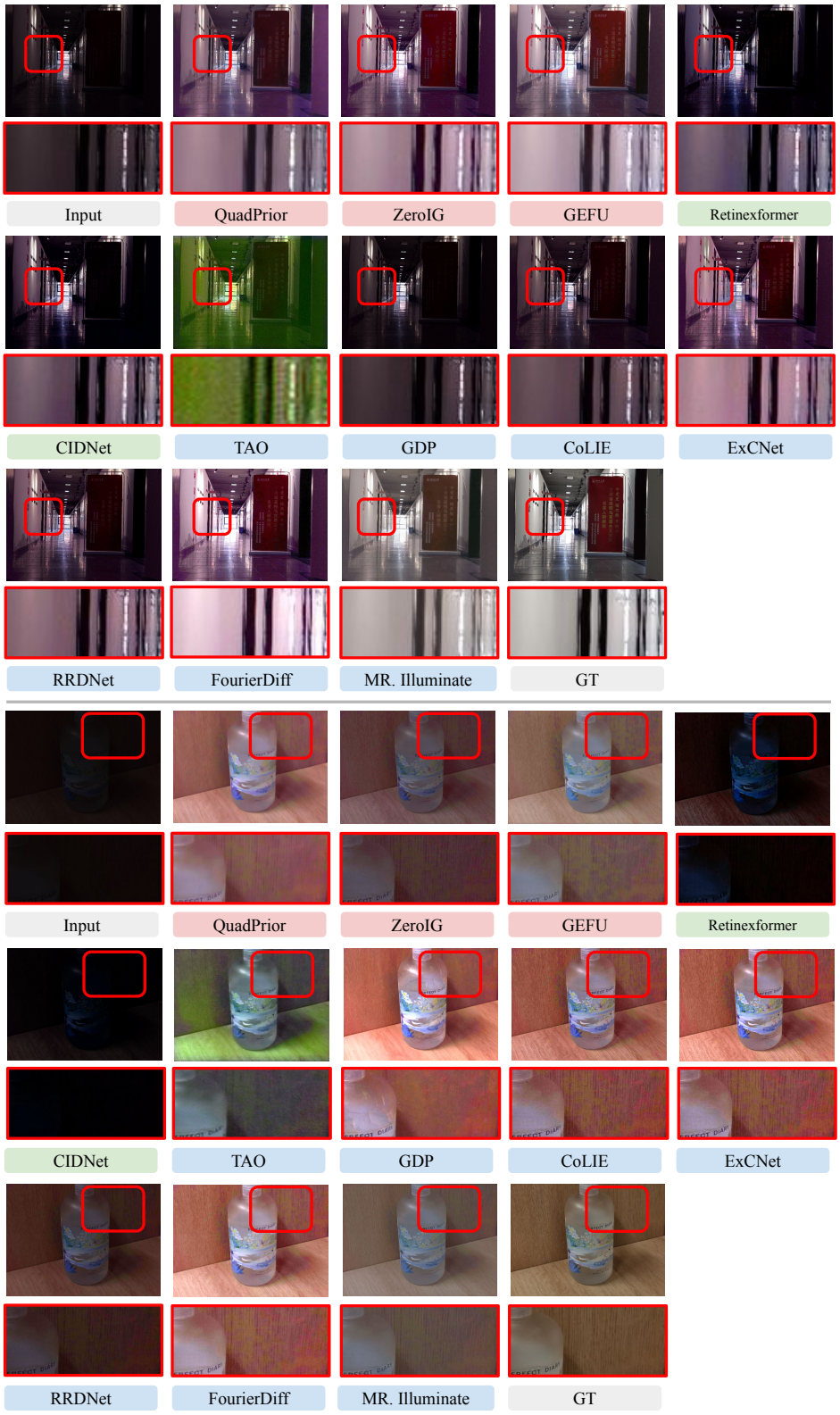


Figure U. **Additional qualitative comparisons on the LSRW dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

