Decoupling Direction and Magnitude: Language-Steered Flow Matching for Super-Resolution in the Dark

Jiaxin Gao[1] Ziyu Yue[2] Yaohua Liu[3,*]
Danchen Cui[4] Zhihui Zhao[4] Zhixun Su[2]
[1]The Hong Kong Polytechnic University [2]Dalian University of Technology
[3]The University of Hong Kong [4]Taiyuan University of Technology
[*]Corresponding author: liuyaohua.918@gmail.com

Abstract

Image super-resolution *in the dark* is fundamentally challenged by extreme spatial heterogeneity: severely underexposed and noisy regions demand aggressive generative enhancement to reconstruct missing details, while relatively well-exposed areas require conservative updates to prevent hallucinated textures. Standard diffusion and flow-matching models, however, rely on globally uniform integration trajectories, making them sub-optimal for such spatially varying degradations. In this paper, we propose **DeLang-SR** (**De**gradation-**Lang**uage steered **S**uper-**R**esolution), a one-step flow-matching framework that explicitly decouples the restoration *direction* and *magnitude*. Rather than treating low-light corruption as arbitrary latent features, DeLang-SR translates measurable physical statistics into structured degradation language, harnessing the semantic prior of pretrained text-to-image models. The degradation-language intent, complemented by image-specific visual tokens, steers the velocity field (direction) toward restoration-relevant manifolds. Simultaneously, a continuous spatial update-strength field acts as a locally varying Euler step-size field (magnitude), applying stronger integration steps in heavily degraded shadows while enforcing conservative updates in reliable regions. Experiments on paired and real dark images, together with ablations and diagnostic analyses, show that DeLang-SR provides an effective and interpretable one-step route to super-resolution in the dark. Code and pretrained models are available at https://github.com/moriyaya/DeLang-SR.

1 Introduction

Generative restoration models [39, 46, 9] are often built upon an implicit simplification: a single conditioned trajectory is expected to repair an entire degraded image. This simplification becomes problematic for image super-resolution *in the dark* [1, 2]. A dark low-resolution image is not uniformly corrupted. Some regions are dominated by underexposure and shadow noise, where visibility and structures are severely suppressed; other regions remain relatively reliable and should stay close to the observation. Applying the same restoration strength everywhere is therefore intrinsically risky: weak updates fail in shadows, while strong updates hallucinate textures or distort reliable details. Dark super-resolution thus exposes a clear mismatch between globally driven restoration trajectories and spatially heterogeneous degradations.

Existing low-light enhancement and super-resolution methods usually address this challenge through cascaded pipelines or task-specific architectures [18, 57, 54]. While effective in improving brightness and resolution, they are often constrained by deterministic reconstruction objectives and limited

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

generative capability under severe degradation. Pretrained diffusion and flow-matching models [50, 37] offer a powerful alternative by injecting rich natural-image priors into restoration. However, these priors are usually activated through generic prompts, degradation-agnostic conditions, or globally shared sampling schedules [36, 30]. Such mechanisms can indicate that an image should be restored, but they do not explicitly answer two more important questions: *toward which restoration direction should the latent state move, and how far should each region move?*

We revisit dark super-resolution from the perspective of flow dynamics [27, 23]. In a one-step flow update, the predicted velocity determines the *direction* of restoration, while the update magnitude determines how far each latent location moves along that direction. Existing restoration pipelines often entangle these two factors by relying on a global condition and a uniform update strength. However, dark images require direction and magnitude to be handled differently. The restoration direction should be guided by semantic degradation intent and image-specific evidence, while the update magnitude should vary across space according to local degradation severity. This motivates a direction–magnitude decoupled formulation for super-resolution in the dark.

In this paper, we propose **DeLang-SR**[1] (**De**gradation-**Lang**uage Steered **S**uper-**R**esolution), a language-steered one-step flow-matching framework. Rather than treating low-light corruption as arbitrary latent features or relying on manually written prompts, DeLang-SR converts measurable low-light statistics into structured degradation language. This produces a coarse restoration intent, discrete degradation semantics, and a continuous degradation descriptor. The language-derived condition is further complemented by visual degradation tokens extracted from the input image, providing instance-specific evidence about local degradation patterns. Together, these conditions steer the SD3 [8] velocity field toward restoration-relevant directions.

To control the magnitude of restoration, DeLang-SR predicts a degradation-aware update-strength map from the low-quality input and the continuous descriptor. This map acts as a locally varying Euler step-size field in the one-step flow update. As a result, heavily degraded shadow regions can receive larger corrective updates, whereas reliable regions are updated conservatively to preserve faithful details. The final restoration is therefore performed as a single direction-steered and magnitude-controlled flow update, avoiding iterative sampling and globally uniform restoration strength. This formulation provides both efficient inference and interpretable intermediate signals, including degradation-language conditions, visual tokens, and update-strength maps.

Our contributions are summarized as follows:

- We identify direction–magnitude entanglement as a key limitation of globally driven generative restoration for super-resolution in the dark, and formulate one-step restoration as a language-steered flow update with spatially varying update strength.
- We propose **DeLang-SR**, which translates measurable low-light statistics into structured degradation language and combines language-derived conditions with visual degradation tokens to steer the restoration direction of a pretrained SD3 velocity field.
- We design a degradation-aware update-strength field that serves as a locally varying Euler step-size map, enabling magnitude-controlled restoration across severely degraded and relatively reliable regions.
- We conduct extensive paired and no-reference evaluations, controlled ablations, efficiency analysis, and diagnostic visualizations, demonstrating effective one-step super-resolution in the dark without oracle degradation annotations or iterative sampling.

2 Related Work

Dark and Real-World Image Super-Resolution. Image super-resolution has evolved from distortion-oriented reconstruction to real-world restoration, where the degradation process is unknown, spatially variant, and often entangled with noise, blur, compression, or illumination corruption. Early real-world SR methods improve robustness by synthesizing complex degradation pipelines or learning stronger restoration backbones [45, 21, 3, 57]. Recent transformer-based SR models further improve long-range feature aggregation and high-frequency reconstruction [62, 5, 60, 28, 26]. In parallel, low-light enhancement methods focus on recovering visibility and illumination from underexposed

[1] Project page: `https://moriyaya.github.io/DeLang-SR/`.

observations [24, 31, 44]. However, super-resolution *in the dark* is not a simple combination of enhancement and upsampling: photometric degradation and structural detail loss are coupled, and cascaded pipelines often propagate enhancement artifacts into the SR stage. RELLISUR provides a paired benchmark for joint low-light enhancement and super-resolution [1]. Recent dark SR methods jointly model brightening and resolution amplification through task-specific architectures [54, 12]. Illumination-aware modulation and conditional-flow modeling provide further alternatives for handling the coupled degradation [13, 55]. Hierarchical and nested optimization frameworks additionally coordinate the coupled task objectives or adapt their configurations [10, 14]. Nevertheless, most existing methods still learn a monolithic restoration mapping. They lack an explicit mechanism to determine how strongly different spatial regions should be restored under heterogeneous darkness.

Diffusion- and Flow-Based Super-Resolution. Diffusion and flow models have turned large-scale generative priors into a practical foundation for perceptual image restoration. Starting from denoising diffusion models [16], latent diffusion models [38] improve high-resolution synthesis efficiency by operating in a compressed latent space, enabling large-scale text-to-image systems such as Stable Diffusion and SDXL [35]. More recently, flow matching and rectified flow formulate generation through continuous velocity fields [23, 27], and Stable Diffusion 3 scales this paradigm with rectified-flow transformers for high-resolution text-to-image synthesis [8]. These advances make pretrained diffusion and flow priors increasingly attractive for image restoration and super-resolution.

A growing body of SR and restoration works adapts such priors to real-world degradations. StableSR [43] and DiffBIR [22] exploit latent diffusion priors for real-world SR and blind restoration, ResShift [56] shortens the diffusion trajectory through residual shifting, and SeeSR [49] introduces degradation-aware semantic guidance for generative SR. To reduce the cost of iterative sampling, recent methods further compress diffusion restoration into few-step or one-step processes, including SinSR [47], TSD-SR [7], S3Diff [58], DOD [41], DHGM [20], and OMGSR [50]. These methods improve perceptual realism and inference efficiency, but most still rely on globally organized sampling trajectories or shared update schedules. SARGD [61] is a spatially adaptive diffusion approach that uses a binary artifact mask to guide latent replacement during iterative DDPM sampling. In contrast, DeLang-SR formulates heterogeneous dark SR as a direction–magnitude decoupled one-step flow update: degradation-language conditions steer the velocity direction, while a continuous degradation-conditioned field controls the local integration magnitude.

Language- and Degradation-Conditioned Restoration. Language and degradation-aware representations have emerged as effective interfaces for controlling image restoration. More broadly, adaptive hyper-guidance and constraint-learning formulations provide optimization tools for coupled image-enhancement problems [11, 25]. PromptIR [36] uses learnable prompts to adapt a blind restoration model to different corruptions, while DA-CLIP [30] exploits vision-language priors to obtain degradation-aware task guidance. For generative SR and restoration, methods such as SeeSR [49], PromptSR [4], S3Diff [58], and DeFusion [29] further introduce semantic prompts, degradation descriptions, or visual instructions to better activate pretrained diffusion priors. These works validate the value of language-level conditions, but their role is usually limited to conditioning the restoration model. Such conditioning often specifies *what* should be restored at a global level, yet it does not explicitly regulate *how strongly* the generative prior should be applied across spatially varying regions. DeLang-SR differs by assigning mathematically distinct roles to degradation-control signals: translated degradation language and visual tokens steer the SD3 velocity direction, while a degradation-aware spatial map controls the local Euler integration magnitude. This direction-magnitude separation moves beyond prompt-conditioned restoration, embedding degradation language into the mechanics of the flow update itself.

