Temporal-Synergistic Policy Optimization for Unsupervised Low-Light Image Enhancement

Yuanfei Bao , Dong Li , Jie Huang , Xingbo Wang , Xueyang Fu*
University of Science and Technology of China, Hefei, China
{baoyuanfei, dongli6}@mail.ustc.edu.cn, xyfu@ustc.edu.cn

Abstract

Diffusion models show significant potential for low-light image enhancement. However, this task requires satisfying human perceptual preferences and content fidelity transcending simple brightness and color improvement. Existing methods rely on heuristic physical priors or incorporate perceptual metrics directly into timestep-wise training objectives. Such proxy constraints fail to provide reliable trajectory-level guidance for perceptual alignment. Furthermore, they often lead to artifacts or unnatural visual effects in this ill-posed inverse problem. To address these issues, we propose an unsupervised low-light enhancement framework based on Group Relative Policy Optimization (GRPO), which utilizes perceptual preferences to directly optimize the diffusion policy. We introduce a Sliding Window Hybrid ODE-SDE Sampling strategy that confines stochasticity to dynamic subintervals, thereby achieving efficient coarse-to-fine exploration and precise advantage attribution. To meet strict fidelity constraints, we construct a Synergistic Perception-Fidelity Reward and introduce an Independent Advantage Estimation strategy to mitigate signal collapse and gradient suppression in multi-objective optimization. Furthermore, we design a Temporal-Aware Dynamic Weighting Mechanism that adaptively adjusts perceptual weights across denoising stages to balance visual enhancement with structural preservation. Extensive experiments on multiple real-world benchmarks demonstrate that our method effectively improves the perceptual performance of existing models, yielding results that better align with human aesthetics.

1 Introduction

Real-world low-light scenarios suffer from coupled degradation factors, including low signal-to-noise ratio, uneven illumination, and significant color deviation, rendering low-light image enhancement a highly ill-posed inverse problem. Beyond merely improving global brightness, the enhancement process must simultaneously suppress noise while preserving structural consistency and color naturalness. Traditional methods predominantly rely on fixed hand-crafted priors, which struggle to adapt to the complex and varying conditions of real-world scenes [Wu *et al.*, 2022]. While supervised learning approaches have achieved remarkable progress [Wu *et al.*, 2023], their application is severely constrained by the high cost of acquiring paired labeled data. Furthermore, synthetic degradation fails to accurately model the complex imaging characteristics of real-world environments. The resulting significant domain gap severely impairs the model's generalization capability.

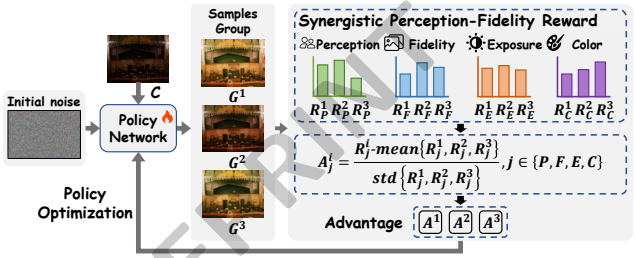

Figure 1: Schematic overview of our proposed online RL-based unsupervised low-light enhancement framework guided by the Synergistic Perception-Fidelity Reward.

In recent years, diffusion models have demonstrated significant potential in low-light image enhancement tasks. Existing methods typically rely on heuristic physical properties for conditional guidance or directly incorporate no-reference perceptual metrics into timestep-wise training objectives [Yi *et al.*, 2023]. Fundamentally, these methods constitute proxy perceptual constraints: perceptual evaluation is performed in the final state, whereas the optimization process is distributed across the entire denoising trajectory. In unsupervised settings, this misalignment not only tends to amplify metric noise, leading to training instability, but also induces the model to generate visual artifacts in an attempt to exploit these metrics. Consequently, how to stably and effectively integrate human perceptual preferences into the diffusion generation process, without relying on paired data, remains a critical problem to be addressed.

In tasks such as text-to-image generation, where human subjective evaluation is difficult to explicitly model, rein-

*Corresponding author.

forcement learning (RL) methods based on preference modeling have recently achieved significant success [Xu *et al.*, 2023]. By formulating quality assessment as a terminal preference reward, RL can directly guide the model at the generative distribution level to align with human perception, bypassing the need to explicitly model the perceptual process. However, traditional RLHF methods, such as Proximal Policy Optimization (PPO) [Schulman *et al.*, 2017], rely on an additional value network for advantage estimation, incurring significant computational and memory overheads. Group Relative Policy Optimization (GRPO) [Shao *et al.*, 2024] mitigates this by leveraging multiple samples from the same input to estimate advantages based on intra-group relative performance, thereby eliminating the dependency on an explicit value network. This characteristic renders GRPO particularly suitable for integration with diffusion models to directly optimize terminal generation quality.

However, directly applying GRPO to low-light enhancement entails two critical challenges. First, the denoising process exhibits distinct temporal characteristics: global structures are established in the early stages, while details and textures are refined in the later stages. Consequently, full-trajectory stochastic exploration proves inefficient. Furthermore, the trajectory-level advantage, derived solely from terminal feedback, is approximately uniformly distributed across all time steps. This ambiguity obscures the causal relationship between the terminal reward and specific denoising behaviors, leading to erroneous credit assignment [Luo *et al.*, 2025]. Second, relying solely on a single human-preference reward makes the policy prone to reward hacking, where the model caters to perceptual metrics by deviating from original structures or introducing unnatural artifacts. Moreover, directly aggregating multi-dimensional rewards within the GRPO framework presents inherent limitations. Naive aggregation tends to blur subtle discrepancies among different reward components, resulting in reward signal collapse [Liu *et al.*, 2026b]. Simultaneously, it allows high-variance rewards to dominate the optimization direction, thereby suppressing gradient signals from other critical objectives [Lyu *et al.*, 2025].