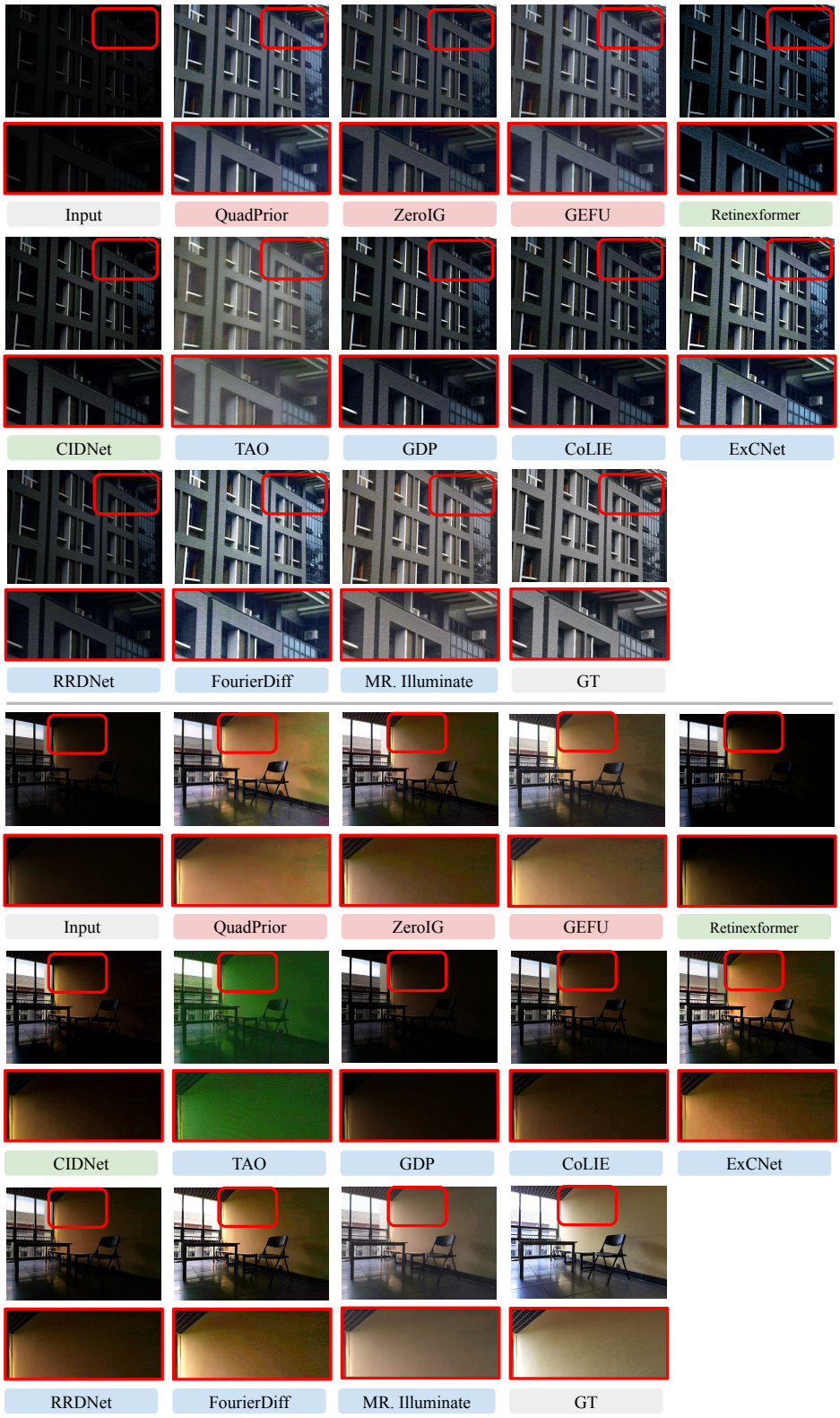


Figure V. **Additional qualitative comparisons on the LSRW dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