3 Method

3.1 Overview

Given a low-light low-resolution observation y, DeLang-SR aims to recover a normally exposed high-resolution image $\hat{x}$. Super-resolution in the dark is inherently spatially heterogeneous: deeply underexposed regions require stronger generative correction to recover visibility and suppressed structures, whereas relatively reliable regions should be updated conservatively to avoid over-correction and hallucinated textures. We therefore formulate the task as a *degradation-conditioned one-step*

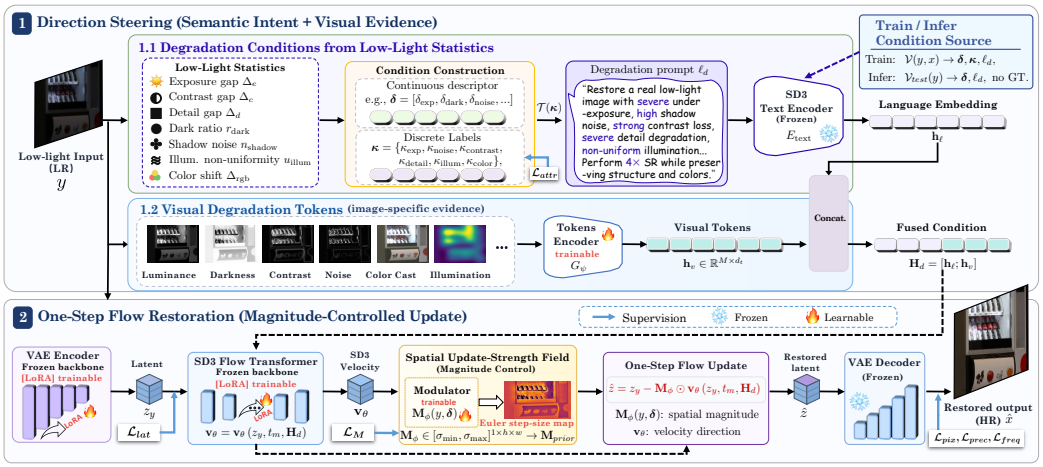

Figure 1: Pipeline of DeLang-SR. Degradation language and visual tokens form $\mathbf{H}_d$ to steer the SD3 velocity direction, while the continuous descriptor and local evidence predict $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ to control spatial update magnitude. During inference, $(\boldsymbol{\delta}, \ell_d)$ are estimated by $\mathcal{V}_{test}(y)$ without GT, and the restored image is obtained by a single direction-steered and magnitude-controlled flow update.

flow restoration problem, where the restoration direction and spatial update magnitude are explicitly decoupled.

Let x denote the paired high-quality target during training. We write the SD3 VAE latents as $z_y = \mathcal{E}_{vae}(y)$ and $z_x = \mathcal{E}_{vae}(x)$. Under the SD3 flow-matching formulation [8, 23, 27], the transformer predicts a velocity field $\mathbf{v}_\theta(z_t, t, \mathbf{H})$. Instead of iterative sampling, DeLang-SR performs a single degradation-controlled Euler update:

$$\hat{z} = z_y - \mathbf{M}_\phi(y, \boldsymbol{\delta}) \odot \mathbf{v}_\theta\left(z_y, t_m, \mathbf{H}_d\right), \qquad \hat{x} = \mathcal{D}_{vae}(\hat{z}), \tag{1}$$

where t_m is a fixed operating timestep, $\boldsymbol{\delta}$ is a continuous degradation descriptor, $\mathbf{H}_d$ is the fused degradation condition, and $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ is a spatial update-strength field. Here, $\mathbf{H}_d$ steers the SD3 velocity direction, while $\mathbf{M}_\phi$ controls the local integration magnitude. Figure 1 provides an overview of the degradation-condition construction, direction steering, magnitude control, and one-step flow update. Details of the one-step convention are provided in **Appx. C**.

3.2 Degradation Conditions from Low-Light Statistics

A generic prompt such as "restore a low-light image" cannot distinguish underexposure, shadow noise, contrast collapse, non-uniform illumination, color cast, and detail degradation. Meanwhile, raw numerical statistics require an interface to interact with the language-conditioned prior of pretrained text-to-image models. We therefore convert measurable low-light statistics into structured degradation conditions.

For a paired sample (y, x), we compute $\mathbf{s}(y, x) = [\Delta_e, \Delta_c, \Delta_d, r_{\text{dark}}, n_{\text{shadow}}, u_{\text{illum}}, \Delta_{\text{rgb}}]$, covering exposure, contrast, detail, darkness, shadow noise, illumination non-uniformity, and color shift. The statistics are mapped into structured degradation conditions:

$$(\boldsymbol{\delta}, \boldsymbol{\kappa}, \ell_d) = \mathcal{V}(y, x), \quad \begin{cases} \boldsymbol{\delta} = [\delta_{\text{exp}}, \delta_{\text{dark}}, \delta_{\text{noise}}, \delta_{\text{contrast}}, \delta_{\text{detail}}, \delta_{\text{illum}}, \delta_{\text{scale}}] \in \mathbb{R}^7, \\ \boldsymbol{\kappa} = \{\kappa_{\text{exp}}, \kappa_{\text{noise}}, \kappa_{\text{contrast}}, \kappa_{\text{detail}}, \kappa_{\text{illum}}, \kappa_{\text{color}}\}, \quad \ell_d = \mathcal{T}(\boldsymbol{\kappa}). \end{cases} \tag{2}$$

Here $\boldsymbol{\delta}$ is a continuous descriptor that guides the spatial update-strength field, $\boldsymbol{\kappa}$ denotes discrete degradation labels for prompt generation and attribute alignment, and ℓ_d is the resulting degradation-language prompt. Color cast is represented as a discrete label rather than a continuous entry of $\boldsymbol{\delta}$. The discrete labels $\boldsymbol{\kappa}$ are obtained by quantizing the corresponding statistics and inserted into a deterministic template $\ell_d = \mathcal{T}(\boldsymbol{\kappa})$. For example, a generated degradation-language prompt is:

> **Prompt.** Restore a real low-light image with **severe** underexposure, **high** shadow noise, **strong** contrast loss, **severe** detail degradation, **non-uniform** illumination, and a **warm** color cast. Perform 4× super-resolution while preserving scene structure and natural colors.

The highlighted slots are automatically instantiated from the discretized degradation labels $\boldsymbol{\kappa}$, which are derived from measurable low-light statistics rather than manual annotations. Detailed degradation statistics, descriptor-label correspondences, and prompt construction rules are provided in **Appx. B**, with the main definitions summarized in Tables 5 and 6.

At inference time, the high-quality reference x is unavailable. We estimate degradation conditions from the low-quality input itself, i.e., $(\boldsymbol{\delta}, \ell_d) = \mathcal{V}_{test}(y)$. Since LQ-only estimation is approximate, ℓ_d is treated as a coarse restoration intent rather than an oracle annotation. Thus, DeLang-SR requires no manual prompt, oracle degradation label, or high-quality reference during inference.

3.3 Direction Steering and Magnitude Control

Language and Visual Evidence for Direction Steering. The degradation prompt ℓ_d provides a global restoration intent, but language alone cannot localize degradation patterns in the current image. We therefore introduce visual degradation tokens to provide image-specific evidence. The prompt is encoded by the frozen SD3 text encoder as $\mathbf{h}_\ell = E_{\text{text}}(\ell_d)$, while a visual token encoder G_ψ produces M learnable tokens $\mathbf{h}_v = G_\psi(y) \in \mathbb{R}^{M \times d_t}$ from the RGB input, luminance, darkness, local contrast, and global low-light statistics. The detailed architecture of G_ψ is provided in **Appx. C.1**. The language embedding and visual tokens are concatenated as:

$$\mathbf{H}_d = [\mathbf{h}_\ell; \mathbf{h}_v]. \tag{3}$$

Here $\mathbf{h}_\ell$ specifies the coarse semantic intent of what degradations should be corrected, while $\mathbf{h}_v$ grounds this intent with image-specific evidence about where such degradations appear. The fused condition $\mathbf{H}_d$ is injected into the SD3 flow transformer to steer the restoration direction.

Spatial Update-Strength Field for Magnitude Control. While $\mathbf{H}_d$ controls the restoration direction, one-step flow restoration also requires controlling how strongly the velocity is applied at each latent location. We therefore predict a spatial update-strength field:

$$\mathbf{M}_\phi(y, \boldsymbol{\delta}) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \cdot \text{Sigmoid}\left(g_\phi\left(E_\phi(y), P_\phi(\boldsymbol{\delta})\right)\right). \tag{4}$$

Here E_ϕ and P_ϕ denote the visual evidence extractor and descriptor-conditioning branch, respectively; their FiLM-based implementation is detailed in **Appx. C.2**.The resulting field $\mathbf{M}_\phi(y, \boldsymbol{\delta}) \in [\sigma_{\min}, \sigma_{\max}]^{1 \times h \times w}$ is predicted at the latent resolution and broadcast along latent channels.

From the one-step flow perspective, $\mathbf{M}_\phi$ acts as a spatially varying Euler step-size field. It does not alter the predicted direction; instead, it controls the local integration magnitude. Thus, the pretrained prior is not applied through a uniform global update, but locally scaled according to degradation severity and visual evidence. Implementation details are provided in **Appx. C.2**.

3.4 One-Step Flow Restoration Dynamics

Given z_y, $\mathbf{H}_d$, and $\mathbf{M}_\phi(y, \boldsymbol{\delta})$, DeLang-SR predicts the SD3 velocity at a fixed operating timestep t_m:

$$\mathbf{v}_\theta = \mathbf{v}_\theta\left(z_y, t_m, \mathbf{H}_d\right). \tag{5}$$

The restored latent is then obtained by the one-step update in Eq. (1). Equivalently, $\mathbf{v}_\theta(z_y, t_m, \mathbf{H}_d)$ determines the restoration direction in the pretrained flow space, while $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ determines how far each latent location moves along that direction. The operating timestep t_m provides a fixed point on the pretrained SD3 flow trajectory. DeLang-SR does not use t_m to encode image-specific degradation; instead, region-dependent restoration strength is controlled by $\mathbf{M}_\phi$. For parameter-efficient adaptation, we freeze the SD3 text encoder and VAE decoder, insert LoRA adapters [17] into the VAE encoder and SD3 flow transformer, and train the degradation-control modules from scratch.