In this paper, we propose an unsupervised low-light image enhancement framework based on Group Relative Policy Optimization (GRPO), as shown in Fig. 1, with the advantage computation detailed in Eqs. (14) and (17). We introduce online reinforcement learning into the low-light enhancement task, modeling human preferences as terminal rewards while requiring only low-light image inputs. By directly optimizing the generative distribution, the model generates high-quality results that are aesthetically pleasing, well-exposed, and vividly colored. To mitigate the issues of phasic exploration and credit assignment during diffusion denoising, and inspired by its success in generation tasks [Li *et al.*, 2025], we introduce the Sliding Window Hybrid ODE-SDE Sampling strategy. This strategy defines a dynamic time window that slides along the denoising direction, restricting stochastic exploration to the sub-interval covered by the window while maintaining deterministic sampling in the remaining stages. Through this mechanism, we achieve an ordered, coarse-to-fine exploration and enable the precise attribution of terminal reward differences to the exploratory noise within the window, thereby significantly alleviating the credit assignment problem.

We construct a Synergistic Perception-Fidelity Reward that encompasses human perceptual preferences, content fidelity, illumination uniformity, and color naturalness, to effectively constrain the model's exploration space while enhancing visual quality. To achieve stable and balanced multi-reward optimization, we adopt Independent Advantage Estimation. This strategy calculates normalized relative advantages separately within each sampling group, mapping distinct reward components to a consistent magnitude. This ensures that policy updates rely on relative performance rather than raw numerical discrepancies, thereby effectively mitigating the gradient suppression problem in multi-objective optimization. Furthermore, based on our experimental setup, we perform a Trajectory Branching Analysis (Fig. 3) to quantify the terminal reward variance induced by independent stochastic noise injection at specific timesteps. Experimental results indicate that this variance exhibits a distinct non-uniform distribution along the time axis. Building on this observation, we propose a Temporal-Aware Dynamic Weighting Mechanism. By assigning higher weights to these critical stages during optimization, we guide the model to prioritize the denoising intervals that most significantly impact final perceptual quality, thereby efficiently achieving synergistic optimization of multiple rewards. Our contributions can be summarized as follows:

1. We pioneer the application of GRPO to unsupervised low-light image enhancement and, by integrating the Sliding Window Hybrid ODE-SDE Sampling strategy, achieve efficient coarse-to-fine exploration and precise advantage attribution.
2. We construct a Synergistic Perception-Fidelity Reward and, coupled with Independent Advantage Estimation and the Temporal-Aware Dynamic Weighting Mechanism, achieve stable and precise synergistic optimization of multiple rewards.
3. Extensive experiments on multiple public real-world low-light datasets demonstrate that the proposed framework significantly improves the perceptual performance of existing generative models, yielding enhanced results that better align with human aesthetics.

2 Related Work

2.1 Low-light Image Enhancement

Traditional low-light image enhancement (LLIE) methods, such as histogram equalization [Wu *et al.*, 2024], gamma correction [Wang *et al.*, 2023], and Retinex models [Ren *et al.*, 2020], rely on hand-crafted priors or fixed transformations. When applied to perceptually challenging real-world scenes [Ji *et al.*, 2024], these approaches often suffer from over-enhancement or color distortion [Wu *et al.*, 2022]. While deep learning-based LLIE methods improve accuracy and robustness [Li *et al.*, 2024], supervised paradigms depend on large-scale paired datasets, which are costly to collect and often limit generalization under real-world distribu-