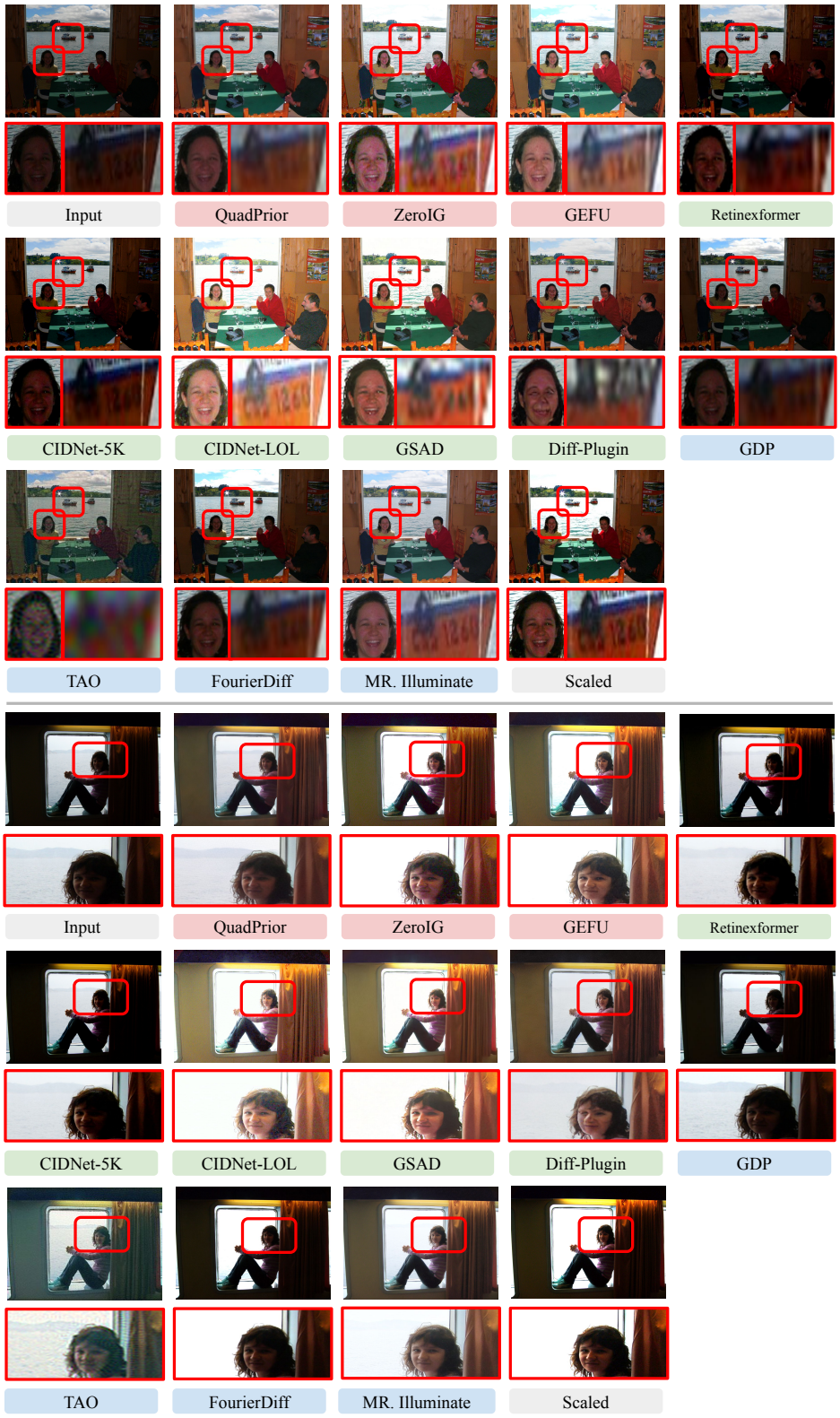


Figure W. **Additional qualitative comparisons on the Unpaired dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