Learning Objective. To train the direction-steered and magnitude-controlled update, we supervise three complementary aspects: latent flow alignment, image reconstruction, and degradation-control regularization. For each paired sample (y, x), we construct $(\boldsymbol{\delta}, \boldsymbol{\kappa}, \ell_d) = \mathcal{V}(y, x)$, encode $z_x = \mathcal{E}_{fix}(x)$ and $z_y = \mathcal{E}_{adapt}(y)$, and build a spatially degradation-dependent flow state. Here, $\mathcal{E}_{fix}$ denotes the frozen pretrained VAE encoder used to obtain target latents, and $\mathcal{E}_{adapt}$ denotes its LoRA-adapted counterpart for low-light inputs:

$$z_{\mathbf{M}} = (1 - \mathbf{M}_\phi(y, \boldsymbol{\delta})) \odot z_x + \mathbf{M}_\phi(y, \boldsymbol{\delta}) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{6}$$

with $\mathcal{L}_{lat} = \|z_y - z_{\mathbf{M}}\|_2^2$. The restored image is supervised by $\mathcal{L}_{pix} = \|\hat{x} - x\|_1$, a DISTS-style perceptual loss $\mathcal{L}_{perc} = \mathrm{DISTS}_{DINOv3}(\hat{x}, x)$ [6, 40], and a frequency-decoupled loss:

$$\mathcal{L}_{freq} = \|\mathrm{Low}(\hat{x}) - \mathrm{Low}(x)\|_1 + \|\mathrm{High}(\hat{x}) - \mathrm{High}(x)\|_1. \tag{7}$$

We further regularize the update-strength field with $\mathcal{L}_M = \|\mathbf{M}_\phi(y, \boldsymbol{\delta}) - \mathbf{M}_{prior}(y)\|_1$, where $\mathbf{M}_{prior}(y) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})d_y$. Finally, an attribute head predicts the discrete labels from the visual tokens, yielding $\mathcal{L}_{attr} = \sum_{k=1}^{K} \mathrm{CE}(\hat{\kappa}^k, \kappa^k)$, where $K = 6$ and CE denotes cross-entropy. The total objective is:

$$\mathcal{L} = \lambda_{lat}\mathcal{L}_{lat} + \lambda_{pix}\mathcal{L}_{pix} + \lambda_{freq}\mathcal{L}_{freq} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_M\mathcal{L}_M + \lambda_{attr}\mathcal{L}_{attr}. \tag{8}$$

At inference time, DeLang-SR estimates $(\boldsymbol{\delta}, \ell_d)$ from y, constructs $\mathbf{H}_d$, $\mathbf{M}_\phi$, and performs one-step update in Eq. (1) without high-quality references, oracle labels, manual prompts, or iterative sampling.

4 Experiments

4.1 Experimental Protocol

Datasets. We evaluate DeLang-SR on both paired and real dark super-resolution settings. For paired evaluation, we use RELLISUR [1], which contains 3,610 training pairs and 425 testing pairs under both $\times 2$ and $\times 4$ settings. The low-light image is used as the degraded observation y, and the paired normal-light high-resolution image is used as the target x. To assess generalization to real nighttime scenes, we further evaluate on 100 images from DarkFace [53] without reference.

Metrics. For paired evaluation, we report PSNR, SSIM [48], LPIPS [59], and DISTS [6] to measure fidelity, structure, and perceptual consistency. Since dark SR also requires natural illumination and chromatic fidelity, we additionally report exposure error (ExpErr) and color consistency error (ColorErr), computed against the paired high-quality target. For no-reference evaluation, we report NIQE [32], MUSIQ [19], MANIQA [51], and CLIP-IQA [42] on RELLISUR outputs and real DarkFace images. We also report runtime and flow-update steps to assess inference efficiency.

Baselines. We compare DeLang-SR with three groups of baselines: (i) general SR/restoration backbones, including Restormer [57], SwinIR [21], SRFormer [62], HAT [3], RGT [5], HiT-SR [60], PFT-SR [28], ASID [33], and CATANet [26]; (ii) cascaded LLE+SR pipelines built from RUAS [24], SCI [31], and LLFormer [44] with strong SR backbones; and (iii) recent diffusion- or degradation-guided restoration/SR methods, including S3Diff [58], DOD [41], and DHGM [20]. All methods are evaluated on the same test images and output resolutions.

Implementation. DeLang-SR is initialized from a pretrained Stable Diffusion 3 (SD3) model [8]. We freeze the SD3 text encoder and VAE decoder, and insert lightweight LoRA adapters [17] into the VAE encoder and SD3 flow transformer. The newly introduced degradation-control modules, including the visual token encoder G_ψ, spatial update-strength predictor $\mathbf{M}_\phi$, and attribute prediction head, are optimized jointly with the LoRA adapters. We use $M = 4$ visual degradation tokens and a 7-dimensional descriptor $\boldsymbol{\delta}$. At inference time, DeLang-SR estimates $(\boldsymbol{\delta}, \ell_d) = \mathcal{V}_{test}(y)$ from the input y and restores the image with a single direction-steered and magnitude-controlled SD3 flow update. Unless otherwise specified, main comparisons use the best full checkpoint, while controlled ablations are conducted on RELLISUR $\times 2$ under the same 5K-step training budget.

4.2 Diagnostic Analysis of Degradation-Control Mechanisms

Spatial update-strength field. Before comparing with competing methods, we first visualize the proposed spatial update-strength field $\mathbf{M}_\phi(y, \boldsymbol{\delta})$, which is the key component for magnitude control in DeLang-SR. As shown in Fig. 2, we present the input image, darkness map, predicted update-strength field, overlay visualization, restored output, and ground truth. The predicted field is spatially non-uniform and provides an interpretable local step-size map for one-step SD3 restoration. This supports our direction–magnitude decoupling view: the fused condition steers the restoration direction, while $\mathbf{M}_\phi$ controls the local update magnitude according to degradation severity. Additional update-strength visualizations are provided in **Appx. G.3**.

LQ-only condition construction. Since oracle degradation statistics are unavailable at test time, we evaluate three condition sources with the same checkpoint: a generic prompt, the proposed LQ-only

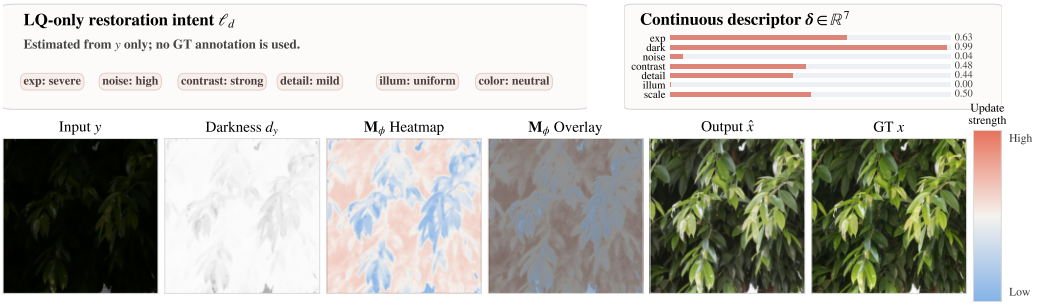

Figure 2: Visualization of the update-strength map $\mathbf{M}_\phi(y, \boldsymbol{\delta})$. Both ℓ_d and $\boldsymbol{\delta}$ are estimated from y without GT. The map provides an interpretable spatial step-size field for one-step SD3 restoration.

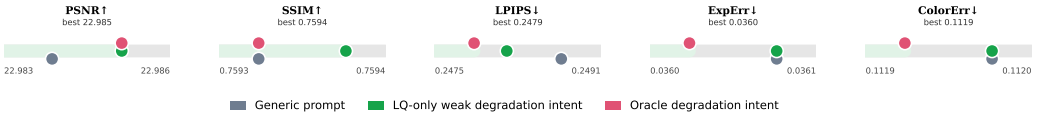

Figure 3: Inference-time condition source analysis. The close overlap among generic, LQ-only, and oracle intents shows that DeLang-SR does not require oracle degradation annotations.

Table 1: Paired quantitative comparison on ***RELLISUR***. **Pink**, **orange**, and **yellow** mark the top three results. ♣, ▲, and ♦ denote cascaded LLE+SR, general SR, and diffusion/degradation-based methods, respectively. †/‡ indicate LLE-trained/normal-light SR-trained models.