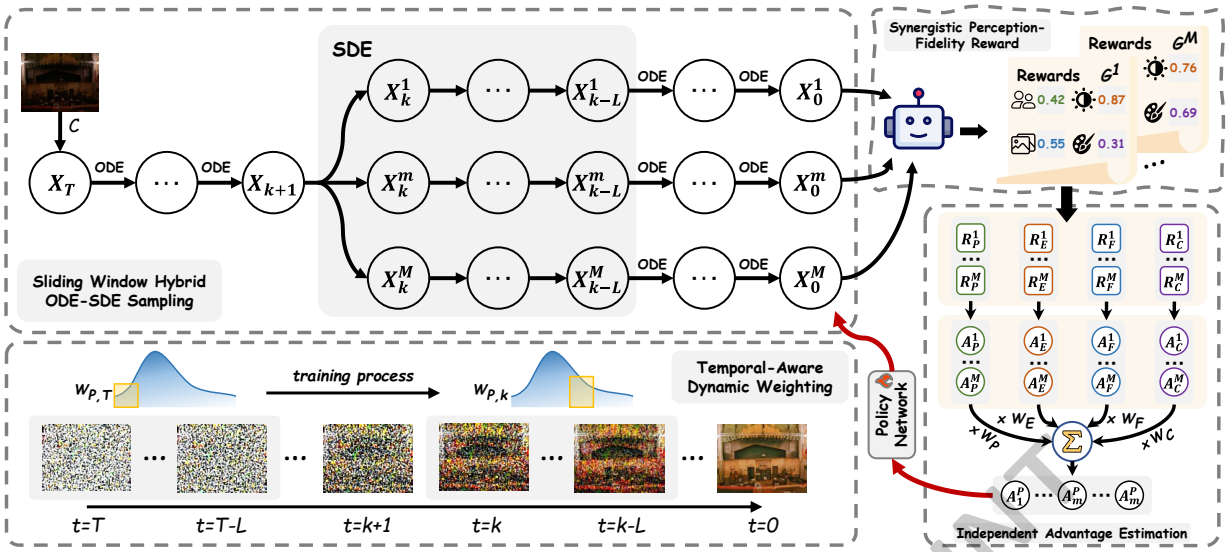

Figure 2: Unsupervised low-light enhancement framework based on Sliding Window Hybrid ODE-SDE Sampling, Independent Advantage Estimation, and Temporal-Aware Dynamic Weighting.

tion shifts [Mumuni *et al.*, 2024; Lu *et al.*, 2025]. In contrast, unsupervised methods [Ma *et al.*, 2022] eliminate the need for paired training data by introducing illumination [Shi *et al.*, 2024] or physical priors [Wang *et al.*, 2024] as constraints, or by adopting generative architectures [Jiang *et al.*, 2021]. Consequently, unsupervised approaches often generalize better in complex real-world environments.

2.2 Preference Alignment

Preference-driven reinforcement learning has been effective in image generation [Xu *et al.*, 2023]. DanceGRPO [Xue *et al.*, 2025] and FlowGRPO [Liu *et al.*, 2026a] apply GRPO to visual generation by reformulating the sampling process via SDEs. Recent studies have explored optimization strategies for the advantage attribution mechanism within the GRPO framework. Chunk-GRPO [Luo *et al.*, 2025] partitions denoising sequences into independent sub-chunks for localized estimation, effectively mitigating high variance and credit assignment ambiguity caused by global attribution. TEMPFLOW-GRPO [He *et al.*, 2025] implements time-sensitive advantage assignment by utilizing noise-aware weighting and trajectory branching to account for the varying contributions of different timesteps to generation quality. Dynamic-TreeRPO [Fu *et al.*, 2025] uses structured tree-based sampling in the Markov Decision Process, providing accurate reference baselines through shared nodes and breaking the attribution bottleneck of independent trajectories.

3 Method

The proposed online reinforcement learning enhancement framework initializes its policy network with LightenDiffusion, a pre-trained unsupervised diffusion-based model. The overall pipeline is illustrated in Fig. 2.

3.1 Diffusion Policy Optimization

RL Formulation. Diffusion models define a forward process that gradually adds noise to perturb original data into a simple distribution. The model is then trained to approximate the reverse generation process, which reconstructs the original data through progressive denoising. DDIMs [Song *et al.*, 2020] are defined as a class of Non-Markovian diffusion processes.

Specifically, given a noisy sample x_t and a trained noise prediction network $\epsilon_\theta(x_t, t)$, the reverse denoising process of DDIM from t to $t-1$ can be expressed as:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta(x_t, t) + \sigma_t \epsilon_t \tag{1}$$

Here, α_t denotes the cumulative product of the noise schedule, and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents standard Gaussian noise. The parameter σ_t controls the stochasticity of the sampling and is defined as:

$$\sigma_t = \eta \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}} \sqrt{1 - \frac{\alpha_t}{\alpha_{t-1}}} \tag{2}$$

where $\eta \in [0, 1]$ is a hyperparameter. When $\eta = 0$, the sampling process is deterministic. Conversely, when $\eta > 0$, a random noise term is introduced to add stochasticity to the generation trajectory. Such controlled stochasticity is pivotal for the integration of reinforcement learning, offering the requisite diversity for policy exploration.

Following current diffusion-based reinforcement learning methods [Black *et al.*, 2024], we utilize the denoising network of the diffusion model as the policy $\pi(a_t|s_t)$. We reframe the diffusion process as a multi-step Markov Decision Process (MDP) defined by the tuple $(S, A, \rho_0, \pi, P, R)$:

$$
\begin{aligned}
s_t &\triangleq (c, t, x_t), \quad \rho_0(s_0) \triangleq (p(c), \delta_T, \mathcal{N}(\mathbf{0}, \mathbf{I})), \\
a_t &\triangleq x_{t-1}, \quad \pi(a_t|s_t) \triangleq p_\theta(x_{t-1}|x_t, t, c), \\
P(s_{t+1}|s_t, a_t) &\triangleq (\delta_c, \delta_{t-1}, \delta_{a_t}), \\
R(s_t, a_t) &\triangleq \begin{cases} r(x_0, c), & \text{if } t = 0, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned} \tag{3}
$$

Here, state s_t consists of the condition c, current diffusion timestep t, and latent variable x_t. The action a_t is defined as the denoised latent variable at the next reverse diffusion step, x_{t-1}. The initial MDP state s_0 is sampled from the condition distribution $p(c)$, the initial diffusion timestep δ_T, and standard Gaussian noise $\mathcal{N}(\mathbf{0}, \mathbf{I})$, corresponding to the starting point of the denoising process.