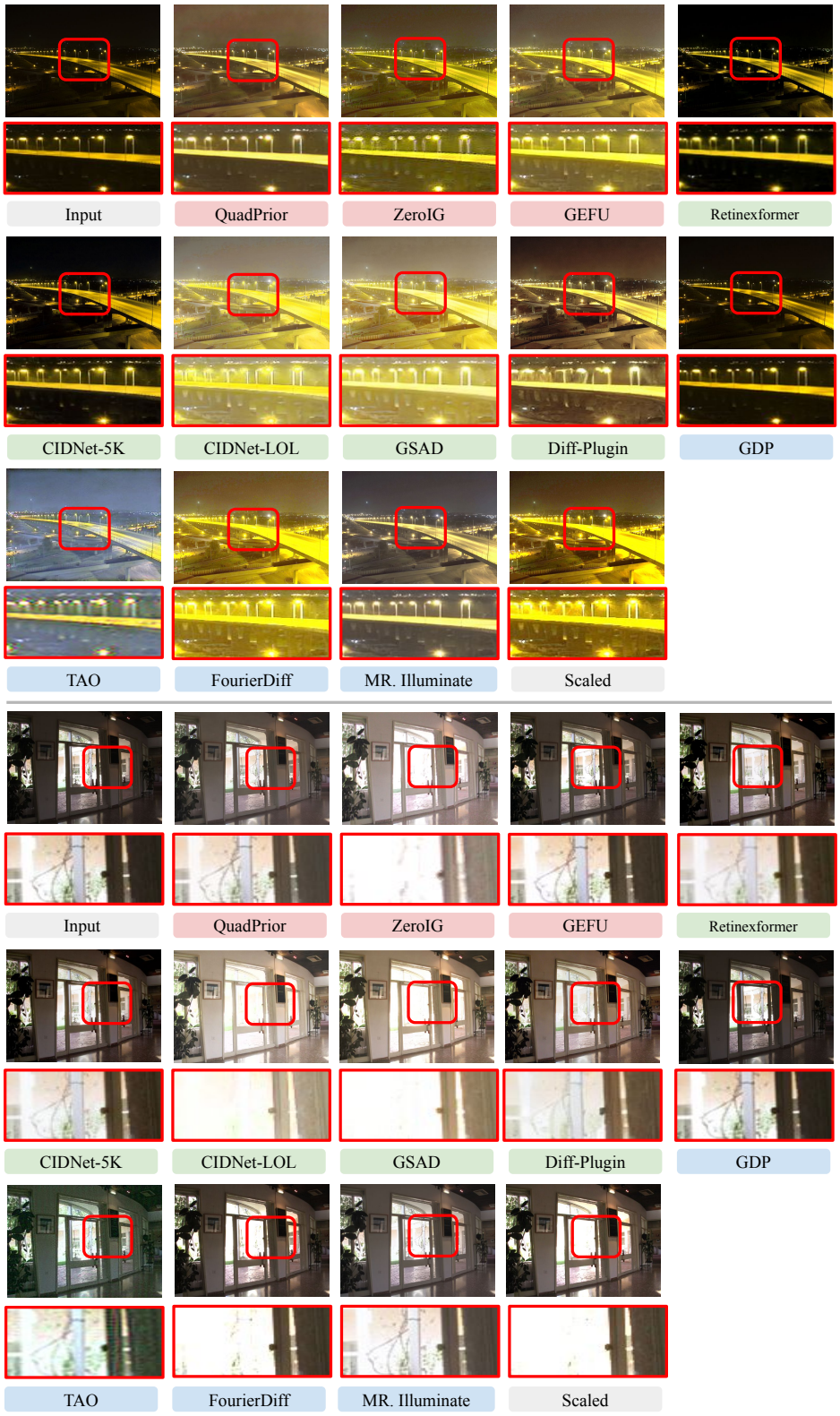


Figure X. **Additional qualitative comparisons on the Unpaired dataset.** Red: unsupervised methods, Green: supervised methods, Blue: zero-shot methods.