Method	Type	*RELLISUR* @ $\times 2$						*RELLISUR* @ $\times 4$					
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	ExpErr↓	ColorErr↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	ExpErr↓	ColorErr↓
†RUAS→‡SwinIR	♣	17.160	0.6170	0.5811	0.3048	0.1024	0.3074	17.058	0.6624	0.6633	0.3756	0.0871	0.2611
†SCI→‡HAT	♣	14.963	0.4389	0.5968	0.3381	0.1644	0.5038	14.750	0.4523	0.7040	0.3591	0.1644	0.5057
†LLFormer→‡HAT	♣	**21.221**	0.7205	0.4617	0.2660	**0.0549**	**0.1683**	20.123	0.7182	0.5847	0.3639	0.0636	**0.1951**
SwinIR (ICCVW'21)	▲	18.389	0.6407	0.5866	0.2862	0.0666	0.2049	17.513	0.6637	0.6968	0.3922	0.0797	0.2445
Restormer (CVPR'22)	▲	21.219	**0.7279**	0.3901	**0.2441**	0.0593	0.1800	**20.276**	**0.7207**	**0.5029**	**0.3179**	0.0660	0.2025
SRFormer (ICCV'23)	▲	19.559	0.7049	0.4724	0.2564	0.0565	0.1735	18.708	0.7055	0.6228	0.3384	**0.0630**	0.1953
HAT (CVPR'23)	▲	20.217	0.7193	0.4598	0.2707	0.0574	0.1753	19.734	0.7154	0.5730	0.3567	0.0661	0.2021
RGT (ICLR'24)	▲	18.599	0.6702	0.5479	0.2861	0.0640	0.1957	19.154	0.7163	0.5464	0.3389	0.0669	0.2016
HiT-SR (ECCV'24)	▲	16.291	0.6242	0.5903	0.4014	0.0755	0.2311	17.647	0.6826	0.6432	0.3767	0.0705	0.2218
ASID (AAAI'25)	▲	19.045	0.6971	0.4470	0.2528	0.0564	0.1723	18.155	0.7002	0.5634	0.3309	**0.0634**	**0.1943**
PFT-SR (CVPR'25)	▲	17.901	0.6493	0.5491	0.2841	0.0703	0.2151	18.534	0.7130	0.5577	0.3579	0.0671	0.2034
CATANet (CVPR'25)	▲	19.812	0.7108	0.4216	0.2559	0.0612	0.1866	18.892	0.7075	0.5505	0.3459	0.0679	0.2075
DHGM (AAAI'26)	♦	**22.663**	**0.7625**	**0.3726**	0.2484	**0.0437**	**0.1368**	**20.146**	**0.7328**	0.5443	0.3505	0.0654	0.1999
DOD (AAAI'26)	♦	20.806	0.7199	**0.2557**	**0.1570**	0.0665	0.2000	17.844	0.6218	**0.4880**	**0.2705**	0.0734	0.2235
DeLang-SR (Ours)	♦	**23.865**	**0.7700**	**0.2186**	**0.1462**	**0.0268**	**0.0856**	**21.259**	**0.7242**	**0.3290**	**0.2121**	**0.0441**	**0.1361**

degradation intent, and an oracle intent constructed from paired data. As shown in Fig. 3, their close overlap indicates that DeLang-SR does not rely on high-quality references or oracle annotations, and the generated degradation language serves as a coarse restoration intent rather than a precise label.

4.3 Main Comparison on Super-Resolution in the Dark

We then evaluate the overall restoration performance of DeLang-SR. Table 1 reports paired quantitative results on RELLISUR under $\times 2$ and $\times 4$ settings, where fidelity, perceptual quality, illumination naturalness, and color consistency are jointly evaluated. Table 2 reports no-reference perceptual quality on RELLISUR, Figure 4 shows the distribution of no-reference scores on real DarkFace images, and Table 3 provides the computational profile under the same output resolution. As an SD3-based method, DeLang-SR has a larger computational footprint than lightweight SR backbones, but it avoids iterative sampling and adapts the pretrained prior with LoRA and lightweight control modules.

More RELLISUR visual comparisons under both $\times 2$ and $\times 4$ settings are provided in **Appx. G.6**, with examples shown in Figs. 11 and 12. Compared with diffusion- or degradation-guided baselines, DeLang-SR performs restoration through a single direction-steered and magnitude-controlled SD3 flow update, achieving a stronger balance between fidelity, perceptual quality, exposure recovery, and color consistency. Figures 5 and 6 provide qualitative comparisons on RELLISUR and DarkFace,

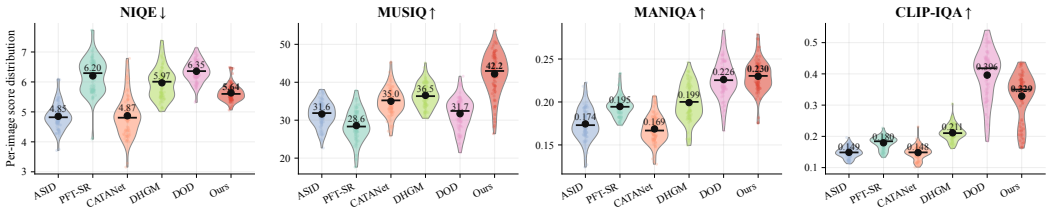

Figure 4: Distribution of no-reference perceptual quality scores on DarkFace under the $\times 2$ setting.

Table 2: No-reference perceptual quality comparison. Top-three results are highlighted.

Method	Type	*RELLISUR* @ $\times 2$				*RELLISUR* @ $\times 4$			
		NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑	NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑
HiT-SR (ECCV'24)	▲	**7.527**	26.2128	**0.2743**	**0.2580**	9.391	21.2099	0.2231	0.1851
PFT-SR (CVPR'25)	▲	8.298	31.3606	0.2217	0.2283	9.980	23.3443	**0.2460**	**0.2844**
CATANet (CVPR'25)	▲	8.015	35.5178	0.2003	0.1891	**8.735**	21.5868	0.2094	0.1966
DHGM (AAAI'26)	♦	8.795	**41.7819**	0.2140	0.2050	9.929	**25.3630**	0.2297	0.2277
DOD (AAAI'26)	♦	**7.357**	**50.2245**	**0.2819**	**0.4224**	**8.610**	**29.0542**	**0.2589**	**0.4935**
DeLang-SR (Ours)	♦	**7.285**	**54.7134**	**0.2967**	**0.4188**	**7.113**	**40.9621**	**0.2800**	**0.4253**

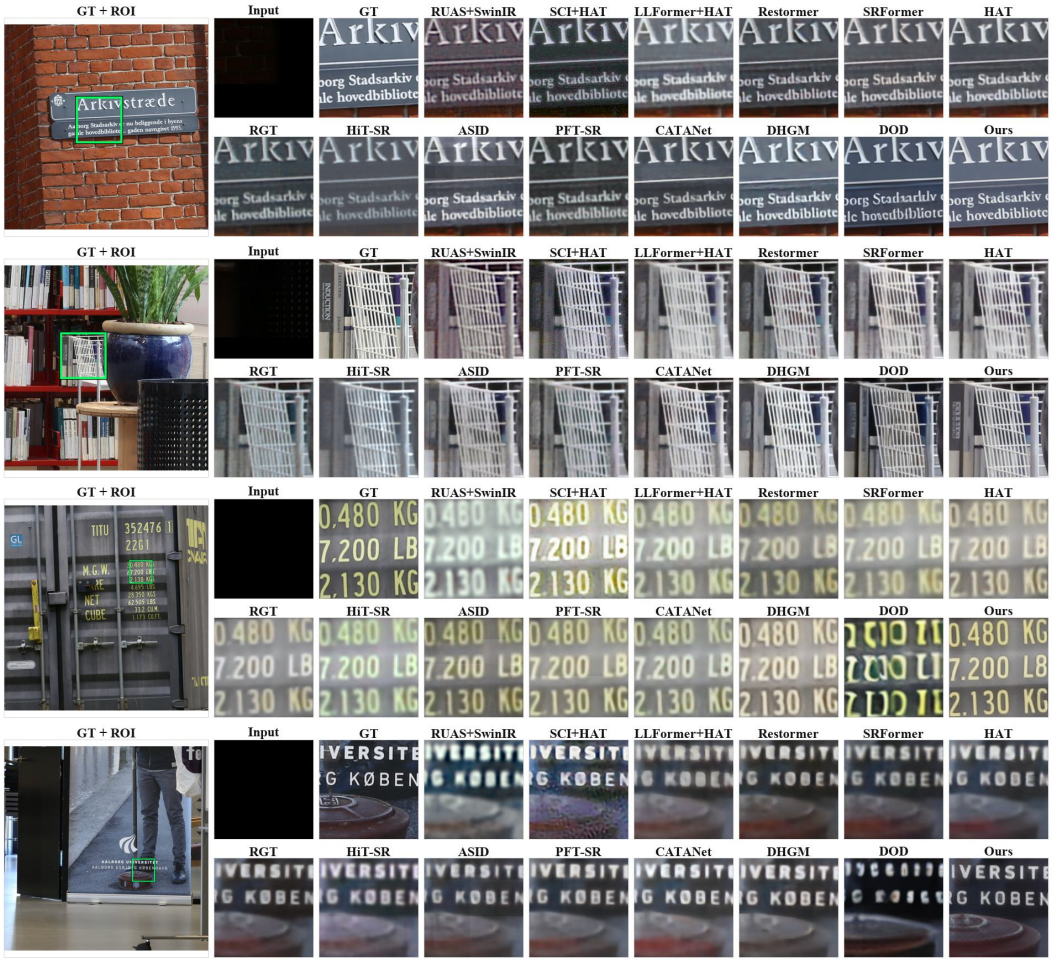

Figure 5: Visual comparison of all competing methods on RELLISUR (Top two:×2, Bottom two:×4). More RELLISUR visual comparisons under $\times 2$ and $\times 4$ settings are provided in **Appx. G.6**.

respectively. DeLang-SR preserves sharper local structures and more faithful contrast under severe dark degradation, and generalizes well to real nighttime scenes.

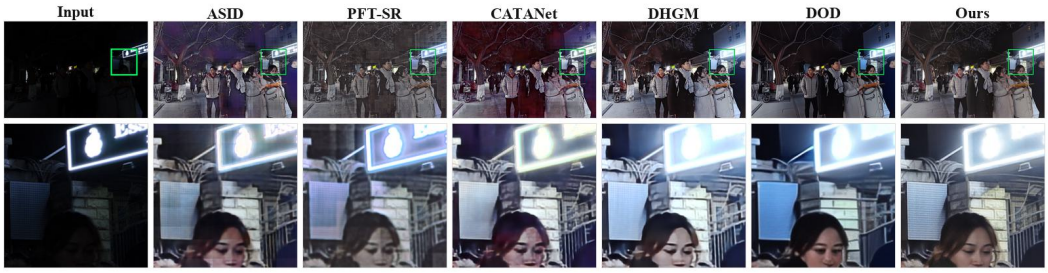

Figure 6: Visual comparison on ***DarkFace*** under the $\times 2$ setting.