The state transition P is deterministic. Once the policy takes action a_t by selecting the next reverse denoising state, the system deterministically transitions from s_t to $s_{t-1} = (c, t-1, a_t)$. The reward function $R(s_t, a_t)$ is defined as a sparse reward. It returns the quality score $r(x_0, c)$ of the generated image only at the end of the denoising process ($t = 0$) and remains zero at all other time steps.

We optimize the diffusion model to learn a policy π_θ that maximizes the expected reward of the generated samples at the final step. This is typically formulated as maximizing the following objective function:

$$
J(\theta) = \mathbb{E}_{p(c)} \mathbb{E}_{p_\theta(x_0|c)}[R(x_0, c)] \tag{4}
$$

where $p(c)$ represents the distribution of input conditions, and $p_\theta(x_0|c)$ denotes the sample distribution generated by the model.

Standard GRPO. To efficiently maximize the objective function defined in Eq. (4), we employ Group Relative Policy Optimization (GRPO). GRPO eliminates the need for a parametric value function by leveraging group-based statistics to estimate the baseline and advantage directly from sampled trajectories.

Specifically, for each input condition c, we generate a group of M independent trajectories $\{\tau_i\}_{i=1}^M$ using the current policy π_θ, resulting in a set of enhanced images $\{x_0^i\}_{i=1}^M$. We then compute the reward $R(x_0^i, c)$ for each sample. The advantage A^i for the i-th sample is estimated by normalizing its reward against the group statistics:

$$
A^i = \frac{R(x_0^i, c) - \mu(R)}{\sigma(R) + \delta} \tag{5}
$$

where $\mu(R)$ and $\sigma(R)$ denote the mean and standard deviation of the rewards within the group, and δ is a small constant for numerical stability.

To update the policy, we optimize the following surrogate objective, which combines the clipped advantage for stability and a KL-divergence penalty to ensure the model stays close to the reference pre-trained distribution π_{ref}:

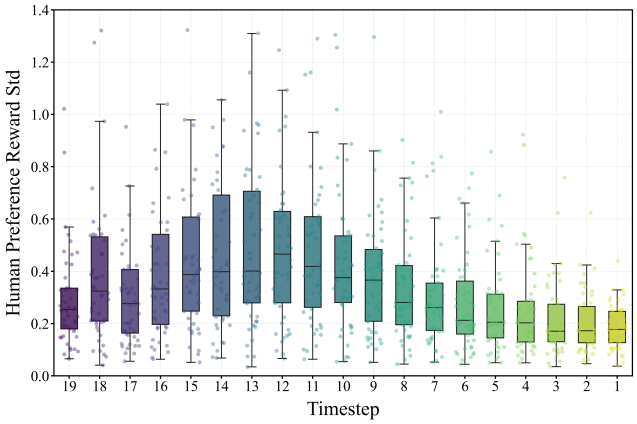

Figure 3: Reward Variance Analysis. We plot the standard deviation of the Human Preference Reward at each denoising step for 50 images randomly sampled from the LOL-v1 training set, with a group size of 12. The results reveal that the reward variance peaks during the early-to-intermediate stage, indicating that exploration in this stage contributes most to the terminal reward.

$$
\begin{aligned}
J(\theta) = \mathbb{E}_{c\sim p(c)} \Bigg[\frac{1}{M} \sum_{i=1}^{M} \Bigg(\frac{1}{T} \sum_{t=1}^{T} \min \big(\rho_t^i A^i, \\
\text{clip}(\rho_t^i, 1-\epsilon, 1+\epsilon) A^i \big) - \beta D_{\text{KL}}(\pi_\theta(\cdot|c) || \pi_{\text{ref}}(\cdot|c)) \Bigg) \Bigg]
\end{aligned} \tag{6}
$$

Here, $\rho_t^i(\theta) = \frac{\pi_\theta(a_t^i|s_t^i)}{\pi_{\theta_{\text{old}}}(a_t^i|s_t^i)}$ is the probability ratio, T denotes the total number of denoising timesteps, ϵ is the clipping threshold, and β controls the strength of the KL regularization.

3.2 Sliding Window Hybrid ODE-SDE Sampling

The standard diffusion-based GRPO method injects randomness at every timestep $t \in \{T, T-1, \ldots, 0\}$ of the denoising process. The trajectory generated by this sampling method is defined as $\tau = (s_T, a_T, s_{T-1}, a_{T-1}, \ldots, s_0, a_0)$. The sparse terminal reward $r(x_0, c)$, calculated at the end of the trajectory, is uniformly distributed to the actions of all timesteps. This time-uniform attribution method overlooks the inherent temporal characteristics of the diffusion denoising process. To achieve precise credit assignment, inspired by previous work [Li *et al.*, 2025], we adopt a sliding window mixed ODE-SDE sampling strategy.

We concentrate the injection of stochasticity within a dynamic sub-interval of the denoising timesteps. We define a sliding window $W(k)$ that covers a portion of the timesteps in the denoising process:

$$
W(k) = \{k, k-1, \ldots, k-L\}, \quad L \le k \le T \tag{7}
$$

where k is the starting timestep index of the sliding window, and L is the window span.