## References


[1] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *CVPR*, 2019. 9, 10, 11

[2] Mahmoud Afifi, Marcus A. Brubaker, and Michael S. Brown. Auto white-balance correction for mixed-illuminant scenes. In *WACV*, 2022. 9, 10

[3] Sara Aghajanzadeh and David Forsyth. Long scale error control in low light image and video enhancement using equivariance. *arXiv:2206.01334*, 2022. 4

[4] Sara Aghajanzadeh and David Forsyth. Towards robust low light image enhancement. *arXiv:2205.08615*, 2022. 4

[5] Nikola Banić and Sven Lončarić. Unsupervised learning for color constancy. In *VISIGRAPP*, 2017. 10, 11

[6] Simone Bianco and Claudio Cusano. Quasi-unsupervised color constancy. In *CVPR*, 2019. 9, 10

[7] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Fredo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 4, 7

[8] 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 *ICCV*, 2023. 4, 7

[9] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. *ICCV*, 2019. 4, 7

[10] Dongliang Cheng, Dilip K Prasad, and Michael S Brown. Illuminant estimation for color constancy: Why spatial-domain methods work and the role of the color distribution. *Journal of the Optical Society of America*, 2014. 10

[11] Heng-Da Cheng and Xiangjun Shi. A simple and effective histogram equalization approach to image enhancement. In *Digital Signal Processing*, 2004. 4

[12] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 4

[13] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, 2023. 8, 10

[14] Daniel Feijoo, Juan C. Benito, Alvaro Garcia, and Marcos V. Conde. Darkir: Robust low-light image restoration. In *CVPR*, 2025. 4

[15] 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 *CVPR*, 2023. 4

[16] Peter Vincent Gehler, Carsten Rother, Andrew Blake, Tom Minka, and Toby Sharp. Bayesian color constancy revisited. In *CVPR*, 2008. 10

[17] Theo Gevers, Arjan Gijsenij, Joost Van de Weijer, and Jan-Mark Geusebroek. *Color in Computer Vision: Fundamentals and Applications*. John Wiley & Sons, 2012. 4

[18] Yuanbiao Gou, Haiyu Zhao, Boyun Li, Xinyan Xiao, and Xi Peng. Test-time degradation adaptation for open-set image restoration. In *ICML*, 2024. 9, 10

[19] Yuxuan Gu, Haoxuan Wang, Pengyang Ling, Zhixiang Wei, Huaian Chen, Yi Jin, and Enhong Chen. Improving visual and downstream performance of low-light enhancer with vision foundation models collaboration. In *CVPR*, 2025. 4

[20] 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 *CVPR*, 2020. 4

[21] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 4

[22] Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. In *NeurIPS*, 2023. 4

[23] Hai Jiang, Ao Luo, Songchen Han, Haoqiang Fan, and Shuaicheng Liu. Low-light image enhancement with wavelet-based diffusion models. *ACM TOG*, 2023.

[24] Hai Jiang, Ao Luo, Xiaohong Liu, Songchen Han, and Shuaicheng Liu. Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In *ECCV*, 2024. 4

[25] 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. In *IEEE TIP*, 2021. 4

[26] Edwin H. Land. The retinex theory of color vision. *Scientific American*, 1977. 4

[27] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. In *IEEE TPAMI*, 2021. 4

[28] Xueyang Li, Yuhui Zhang, Yutong Ma, Boxin Shi, and Wen Gao. Cwnet: Low-light image enhancement with causal wavelet network. *ICCV*, 2025. 4

[29] Ze Li, Feng Zhang, Xiatian Zhu, Meng Zhang, Yanghong Zhou, and P. Y. Mok. Robust low-light scene restoration illumination transition. In *ICCV*, 2025. 4