Table 3: Computational profile under the same output resolution. Trainable parameters exclude frozen pretrained priors where applicable.

Metric	HiT-SR	ASID	PFT-SR	CATANet	RGT	DHGM	DOD	DeLang-SR
Trainable Params. (M)↓	0.772	0.298	19.624	0.477	10.051	10.75	958.684	4.48
FLOPs (G)↓	18.290	3.965	106.359	21.266	77.919	87.78	826.185	20390.77
Inference Time (s)↓	0.044	0.007	0.362	0.042	0.615	0.062	0.088	0.448

Table 4: Controlled ablation on ***RELLISUR*** $\times 2$ under the same 5K-step training budget. ℓ_d, $\mathbf{h}_v$, and $\boldsymbol{\delta}$ denote degradation language, visual degradation tokens, and descriptor conditioning of the update-strength predictor, respectively.

Variant	*Degradation-Control Signals*				*Quantitative Metrics*				
	ℓ_d	$\mathbf{h}_v$	$\boldsymbol{\delta}$	Attr.	PSNR↑	SSIM↑	LPIPS↓	ExpErr↓	ColorErr↓
Base model	✗	✗	✗	✗	21.358	0.7529	0.2796	0.0531	0.1614
Full DeLang-SR @5K	✓	✓	✓	✓	**22.986**	**0.7594**	0.2482	0.0360	0.1120
w/o degradation language	✗	✓	✓	✓	22.433	0.7583	0.2786	0.0383	0.1236
w/o visual degradation tokens	✓	✗	✓	✗	22.398	0.7500	0.2589	0.0423	0.1270
w/o descriptor conditioning in $\mathbf{M}_\phi$	✓	✓	✗	✓	22.754	0.7577	0.2534	**0.0335**	**0.1023**
w/o attribute alignment	✓	✓	✓	✗	22.516	0.7539	**0.2456**	0.0425	0.1274

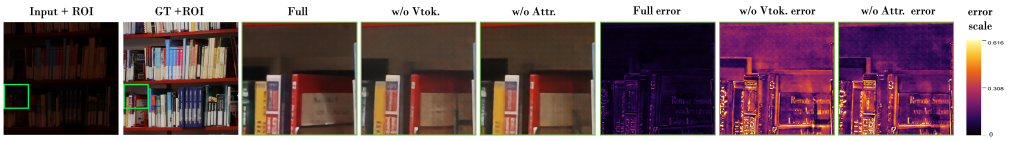

Figure 7: Qualitative ablation among different cases in Table 4. ROI crops and error maps show that visual tokens and attribute alignment improve local texture recovery.

4.4 Component Analysis and Diagnostic Ablation

We analyze different degradation-control components under a controlled 5K-step training budget on RELLISUR $\times 2$. Table 4 includes a base model implemented from the S3Diff codebase [58]. Full DeLang-SR clearly improves over this base model across fidelity, perceptual, exposure, and color metrics. Removing degradation language, visual tokens, or attribute alignment weakens the overall balance, confirming the complementary roles of language intent, image-specific evidence, and semantic token supervision. Because the attribute head operates on $\mathbf{h}_v$, the variant without visual tokens necessarily disables attribute alignment as well. Removing descriptor conditioning from $\mathbf{M}_\phi$ improves exposure/color errors slightly but reduces fidelity and perceptual quality, indicating its role in balancing conservative correction and structure preservation. Overall, the full model achieves the best trade-off across multiple metrics.

Qualitative error maps in Fig. 7 further show that removing visual tokens or attribute alignment causes weaker local texture recovery and larger reconstruction errors. Finally, Fig. 8 shows that the metrics become stable after 15K iterations, indicating that the final performance is not tied to a specific checkpoint.

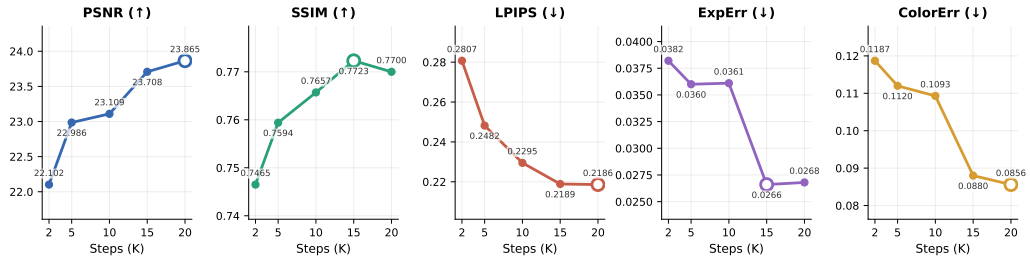

Figure 8: Checkpoint sensitivity on ***RELLISUR*** $\times 2$. Metrics become stable after 15K iterations, indicating that the final performance is not tied to a specific checkpoint.

5 Conclusion

We presented **DeLang-SR**, a language-steered one-step flow-matching framework for super-resolution in the dark. By decoupling restoration direction and spatial update magnitude, DeLang-SR uses degradation language and visual tokens to steer the SD3 velocity field, while an update-strength field controls local Euler step sizes across heterogeneous dark regions. Experiments, ablations, and diagnostics demonstrate effective and interpretable dark super-resolution without oracle degradation annotations or iterative sampling. Nevertheless, DeLang-SR still inherits the computational footprint of a large pretrained SD3 prior and may hallucinate plausible details when scene structures are nearly invisible. Future work will explore lightweight flow priors, uncertainty-aware degradation estimation, and confidence-calibrated update-strength control.

Acknowledgements and Disclosures

Funding. This research received no external funding. **Competing interests.** The authors declare no competing interests.

References

[1] Andreas Aakerberg, Kamal Nasrollahi, and Thomas B. Moeslund. Rellisur: A real low-light image super-resolution dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[2] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, pages 3291–3300, 2018.

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pages 22367–22377, 2023.

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, Xin Yuan, Linghe Kong, Guihai Chen, and Xiaokang Yang. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2311.14282*, 2023.

[5] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, and Xiaokang Yang. Recursive generalization transformer for image super-resolution. In *ICLR*, 2024.

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

[7] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *CVPR*, pages 23174–23184, 2025.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, pages 12606–12633, 2024.

[9] 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*, pages 9935–9946, 2023.

[10] Jiaxin Gao and Yaohua Liu. Enhancing images with coupled low-resolution and ultra-dark degradations: A tri-level learning framework. In *ACM MM*, pages 8642–8651, 2024.

[11] Jiaxin Gao, Xiaokun Liu, Risheng Liu, and Xin Fan. Learning adaptive hyper-guidance via proxy-based bilevel optimization for image enhancement. *The Visual Computer*, 39(4):1471–1484, 2023.

[12] Jiaxin Gao, Yaohua Liu, Ziyu Yue, Xin Fan, and Risheng Liu. Collaborative brightening and amplification of low-light imagery via bi-level adversarial learning. *Pattern Recognition*, 154: 110558, 2024.

[13] Jiaxin Gao, Ziyu Yue, Yaohua Liu, Sihan Xie, Xin Fan, and Risheng Liu. A dual-stream-modulated learning framework for illuminating and super-resolving ultra-dark images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):7500–7513, 2025.

[14] Jiaxin Gao, Yaohua Liu, Danchen Cui, and Zhihui Zhao. SNOC: Subtle nested objective configuration for joint ultra-low-light enhancement and super-resolution. In *ICMR*, pages 2172–2181, 2026.

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

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020.

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[18] 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.

[19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5148–5157, 2021.

[20] Wenjie Li, Jinglei Shi, Jin Han, Heng Guo, and Zhanyu Ma. Seeing through the rain: Resolving high-frequency conflicts in deraining and super-resolution via diffusion guidance. In *AAAI*, volume 40, pages 6468–6476, 2026.

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV Workshops*, pages 1833–1844, 2021.

[22] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pages 430–448, 2024.

[23] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

[24] 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 *CVPR*, pages 10561–10570, 2021.

[25] Risheng Liu, Jiaxin Gao, Xuan Liu, and Xin Fan. Learning with constraint learning: New perspective, solution strategy and various applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):5026–5043, 2024.

[26] Xin Liu, Jie Liu, Jie Tang, and Gangshan Wu. Catanet: Efficient content-aware token aggregation for lightweight image super-resolution. In *CVPR*, pages 17902–17912, 2025.

[27] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[28] Wei Long, Xingyu Zhou, Leheng Zhang, and Shuhang Gu. Progressive focused transformer for single image super-resolution. In *CVPR*, pages 2279–2288, 2025.

[29] Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *CVPR*, pages 12764–12777, 2025.

[30] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Controlling vision-language models for multi-task image restoration. In *ICLR*, 2024.

[31] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, pages 5637–5646, 2022.

[32] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012.

[33] Karam Park, Jae Woong Soh, and Nam Ik Cho. Efficient attention-sharing information distillation transformer for lightweight single image super-resolution. In *AAAI*, volume 39, pages 6416–6424, 2025.

[34] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32, 2018.

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[36] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *NeurIPS*, pages 71275–71293, 2023.

[37] Haina Qin, Wenyang Luo, Libin Wang, Dandan Zheng, Jingdong Chen, Ming Yang, Bing Li, and Weiming Hu. Reversing flow for image restoration. In *CVPR*, pages 7545–7558, 2025.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[39] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[40] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[41] Ni Tang, Xiaotong Luo, Zihan Cheng, Liangtai Zhou, Dongxiao Zhang, and Yanyun Qu. Diffusion once and done: Degradation-aware lora for all-in-one image restoration. In *AAAI*, volume 40, pages 9448–9456, 2026.