Based on this window, we construct a hybrid ODE-SDE sampling method. Starting from the initial noise x_T, the trajectory evolves deterministically until reaching the specified

window. To implement the dynamic switching of sampling, we define the hyperparameter η in Eq. (2) in the following time-step dependent form:

$$\eta_t = \begin{cases} \eta_{\text{set}}, & \text{if } t \in W(k) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Specifically, upon entering the window ($t \in W(k)$), we set η_t to a preset constant $\eta_{\text{set}} > 0$. According to Eq. (2), this yields $\sigma_t > 0$. This also activates the random noise term $\sigma_t \epsilon_t$ in Eq. (1). The model shifts to performing SDE sampling to explore the policy space within the current timestep window. After leaving the window, deterministic sampling is resumed until the final image x_0 is generated. This strategy shortens the effective sequence length of the Markov Decision Process (MDP) to the subset $\tau_k = (s_k, a_k, s_{k-1}, a_{k-1}, \ldots, s_{k-L}, a_{k-L})$.

The window position k slides dynamically from T to L to guide an ordered coarse-to-fine exploration. Governed by the sliding interval I and step size Δk, the update rule is defined as:

$$k_{\text{iter}} = T - \left\lfloor \frac{\text{iter}}{I} \right\rfloor \Delta k \tag{9}$$

For brevity, we omit the subscript $iter$ and use k to represent the current window position in subsequent discussions. By concentrating randomness solely within $W(k)$, any variation in the final image (and resulting reward) is directly attributable to the noise in this specific window, thereby achieving precise credit assignment.

3.3 Synergistic Perception-Fidelity Reward

Human Preference Reward ($R_{\mathbf{P}}$). We employ MUSIQ, a no-reference image quality assessment metric, as the perceptual reward to estimate the visual quality of generated images without requiring paired ground truth. This reward encourages visually plausible and perceptually preferred enhancement results.

Fidelity Reward ($R_{\mathbf{F}}$). We employ LPIPS as a fidelity constraint to better capture human-perceptual structural differences. Due to the lack of ground truth, we compute the negative LPIPS distance between the generated image x_0 and a reference image x_{ref} produced by a pre-trained model.

Exposure Reward ($R_{\mathbf{E}}$). To achieve adaptive dynamic range adjustment, following [Lin *et al.*, 2025], we utilize a spatially variant exposure map E to constrain the local exposure of the generated image $\hat{x}_0$. Unlike global adjustments, E assigns non-uniform target values based on the input luminance Y. This balances underexposed and overexposed regions. Based on the statistics of normal-light images, E is calculated as:

$$E = \lambda_E \cdot \text{Norm}(Y(c) - Y_{\text{avg}}) + b_E \tag{10}$$

We define the exposure reward as the negative squared difference between the mean exposure of the output image and the target map E:

$$R_{\text{E}} = - \left\| \text{Mean}(x_0) - \text{Mean}(E) \right\|_2^2 \tag{11}$$

Color Reward ($R_{\mathbf{C}}$). To prevent color distortion, we rely on the Retinex theory. We utilize the property that reflectance R is independent of illumination changes as a physical consistency prior. We employ a pre-trained decomposition network $\mathcal{F}(\cdot)$ (RNet) to extract reflectance features. We enforce consistency between the reflectance of the generated image x_0 and the input condition c using the following reward:

$$R_{\text{C}} = - \left\| \mathcal{F}(x_0) - \mathcal{F}(c) \right\|_2^2 \tag{12}$$

3.4 Temporal-Aware Independent Advantage Aggregation

Independent Advantage Estimation. Traditional multi-reward reinforcement learning methods typically calculate the total reward by performing a weighted summation of all reward components:

$$R_{\text{total}}(x_0^i, c) = \sum_j w_j R_j(x_0^i, c) \tag{13}$$

Subsequently, group-wise normalization is applied to this total reward using Eq. (5) to calculate the relative advantage. However, significant discrepancies exist in the numerical scales and variance distributions of different reward functions. Direct summation causes reward components with high variance to dominate the optimization direction, suppressing gradient signals from other critical objectives.

To address this issue, we compute the advantage values for each reward independently. Specifically, for each input condition c, we sample a group of independent trajectories $\{\tau_i\}_{i=1}^M$ using the current policy π_θ. Through this sampling process, we obtain a set of enhanced images $\{x_0^i\}_{i=1}^M$. For the j-th reward objective, we first calculate its raw reward value $R_j(x_0^i, c)$ within the group, and then compute its normalized relative advantage A_j^i:

$$A_j^i = \frac{R_j(x_0^i, c) - \mu_j}{\sigma_j + \delta} \tag{14}$$

where μ_j and σ_j denote the mean and standard deviation of that reward within the group, respectively, and δ is a small constant for numerical stability. In this manner, each reward component is rescaled to a consistent magnitude.

Temporal-Aware Dynamic Weighting. After obtaining the independent normalized advantages, direct linear weighted summation fails to reflect the non-uniform contribution of different denoising stages to the final generation quality. To precisely quantify the contribution of different time steps to the final perceptual reward, we employ single-variable perturbation analysis. Specifically, we employ deterministic ODE sampling for the denoising trajectory, introducing stochastic noise exclusively at a targeted single timestep. We then assess the significance of this timestep by measuring the variance of the terminal perceptual rewards. As shown in Fig. 3, the reward variance exhibits a non-uniform distribution along the time axis and reaches a peak during the early-to-intermediate stage. This result reveals the dominant role of this stage in optimizing perceptual quality. To this end, we design a time-step-based dynamic weighting mechanism:

$$w(t) = 1 + \lambda_w \cdot \left(\frac{t}{\mu} \right)^{\alpha} \cdot \exp\left(\alpha \left(1 - \frac{t}{\mu} \right) \right) \tag{15}$$