[30] Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *ICCV*, 2023. 4

[31] Jingxi Liao, Shijie Hao, Richang Hong, and Meng Wang. Gt-mean loss: A simple yet effective solution for brightness mismatch in low-light image enhancement. In *ICCV*, 2025. 4, 7

[32] Minwen Liao, Haobo Dong, Xinyi Wang, Kurban Ubul, Ziyang Yan, and Yihua Shao. Gm-moe: Low-light enhancement with gated-mechanism mixture-of-experts. *ICCV*, 2025. 4

[33] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 4

[34] Yuhao Liu, Zhanghan Ke, Fang Liu, Nanxuan Zhao, and Rynson W.H. Lau. Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In *CVPR*, 2024. 4

[35] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: low-light image/video enhancement using cnns. In *BMVC*, 2018. 4

[36] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022. 4

[37] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[38] Kristina Monakhova, Stephan R. Richter, Laura Waller, and Vladlen Koltun. Dancing under the stars: Video denoising in starlight. In *CVPR*, 2022. 4

[39] Cindy M Nguyen, Eric R Chan, Alexander W Bergman, and Gordon Wetzstein. Diffusion in the dark: A diffusion model for low-light text recognition. In *WACV*, 2024. 4

[40] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 10

[41] Stephen M. Pizer, Elton Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart M. ter Haar Romeny, and John B. Zimmerman. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 1987. 4

[42] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib. An adaptive gamma correction for image enhancement. *EURASIP Journal on Image and Video Processing*, 2016. 4

[43] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Emil Pavlović, Geir Kjetil Sandve, Victor Greiff, David P. Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *ICLR*, 2021. 2, 3

[44] Liu Risheng, Ma Long, Zhang Jiaao, Fan Xin, and Luo Zhongxuan. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *CVPR*, 2021. 4

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 10

[47] Mariam Saeed and Marwan Torki. Lit the darkness: Three-stage zero-shot learning for low-light enhancement with multi-neighbor enhancement factors. In *ICASSP*, 2023. 4

[48] Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. Zero-ig: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *CVPR*, 2024. 4, 7

[49] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 4

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[51] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 4

[52] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 4

[53] Stability AI. Stable diffusion v1.5 model card. https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5, 2022. 9

[54] Lei Sun, Yuhan Bao, Jiajun Zhai, Jingyun Liang, Yulun Zhang, Kaiwei Wang, Danda Pani Paudel, and Luc Van Gool. Low-light image enhancement using event-based illumination estimation. *ICCV*, 2025. 4

[55] Sen Wang, Shao Zeng, Tianjun Gu, Zhizhong Zhang, Ruixin Zhang, Shouhong Ding, Jingyun Zhang, Jun Wang, Xin Tan, Yuan Xie, and Lizhuang Ma. From enhancement to understanding: Build a generalized bridge for low-light vision via semantically consistent unsupervised fine-tuning. In *ICCV*, 2025. 4, 7

[56] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *CVPR*, 2024. 4, 9

[57] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *ICLR*, 2023. 9, 10

[58] 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 *ICCV*, 2023. 4

[59] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 4, 8

[60] Rui Xu, Yuzhen Niu, Yuezhou Li, Huangbiao Xu, Wenxi Liu, and Yuzhong Chen. Urwkv: Unified rwkv model with multi-state perspective for low-light image restoration. In *CVPR*, 2025. 4

[61] 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. *CVPR*, 2025. 4, 7

[62] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *ICCV*, 2023. 4

[63] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE TIP*, 2021. 4

[64] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *ICCV*, 2023. 4

[65] Alan Loddon Yuille and Anand Rangarajan. The concave-convex procedure (cccp). In *NeurIPS*, 2001. 2, 3

[66] Lin Zhang, Lijun Zhang, Xinyu Liu, Ying Shen, Shaoming Zhang, and Shengjie Zhao. Zero-shot restoration of back-lit images using deep internal learning. In *ACM MM*, 2019. 4

[67] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *WACV*, 2022. 4

[68] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. 10, 11