[42] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, pages 2555–2563, 2023.

[43] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

[44] Tao Wang, Kaihao Zhang, Tianrun Shen, Wenhan Luo, Bjorn Stenger, and Tong Lu. Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *AAAI*, volume 37, pages 2654–2662, 2023.

[45] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, pages 1905–1914, 2021.

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

[47] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *CVPR*, pages 25796–25805, 2024.

[48] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

[49] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pages 25456–25467, 2024.

[50] Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. *arXiv preprint arXiv:2508.08227*, 2025.

[51] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR Workshops*, pages 1190–1199, 2022.

[52] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, pages 3063–3072, 2020.

[53] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[54] Ziyu Yue, Jiaxin Gao, and Zhixun Su. Unveiling details in the dark: Simultaneous brightening and zooming for low-light image enhancement. In *AAAI*, pages 6899–6907, 2024.

[55] Ziyu Yue, Jiaxin Gao, Sihan Xie, Yang Liu, and Zhixun Su. LoLiSRFlow: Joint single image low-light enhancement and super-resolution via cross-scale transformer-based conditional flow. *arXiv preprint arXiv:2402.18871*, 2024.

[56] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, pages 13294–13307, 2023.

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022.

[58] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024.

[59] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[60] Xiang Zhang, Yulun Zhang, and Fisher Yu. Hit-sr: Hierarchical transformer for efficient image super-resolution. In *ECCV*, pages 483–500, 2024.

[61] Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, and Hang Xu. Self-adaptive reality-guided diffusion for artifact-free super-resolution. In *CVPR*, pages 25806–25816, 2024.

[62] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *CVPR*, pages 12780–12791, 2023.

A Appendix Overview

This appendix provides additional technical details, implementation protocols, and supplementary experimental results for DeLang-SR. **Appx. B** details the construction of degradation conditions from measurable low-light statistics, including continuous descriptors, discrete degradation labels, prompt generation, and LQ-only inference-time estimation. **Appx. C** describes the visual degradation token encoder, the spatial update-strength predictor, and the one-step flow-update convention used for direction–magnitude decoupling. **Appx. D** provides the training objectives and inference procedure. **Appx. F** summarizes dataset processing, evaluation metrics, baseline protocols, and runtime settings. Finally, **Appx. G** presents condition-source analysis, cross-dataset generalization, update-strength sensitivity and uniform-strength ablations, and qualitative comparisons.

B Details of Degradation Condition Construction

This section supplements Sec. 3.2 by providing concrete definitions of the degradation statistics, continuous degradation descriptor, discrete labels, prompt template, and LQ-only condition construction.

B.1 Low-Light Degradation Statistics

For a paired training sample (y, x), we compute a compact statistic vector $\mathbf{s}(y, x) = [\Delta_e, \Delta_c, \Delta_d, r_{\text{dark}}, n_{\text{shadow}}, u_{\text{illum}}, \Delta_{\text{rgb}}]$. Following the standard RGB luminance conversion, we use $Y(I) = 0.299I_R + 0.587I_G + 0.114I_B$. The detailed definitions are summarized in Table 5.

Table 5: Low-light degradation statistics used for degradation-condition construction. **P** and **I** denote paired-reference and input-only statistics, respectively.

Type	Statistic	Definition	Interpretation
P	Δ_e	$\text{Mean}(Y(x)) - \text{Mean}(Y(y))$	Exposure gap between the target and low-light input.
P	Δ_c	$\text{Std}(Y(x)) - \text{Std}(Y(y))$	Contrast degradation caused by insufficient illumination.
P	Δ_d	$\Delta_d = g(x) - g(y),\ \ g(I) = \text{Mean}(\lvert\nabla_x Y(I)\rvert) + \text{Mean}(\lvert\nabla_y Y(I)\rvert)$	Gradient-based detail loss.
I	r_{dark}	$\frac{1}{HW}\sum_{i,j} \mathbf{1}[Y(y)_{i,j} < \tau_d]$	Ratio of dark pixels in the low-light input.
I	n_{shadow}	$\text{Mean}_{Y(y)<\tau_n}\left(\lvert Y(y) - \mathcal{B}(Y(y))\rvert\right)$	Local residual fluctuation in dark regions.
I	u_{illum}	$\text{Std}\left(\mathcal{B}_{large}(Y(y))\right)$	Large-scale illumination non-uniformity.
P	Δ_{rgb}	$\text{Mean}_{rgb}(x) - \text{Mean}_{rgb}(y)$	RGB color shift between the target and low-light input.

B.2 Continuous Descriptor and Discrete Labels

The statistics are converted into two complementary representations in Table 6: a continuous descriptor $\boldsymbol{\delta}$ for update-strength prediction and discrete labels $\boldsymbol{\kappa}$ for prompt generation and attribute alignment. The continuous descriptor is

$$\boldsymbol{\delta} = [\delta_{\text{exp}}, \delta_{\text{dark}}, \delta_{\text{noise}}, \delta_{\text{contrast}}, \delta_{\text{detail}}, \delta_{\text{illum}}, \delta_{\text{scale}}] \in \mathbb{R}^7.$$

In practice, it is instantiated as

$$\boldsymbol{\delta} = [\bar{\Delta}_e, r_{\text{dark}}, \bar{n}_{\text{shadow}}, \bar{\Delta}_c, \bar{\Delta}_d, \mathbf{1}_{\text{nonuniform}}, r_{\text{scale}}],$$

where bars denote percentile-normalized statistics:

$$\bar{s} = \text{clip}\left(\frac{s - P_{low}(s)}{P_{high}(s) - P_{low}(s) + \epsilon}, 0, 1\right).$$

The discrete labels are

$$\boldsymbol{\kappa} = \{\kappa_{\text{exp}}, \kappa_{\text{noise}}, \kappa_{\text{contrast}}, \kappa_{\text{detail}}, \kappa_{\text{illum}}, \kappa_{\text{color}}\},$$

where each label is obtained by quantizing the corresponding statistic using fixed thresholds or dataset-level percentiles. Color cast is represented as the discrete label κ_{color} rather than a continuous descriptor entry.

Table 6: Continuous descriptor entries and discrete degradation labels. The descriptor $\boldsymbol{\delta}$ guides update-strength prediction, while $\boldsymbol{\kappa}$ fills degradation-language prompt slots and supervises attribute alignment.

Degradation Factor	Continuous Descriptor	Discrete Label / Slot	Role
Underexposure	$\delta_{\text{exp}} \leftarrow \bar{\Delta}_e$	κ_{exp}	Exposure intent and visibility recovery.
Dark regions	$\delta_{\text{dark}} \leftarrow r_{\text{dark}}$	–	Local darkness-aware magnitude control.
Shadow noise	$\delta_{\text{noise}} \leftarrow \bar{n}_{\text{shadow}}$	κ_{noise}	Noise-aware restoration intent.
Contrast loss	$\delta_{\text{contrast}} \leftarrow \bar{\Delta}_c$	κ_{contrast}	Contrast restoration intent.
Detail degradation	$\delta_{\text{detail}} \leftarrow \bar{\Delta}_d$	κ_{detail}	Detail recovery intent.
Illumination non-uniformity	$\delta_{\text{illum}} \leftarrow \mathbf{1}_{\text{nonuniform}}$	κ_{illum}	Non-uniform illumination control.
Color cast	–	κ_{color}	Color-cast prompt slot.
Scale factor	$\delta_{\text{scale}} \leftarrow r_{\text{scale}}$	**T** prompt slot	Target SR scale instruction.

B.3 Prompt Generation and LQ-Only Inference

The degradation prompt ℓ_d is generated from the discrete labels using a deterministic template, i.e., $\ell_d = \mathcal{T}(\boldsymbol{\kappa})$. Specifically, the discrete labels $\boldsymbol{\kappa}$ fill the degradation-specific slots in the prompt template, as illustrated below.

> **Degradation template.** Restore a real low-light image with `[exposure]` underexposure, `[noise]` shadow noise, `[contrast]` contrast loss, `[detail]` detail degradation, `[illumination]` illumination, and a `[color]` color cast. Perform `[scale]`$\times$ super-resolution while preserving scene structure and natural colors.
>
> **Fixed task instruction.** Enhance dark regions, suppress shadow artifacts, recover fine details, and avoid over-exposure or hallucinated content.

For example, a $4\times$ input with severe degradation may generate:

> **Instantiated prompt.** Restore a real low-light image with severe underexposure, high shadow noise, strong contrast loss, severe detail degradation, non-uniform illumination, and a warm color cast. Perform $4\times$ super-resolution while preserving scene structure and natural colors. Enhance dark regions, suppress shadow artifacts, recover fine details, and avoid over-exposure or hallucinated content.

The prompt is encoded by the frozen SD3 text encoder [8]: $\mathbf{h}_\ell = E_{\text{text}}(\ell_d)$. For efficiency, prompt embeddings can be cached offline during training.

B.4 LQ-Only Condition Construction at Inference

At inference time, the high-quality reference x is unavailable. We therefore use an LQ-only construction function $(\boldsymbol{\delta}, \ell_d) = \mathcal{V}_{test}(y)$. Exposure severity is estimated from the mean luminance and dark-region ratio of y; contrast and detail degradation are estimated from luminance statistics and gradient responses; shadow noise and illumination non-uniformity are computed from the input-only statistics defined in Table 5; and color cast is estimated from the mean RGB distribution of y. For statistics that require the paired target during training, the LQ-only variant uses the corresponding input-only proxy estimated from y. This protocol requires no high-quality reference image or oracle degradation label during deployment.

C Architecture and Flow-Update Details

This section supplements Sec. 3.3 and Sec. 3.4 by providing implementation details of the visual degradation token encoder, the spatial update-strength predictor, and the one-step flow-update convention. DeLang-SR is built upon a pretrained SD3 prior [8], whose rectified-flow formulation follows flow matching and rectified flow [23, 27]. We adapt the pretrained prior with lightweight LoRA modules [17], while the newly introduced degradation-control modules are trained from scratch.