Metrics	DICM		MEF		NPE		LIME		VV	
	LD	LD+Ours	LD	LD+Ours	LD	LD+Ours	LD	LD+Ours	LD	LD+Ours
LIQE ↑	3.216	**3.220**	3.425	**3.615**	3.451	**3.568**	2.778	**2.834**	1.338	**1.346**
TOPIQ-IAA ↑	4.773	**4.788**	4.759	**4.796**	4.523	**4.558**	4.628	**4.630**	**4.617**	4.589
TReS ↑	70.57	**73.73**	72.11	**75.78**	74.92	**78.36**	67.48	**70.13**	42.72	**49.44**
CLIPIQA+ ↑	0.573	**0.581**	0.595	**0.620**	0.642	**0.643**	0.552	**0.575**	0.481	**0.491**
MANIQA ↑	0.334	**0.344**	0.333	**0.357**	0.379	**0.392**	0.356	**0.369**	0.222	**0.229**
MUSIQ ↑	61.32	**63.43**	63.16	**64.88**	65.20	**67.31**	59.11	**61.19**	40.37	**40.57**
DBCNN ↑	0.491	**0.516**	0.503	**0.539**	0.518	**0.549**	0.493	**0.519**	0.353	**0.386**
PAQ2PIQ ↑	73.33	**74.50**	72.88	**73.81**	74.31	**75.50**	73.80	**74.54**	67.88	**70.14**
PI ↓	2.618	**2.416**	2.899	**2.691**	2.909	**2.756**	3.205	**3.176**	3.654	**3.221**
NIQE ↓	3.366	**3.165**	3.552	**3.449**	3.806	**3.625**	4.059	**3.987**	3.356	**2.918**

Table 1: Quantitative comparison on five real-world unpaired datasets. "LD" denotes LightenDiffusion. Best results are highlighted in **bold**.

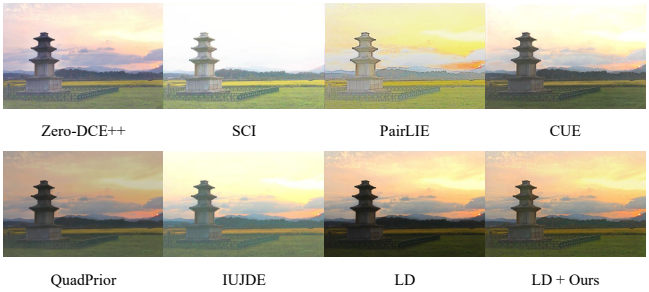

Figure 4: Qualitative comparison on the DICM dataset.

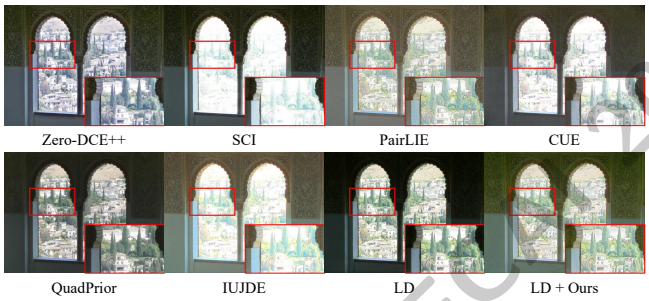

Figure 5: Qualitative comparison on the VV dataset.

Here, t denotes the index of the elapsed sampling step starting from the initial noise. Eq. (15) defines a temporal importance distribution that peaks at the early-to-intermediate stage. Considering that our sampling strategy is based on a sliding window $W(k)$ of width L, the weights of individual time steps need to be aggregated into window-level weights. Therefore, we define the perceptual dynamic weighting coefficient $w_{\mathrm{P}}(k)$ for the current sliding window $W(k)$ as the arithmetic mean of the weights of all discrete time steps within the window:

$$w_{\mathrm{P}}(k) = \frac{1}{L+1}\sum_{l=0}^{L} w\big(T-(k-l)\big) \tag{16}$$

Finally, the aggregated advantage $\mathcal{A}_k^i$ for the i-th trajectory under sliding window k consists of the dynamically adjusted human preference advantage and other constraint terms with fixed weights:

$$\mathcal{A}_k^i = w_{\mathrm{P}}(k)A_{\mathrm{P}}^i + w_{\mathrm{F}}A_{\mathrm{F}}^i + w_{\mathrm{E}}A_{\mathrm{E}}^i + w_{\mathrm{C}}A_{\mathrm{C}}^i \tag{17}$$

3.5 Optimization Objective

We perform exploration within the dynamically sliding denoising sub-interval $W(k)$ and guide policy updates using the advantage $\mathcal{A}_k^i$. This advantage effectively integrates the Synergistic Perception-Fidelity Reward via independent normalization and time-aware dynamic weighting mechanisms. Following DanceGRPO [Xue *et al.*, 2025], we omit the explicit KL regularization term to simplify the optimization objective. The final optimization objective is defined as:

$$J(\theta) = \mathbb{E}_{c\sim p(c)}\Bigg[\frac{1}{M(L+1)}\sum_{i=1}^{M}\sum_{t\in W(k)} \min\Big(\rho_t^i\mathcal{A}_k^i, \mathrm{clip}(\rho_t^i, 1-\epsilon, 1+\epsilon)\mathcal{A}_k^i\Big)\Bigg] \tag{18}$$

The detailed optimization algorithm is provided in the Supplementary Material.

4 Experiments

4.1 Training and Testing Data

The proposed enhancement framework further optimizes the LightenDiffusion backbone using the publicly available EnlightenGAN dataset [Jiang *et al.*, 2021]. The dataset contains over 1,000 unpaired images from low-light and normal-light domains, while only the low-light images are used as training inputs. To comprehensively evaluate the performance of the proposed framework, extensive experiments were conducted on a variety of benchmark datasets. The test suites include paired datasets such as LOL-v1 [Wei *et al.*, 2018] and LOLv2 [Yang *et al.*, 2021], as well as real-world unpaired datasets, including DICM [Lee *et al.*, 2013], LIME [Guo *et al.*, 2016], MEF [Ma *et al.*, 2015], NPE [Wang *et al.*, 2013], and VV [Vonikakis *et al.*, 2018]. Due to space constraints, detailed evaluation results on some datasets are provided in the supplementary materials.

Metrics	Zero-DCE	Zero-DCE++	RUAS	SCI	PairLIE	CUE	Quad Prior	IUJDE	LD	LD+ Ours
PSNR ↑	14.86	14.64	16.45	14.85	18.50	21.68	18.83	19.77	20.28	20.42
SSIM ↑	0.563	0.454	0.506	0.528	0.744	0.770	0.772	0.746	0.871	0.873
LPIPS ↓	0.237	0.248	0.194	0.238	0.164	0.151	0.156	0.170	0.176	0.179
LIQE ↑	2.359	2.459	2.482	2.543	2.132	2.297	2.477	1.738	2.467	2.583
TOPIQ-IAA ↑	4.108	4.244	4.454	4.273	3.841	4.120	4.298	3.869	4.279	4.329
TReS ↑	52.92	52.08	59.08	52.26	66.27	63.42	53.30	50.17	76.38	78.21
MANIQA ↑	0.400	0.403	0.343	0.408	0.318	0.282	0.303	0.240	0.355	**0.367**
MUSIQ ↑	55.87	57.25	59.40	57.05	56.83	56.15	58.78	50.61	63.17	65.11
DBCNN ↑	0.503	0.499	0.474	0.500	0.501	0.444	0.398	0.496	0.496	0.517
NIQE ↓	7.794	7.796	6.315	7.828	4.110	3.645	5.184	4.008	4.027	3.813

Table 2: Quantitative comparison on the LOL-v1 dataset. **Red** and **Blue** indicate the best and second-best results, respectively. The beige background indicates fidelity metrics. Results where our method outperforms the backbone (LD) but is not top two are highlighted in **bold**.

4.2 Comparison Methods and Analysis

To validate the effectiveness and generalization of the proposed framework, we designed comparative experiments from two perspectives. To demonstrate the optimization efficacy in improving perceptual quality, we conducted a direct comparison between the optimized model and the initial policy. To evaluate the overall competitiveness of our algorithm, we selected eight advanced unsupervised low-light image enhancement methods as baselines for performance benchmarking, specifically including: Zero-DCE [Guo *et al.*, 2020], Zero-DCE++ [Li *et al.*, 2021], RUAS [Liu *et al.*, 2021], SCI [Ma *et al.*, 2022], PairLIE [Fu *et al.*, 2023], CUE [Zheng *et al.*, 2023], QuadPrior [Wang *et al.*, 2024], and IUJDE [Li and Wang, 2025].

Table 1 shows that the optimized model improves nearly all perceptual metrics across five unpaired datasets. Visual comparisons in Fig. 4 and Fig. 5 demonstrate that our method generates images that are visually more natural and richer in detail. Furthermore, quantitative results on the LOL-v1 dataset in Table 2 highlight our advantages in perceptual metrics. Although slightly outperformed by other methods in certain metrics, our approach achieves overall improvements over the baseline, especially in perceptual quality. This indicates that the proposed strategy effectively exploits the model's potential, maximizing perceptual quality under constrained base performance.

4.3 Ablation Study

All experiments were conducted on the LOL-v1 test set. As shown in Table 3, relying solely on the perceptual reward R_P yields the highest MUSIQ but extremely low PSNR, confirming the reward hacking phenomenon. Incorporating the fidelity reward R_F significantly improves PSNR, effectively constraining structural consistency. Consequently, the complete synergistic reward mechanism achieves the optimal perception-fidelity balance, validating the effectiveness of Independent Advantage Estimation in mitigating multi-objective signal collapse. Regarding the sampling and weighting strategies in Table 4, our full model improves all three metrics compared with both the w/o SWHS and w/o TADW variants. These results indicate that sliding-window stochastic exploration helps alleviate credit-assignment ambiguity, while Temporal-Aware Dynamic Weighting further strengthens optimization by assigning higher weights to critical denoising stages.

R_P	R_F	R_C	R_E	PSNR ↑	SSIM ↑	MUSIQ ↑
✓				15.17	0.793	**65.49**
✓	✓			20.24	0.872	64.98
✓	✓	✓		**20.49**	0.873	64.97
✓	✓	✓	✓	20.42	**0.873**	65.11

Table 3: Ablation study on Synergistic Perception-Fidelity Reward.