C.1 Visual Degradation Token Encoder

The visual degradation token encoder G_ψ produces image-specific tokens that complement the global degradation-language intent. While the degradation prompt ℓ_d specifies what degradations should be corrected, the visual tokens provide instance-level evidence about where degradations appear in the current input. Given the low-light input y, we compute the luminance map $Y(y)$, darkness map $d_y = 1 - Y(y)$, and local contrast map $\ell_y = |Y(y) - \mathrm{AvgPool}(Y(y))|$. These cues are concatenated with the input image to form a local evidence tensor $\tilde{e}_y = [y, Y(y), d_y, \ell_y]$. A lightweight encoder q_ψ extracts degradation-aware features $\mathbf{F}_{\mathrm{tok}} = q_\psi(\tilde{e}_y)$. The pooled feature is concatenated with global statistics $\mathbf{r}(y)$, yielding $\mathbf{u}_{\mathrm{tok}} = [\mathrm{Pool}(\mathbf{F}_{\mathrm{tok}}); \mathbf{r}(y)]$, and is then projected into M visual degradation tokens:

$$\mathbf{h}_v = G_\psi(y) = \mathrm{reshape}\left(\mathrm{MLP}(\mathbf{u}_{\mathrm{tok}})\right) \in \mathbb{R}^{M \times d_t}. \tag{9}$$

Here d_t denotes the SD3 text-token dimension. We use $M = 4$ tokens to keep the additional sequence length negligible. The resulting $\mathbf{h}_v$ is concatenated with the degradation-language embedding $\mathbf{h}_\ell$ to form $\mathbf{H}_d = [\mathbf{h}_\ell; \mathbf{h}_v]$, which steers the SD3 velocity direction.

C.2 Spatial Update-Strength Predictor

The spatial update-strength predictor estimates $\mathbf{M}_\phi(y, \boldsymbol{\delta})$, which serves as a locally varying Euler step-size field in the one-step flow update. Different from the visual token encoder that provides conditioning tokens for direction steering, this predictor controls how strongly the predicted velocity is applied at each latent location. In Eq. (4), $E_\phi(y)$ denotes the local visual feature extractor and $P_\phi(\boldsymbol{\delta})$ denotes the descriptor-conditioned modulation branch. Concretely, we implement $E_\phi(y)$ by constructing $e_y = [y, d_y, \ell_y]$ and extracting $\mathbf{F}_y = f_\phi(e_y)$. The descriptor branch $P_\phi(\boldsymbol{\delta})$ predicts FiLM parameters [34], i.e., $(\gamma_\delta, \beta_\delta) = P_\phi(\boldsymbol{\delta})$, which modulate the local visual feature as

$$\widetilde{\mathbf{F}}_y = \mathbf{F}_y \odot (1 + \gamma_\delta) + \beta_\delta. \tag{10}$$

The residual form $1 + \gamma_\delta$ keeps the modulation close to identity at initialization and stabilizes training. The spatial update-strength field is then predicted by

$$\mathbf{M}_\phi(y, \boldsymbol{\delta}) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \cdot \mathrm{Sigmoid}\left(g_\phi(\widetilde{\mathbf{F}}_y)\right). \tag{11}$$

The field is resized to the SD3 latent resolution and broadcast along latent channels before multiplying the predicted velocity field. This design allows stronger correction in heavily degraded regions and more conservative updates in relatively reliable regions.

C.3 One-Step Flow Update and Operating Timestep

Under the SD3 flow-matching formulation [8, 23, 27], the transformer predicts a velocity field conditioned on a latent state, timestep, and context. In DeLang-SR, we use the degraded latent z_y as the starting state and perform a single direction-steered and magnitude-controlled Euler update:

$$\hat{z} = z_y - \mathbf{M}_\phi(y, \boldsymbol{\delta}) \odot \mathbf{v}_\theta(z_y, t_m, \mathbf{H}_d). \tag{12}$$

The sign follows the velocity convention used by the SD3 scheduler. In this update, $\mathbf{H}_d$ determines the restoration direction of the predicted velocity field, while $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ controls the local update magnitude.

The fixed operating timestep t_m provides a stable point on the pretrained SD3 flow trajectory. A very small timestep may lead to weak correction, while a very large timestep may over-activate the generative prior and introduce hallucinated details. In our experiments, t_m is fixed for both training and inference, so image- and region-dependent restoration strength is controlled by $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ rather than by changing the operating timestep.

C.4 Frozen and Trainable Components

DeLang-SR adapts a pretrained SD3 prior [8] in a parameter-efficient manner. The SD3 text encoder is frozen to preserve its language-conditioning space, and the VAE decoder is frozen to preserve the pretrained decoding space. We insert lightweight LoRA adapters [17] into the VAE encoder

and SD3 flow transformer to adapt the latent representation and velocity prediction to dark super-resolution. The visual degradation token encoder G_ψ, spatial update-strength predictor $\mathbf{M}_\phi$, and attribute prediction head are trained from scratch. This design preserves the general generative prior of SD3 while enabling task-specific direction steering and magnitude control for super-resolution in the dark.

D Training Objective and Inference Procedure

This section supplements Sec. 3.4 by describing the loss terms used for training.

D.1 Training Objective Details

The training objective supervises DeLang-SR from three complementary aspects: (i) latent flow alignment, which makes the adapted degraded latent compatible with the spatially modulated flow state; (ii) image reconstruction, which constrains the restored image in pixel, frequency, and perceptual spaces; and (iii) degradation-control regularization, which calibrates the spatial update-strength field and visual degradation tokens. For perceptual supervision, we compute a DISTS-style distance [6] using DINOv3 visual features [40].

For each paired sample (y, x), we construct a spatially degradation-dependent latent state:

$$z_{\mathbf{M}} = (1 - \mathbf{M}_\phi(y, \boldsymbol{\delta})) \odot z_x + \mathbf{M}_\phi(y, \boldsymbol{\delta}) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{13}$$

This state provides a spatial flow-alignment target, where $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ determines the local interpolation strength between the high-quality latent z_x and random noise ϵ.

The loss terms are summarized in Table 7. Here, $\mathrm{Low}(\cdot)$ denotes Gaussian low-pass filtering and $\mathrm{High}(I) = I - \mathrm{Low}(I)$. The update-strength prior is defined as $\mathbf{M}_{prior}(y) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min})d_y$, and $K = 6$ corresponds to the six discrete degradation attributes.

Table 7: Summary of training objectives used in DeLang-SR. The losses supervise latent flow alignment, image reconstruction, and degradation-control regularization.

Group	Term	Definition	Role
Latent alignment	$\mathcal{L}_{lat}$	$\|z_y - z_{\mathbf{M}}\|_2^2$	Aligns the degraded latent with the spatially modulated flow state.
Pixel reconstruction	$\mathcal{L}_{pix}$	$\|\hat{x} - x\|_1$	Supervises pixel-level fidelity.
Frequency reconstruction	$\mathcal{L}_{freq}$	$\|\mathrm{Low}(\hat{x}) - \mathrm{Low}(x)\|_1 + \|\mathrm{High}(\hat{x}) - \mathrm{High}(x)\|_1$	Constrains illumination/color and high-frequency details.
Perceptual consistency	$\mathcal{L}_{perc}$	$\mathrm{DISTS}_{DINOv3}(\hat{x}, x)$	Uses DISTS-style supervision with DINOv3 features [6, 40].
Update-strength prior	$\mathcal{L}_M$	$\|\mathbf{M}_\phi(y, \boldsymbol{\delta}) - \mathbf{M}_{prior}(y)\|_1$	Calibrates the spatial update-strength field.
Attribute alignment	$\mathcal{L}_{attr}$	$\sum_{k=1}^{K} \mathrm{CE}(\hat{\kappa}^k, \kappa^k)$	Aligns visual tokens with discrete degradation labels.

D.2 Training and Inference Algorithms

Algorithm 1 summarizes the training procedure of DeLang-SR. During training, paired samples are used to construct degradation conditions and supervise reconstruction, while the SD3 text encoder and VAE decoder remain frozen. Algorithm 2 summarizes the inference procedure, where only the low-light input y is required.

E Implementation Details

DeLang-SR is implemented based on a pretrained Stable Diffusion 3 (SD3) model [8]. The SD3 text encoder and VAE decoder are frozen. The VAE encoder and SD3 flow transformer are adapted with LoRA modules [17]. The visual degradation token encoder, spatial update-strength predictor, and attribute prediction head are randomly initialized and optimized jointly with the LoRA adapters. We use $M = 4$ visual degradation tokens and a 7-dimensional continuous degradation descriptor. The spatial update-strength field is bounded by $\sigma_{\min} = 0.15$ and $\sigma_{\max} = 0.95$, resized to the latent resolution, and broadcast along latent channels before multiplying the predicted flow. During training, degradation prompts are generated from metadata and fixed task instructions, and their text embeddings can be cached for efficiency.

Algorithm 1 Training procedure of DeLang-SR

Require: Paired training data $\{(y, x)\}$, fixed timestep t_m, pretrained SD3 prior.
Ensure: Trainable parameters of LoRA adapters, visual token encoder G_ψ, update-strength predictor $\mathbf{M}_\phi$, and attribute head H.