	PSNR ↑	SSIM ↑	MUSIQ ↑
w/o SWHS	20.28	0.872	64.96
Ours	**20.42**	**0.873**	**65.11**
w/o TADW	20.35	0.870	65.09
Ours	**20.42**	**0.873**	**65.11**

Table 4: Ablation studies. **Top:** Effectiveness of the Sliding Window Hybrid ODE-SDE Sampling (SWHS). **Bottom:** Effectiveness of the Temporal-Aware Dynamic Weighting (TADW).

5 Conclusion

We proposed an unsupervised framework based on GRPO to directly optimize diffusion policies using terminal rewards. To mitigate credit assignment in trajectory-wise optimization, we introduced Sliding Window Hybrid ODE-SDE Sampling, confining stochasticity to dynamic sub-intervals for efficient exploration. Furthermore, to enforce fidelity and prevent reward hacking, we constructed a Synergistic Perception-Fidelity Reward with Independent Advantage Estimation and Temporal-Aware Dynamic Weighting. This approach mitigates signal collapse in multi-objective optimization. Experiments demonstrate that our method generates high-quality results aligning with human preferences without paired data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62422609, and in part by the Fundamental Research Funds for the Central Universities under Grant WK2102026001.

References

[Black *et al.*, 2024] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pages 4965–4987, 2024.

[Fu *et al.*, 2023] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22252–22261, 2023.

[Fu *et al.*, 2025] Xiaolong Fu, Lichen Ma, Zipeng Guo, Gaojing Zhou, Chongxiao Wang, ShiPing Dong, Shizhe Zhou, Ximan Liu, Jingling Fu, Tan Lit Sin, et al. Dynamic-treerpo: Breaking the independent trajectory bottleneck with structured sampling. *arXiv preprint arXiv:2509.23352*, 2025.

[Guo *et al.*, 2016] 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.

[Guo *et al.*, 2020] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020.

[He *et al.*, 2025] Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025.

[Ji *et al.*, 2024] Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, 21(4):617–630, 2024.

[Jiang *et al.*, 2021] 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.

[Lee *et al.*, 2013] 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.

[Li and Wang, 2025] Huaqiu Li and Haoqian Wang. Interpretable unsupervised joint denoising and enhancement for real-world low-light scenarios. In *International Conference on Learning Representations*, volume 2025, pages 39525–39537, 2025.

[Li *et al.*, 2021] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4225–4238, 2021.

[Li *et al.*, 2024] Dong Li, Yidi Liu, Xueyang Fu, Senyan Xu, and Zheng-Jun Zha. Fouriermamba: Fourier learning integration with state space models for image deraining. *arXiv preprint arXiv:2405.19450*, 2024.

[Li *et al.*, 2025] Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

[Lin *et al.*, 2025] 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, pages 5307–5315, 2025.

[Liu *et al.*, 2021] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10561–10570, 2021.

[Liu *et al.*, 2026a] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

[Liu *et al.*, 2026b] Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

[Lu *et al.*, 2025] Fan Lu, Kai Zhu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Likelihood-Aware Semantic Alignment for Full-Spectrum Out-of-Distribution Detection. *Journal of Intelligent Computing and Networking*, 1(1):1–13, 2025.

[Luo *et al.*, 2025] Yifu Luo, Penghui Du, Bo Li, Sinan Du, Tiantian Zhang, Yongzhe Chang, Kai Wu, Kun Gai, and Xueqian Wang. Sample by step, optimize by chunk: Chunk-level grpo for text-to-image generation. *arXiv preprint arXiv:2510.21583*, 2025.

[Lyu *et al.*, 2025] Qiang Lyu, Zicong Chen, Chongxiao Wang, Haolin Shi, Shibo Gao, Ran Piao, Youwei Zeng, Jianlou Si, Fei Ding, Jing Li, et al. Multi-grpo: Multi-group advantage estimation for text-to-image generation with tree-based trajectories and multiple rewards. *arXiv preprint arXiv:2512.00743*, 2025.

[Ma *et al.*, 2015] Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015.

[Ma *et al.*, 2022] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5637–5646, 2022.

[Mumuni *et al.*, 2024] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. A survey of synthetic data augmentation methods in machine vision. *Machine Intelligence Research*, 21(5):831–869, 2024.

[Ren *et al.*, 2020] Xutong Ren, Wenhan Yang, Wen-Huang Cheng, and Jiaying Liu. Lr3m: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Transactions on Image Processing*, 29:5862–5876, 2020.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[Shao *et al.*, 2024] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[Shi *et al.*, 2024] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3015–3024, 2024.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

[Wang *et al.*, 2013] 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.

[Wang *et al.*, 2023] Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13128–13137, 2023.

[Wang *et al.*, 2024] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26057–26066, 2024.

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

[Wu *et al.*, 2022] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2022.

[Wu *et al.*, 2023] Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1662–1671, 2023.

[Wu *et al.*, 2024] Yuanbin Wu, Shengkui Dai, and Zhan Ma. Advanced histogram equalization based on a hybrid saliency map and novel visual prior. *Machine Intelligence Research*, 21(6):1178–1191, 2024.

[Xu *et al.*, 2023] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

[Xue *et al.*, 2025] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

[Yang *et al.*, 2021] 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.

[Yi *et al.*, 2023] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12302–12311, 2023.

[Zheng *et al.*, 2023] Naishan Zheng, Man Zhou, Yanmeng Dong, Xiangyu Rui, Jie Huang, Chongyi Li, and Feng Zhao. Empowering low-light image enhancer through customized learnable priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12559–12569, 2023.