1: **Initialization.** Freeze the SD3 text encoder E_{text} and VAE decoder $\mathcal{D}_{vae}$; insert LoRA adapters into the VAE encoder and SD3 flow transformer.
2: **for** each mini-batch (y, x) **do**
3: **Degradation-condition construction.** Construct degradation conditions

$$(\boldsymbol{\delta}, \boldsymbol{\kappa}, \ell_d) = \mathcal{V}(y, x).$$

4: **Direction steering.** Encode degradation language and visual evidence:

$$\mathbf{h}_\ell = E_{\text{text}}(\ell_d), \qquad \mathbf{h}_v = G_\psi(y), \qquad \mathbf{H}_d = [\mathbf{h}_\ell; \mathbf{h}_v].$$

5: **Magnitude control.** Encode latents and predict the update-strength field:

$$z_x = \mathcal{E}_{fix}(x), \qquad z_y = \mathcal{E}_{adapt}(y), \qquad \mathbf{M} = \mathbf{M}_\phi(y, \boldsymbol{\delta}).$$

6: **One-step flow restoration.** Predict the SD3 velocity field and perform the direction-steered, magnitude-controlled update:

$$\mathbf{v} = \mathbf{v}_\theta(z_y, t_m, \mathbf{H}_d), \qquad \hat{z} = z_y - \mathbf{M} \odot \mathbf{v}, \qquad \hat{x} = \mathcal{D}_{vae}(\hat{z}).$$

7: **Objective computation.** Construct the spatial latent target

$$z_{\mathbf{M}} = (1 - \mathbf{M}) \odot z_x + \mathbf{M} \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

and compute $\mathcal{L}_{lat}$, $\mathcal{L}_{pix}$, $\mathcal{L}_{freq}$, $\mathcal{L}_{perc}$, $\mathcal{L}_M$, and $\mathcal{L}_{attr}$.
8: **Parameter update.** Update all trainable parameters by minimizing the total objective in Eq. (8).
9: **end for**

Algorithm 2 Inference procedure of DeLang-SR

Require: Low-light low-resolution input y, fixed timestep t_m.
Ensure: Restored high-resolution image $\hat{x}$.

1: **LQ-only condition construction.** Estimate degradation conditions from the input only:

$$(\boldsymbol{\delta}, \ell_d) = \mathcal{V}_{test}(y).$$

2: **Direction steering.** Encode degradation language and visual evidence:

$$\mathbf{h}_\ell = E_{\text{text}}(\ell_d), \qquad \mathbf{h}_v = G_\psi(y), \qquad \mathbf{H}_d = [\mathbf{h}_\ell; \mathbf{h}_v].$$

3: **Magnitude control.** Encode the input latent and predict the spatial update-strength field:

$$z_y = \mathcal{E}_{adapt}(y), \qquad \mathbf{M} = \mathbf{M}_\phi(y, \boldsymbol{\delta}).$$

4: **One-step flow restoration.** Predict the SD3 velocity field and perform the final update:

$$\mathbf{v} = \mathbf{v}_\theta(z_y, t_m, \mathbf{H}_d), \qquad \hat{z} = z_y - \mathbf{M} \odot \mathbf{v}, \qquad \hat{x} = \mathcal{D}_{vae}(\hat{z}).$$

5: **return** $\hat{x}$.

The loss weights are $\lambda_{lat} = 5$, $\lambda_{perc} = 5$, $\lambda_{pix} = 0.7$, $\lambda_M = 0.2$, and $\lambda_{attr} = 0.1$; the low- and high-frequency terms are each weighted by 0.5. The LoRA ranks for the VAE encoder and flow transformer are 96 and 128, respectively. For both adapters, the scaling factor α is set to the rank, the weights are initialized with a Gaussian distribution, and no LoRA dropout is applied. The VAE adapters target convolutional and attention projections, while the transformer adapters target the input embedding, attention projections, and feed-forward layers.

For controlled ablations, we train all variants on RELLISUR $\times 2$ under the same 5K-step budget. The full model used for main comparison and diagnostic visualization is selected according to validation

performance. Runtime and memory are measured during inference, excluding model loading and metric computation.

F Experimental Details

F.1 Dataset Processing

For paired evaluation, we use RELLISUR [1], which contains real low-light low-resolution images paired with normal-light high-resolution references. Following the standard split, we use 3,610 paired samples for training and 425 paired samples for testing under both $\times 2$ and $\times 4$ settings. The low-light image is denoted as y, and the paired normal-light image is denoted as x. For real-scene generalization, we additionally evaluate on 100 images selected from DarkFace [53]. Since DarkFace does not provide paired high-quality references, it is used only for no-reference perceptual evaluation and visual comparison.

We additionally evaluate cross-dataset transfer on 100 LOL-v2-Real test pairs and 50 LSRW test pairs. Because both datasets were introduced for paired low-light enhancement rather than super-resolution, we construct a $2\times$ dark-SR protocol by bicubically downsampling each low-light image and retaining the paired normal-light image as the high-resolution target. All methods use models trained on RELLISUR and are evaluated without fine-tuning on LOL-v2-Real or LSRW.

All methods are evaluated at the same output resolution. For paired RELLISUR evaluation, full-reference metrics are computed between the restored image $\hat{x}$ and the paired target x. For DarkFace, no-reference metrics are computed directly on the restored images. Unless otherwise specified, no high-quality reference image, oracle degradation label, or manual prompt is used during inference.

F.2 Evaluation Metrics

For paired evaluation, we report PSNR, SSIM [48], LPIPS [59], and DISTS [6]. PSNR and SSIM measure pixel-level fidelity and structural similarity, while LPIPS and DISTS evaluate perceptual and semantic consistency. To better reflect low-light restoration quality, we further report exposure error and color consistency error. Given the restored image $\hat{x}$ and target image x, exposure error is defined as

$$\mathrm{ExpErr} = \left|\mathrm{Mean}(Y(\hat{x})) - \mathrm{Mean}(Y(x))\right|, \tag{14}$$

where $Y(\cdot)$ denotes the luminance channel. The color consistency error is defined as

$$\mathrm{ColorErr} = \left\|\mathrm{Mean}_{rgb}(\hat{x}) - \mathrm{Mean}_{rgb}(x)\right\|_1. \tag{15}$$

These two metrics are used as task-specific diagnostic measures for illumination naturalness and chromatic fidelity.

For no-reference perceptual evaluation, we report NIQE [32], MUSIQ [19], MANIQA [51], and CLIP-IQA [42]. NIQE is lower better, while MUSIQ, MANIQA, and CLIP-IQA are higher better.

F.3 Baseline Evaluation Protocol

We evaluate three groups of baselines. For general SR/restoration backbones, low-light inputs are directly fed into the corresponding models and evaluated at the target output resolution. For cascaded LLE+SR pipelines, we consider both enhancement-before-SR and SR-before-enhancement orders when applicable. Low-light enhancement models are trained or applied on the $1\times$ setting, while SR models are trained or evaluated under the corresponding $\times 2$ or $\times 4$ scale. For diffusion-guided or degradation-guided methods, we follow their official inference configurations when available and resize the outputs to match the evaluation resolution. All reported metrics are computed using the same test images and evaluation scripts.

F.4 Training and Runtime Settings

DeLang-SR is implemented in PyTorch. The model is initialized from a pretrained SD3 prior [8]. The SD3 text encoder and VAE decoder are frozen. The VAE encoder and SD3 flow transformer are adapted with LoRA modules, while G_ψ, $\mathbf{M}_\phi$, and the attribute prediction head are trained from scratch. We use $M = 4$ visual degradation tokens and a 7-dimensional degradation descriptor

$\boldsymbol{\delta}$. During inference, text embeddings can be cached once the degradation prompt ℓ_d is generated. DeLang-SR performs restoration with a single SD3 flow update, without iterative diffusion sampling.

We train the $\times 2$ and $\times 4$ models at resolutions of 1024×1024 and 864×864, respectively, using AdamW with a learning rate of 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-4}, and gradient clipping with a maximum norm of 1.0. The per-device batch size is 1 and gradient accumulation is set to 2, giving an effective batch size of 2 with one training GPU. Both the $\times 2$ and $\times 4$ models are trained for 20K optimization steps with mixed-precision training, and checkpoints are saved every 1K steps. We use LoRA ranks 96 and 128 for the SD3 VAE encoder and transformer, respectively, while the pretrained SD3 components are otherwise kept frozen. All methods are evaluated at the same output resolution. Runtime and peak memory are measured with batch size 1 after warm-up, excluding model loading, prompt-cache construction, image I/O, and metric computation. The checkpoint used for the main comparison is selected according to validation performance.

G Additional Experimental Results

G.1 Cross-Dataset Evaluation without Fine-Tuning

LOL-v2-Real [52] and LSRW [15] do not provide native SR protocols. We therefore follow the construction described in **Appx. F.1** and evaluate the RELLISUR-trained models without further optimization. Tables 8 and 9 report full-reference and no-reference results, respectively.

Table 8: Full-reference cross-dataset evaluation under the constructed $\times 2$ dark-SR protocol. All models are trained on RELLISUR and evaluated without fine-tuning.

Method	LOL-v2-Real						LSRW					
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	ExpErr↓	ColorErr↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	ExpErr↓	ColorErr↓
ASID	17.879	0.6683	0.4921	0.2721	0.0746	0.2264	17.293	0.5173	0.6048	0.2607	0.0613	0.1936
CATANet	17.488	0.6536	0.4567	0.2710	0.0861	0.2641	16.922	0.5127	0.5906	0.2650	0.0816	0.2580
HiT-SR	18.943	**0.6985**	0.4213	0.2863	0.0854	0.2667	16.174	0.5147	0.6061	0.2879	0.0908	0.2940
PFT-SR	18.927	0.6826	0.5298	0.2966	0.0844	0.2577	17.463	0.5098	0.6545	0.2833	0.0767	0.2378
DHGM	19.103	0.6189	0.4914	0.2789	0.0597	0.1866	17.383	0.5192	0.5036	0.2611	**0.0561**	**0.1805**
DeLang-SR	**19.512**	0.6973	**0.4060**	**0.2382**	**0.0550**	**0.1749**	**17.639**	**0.6579**	**0.4941**	**0.2137**	0.0611	0.1868

Table 9: No-reference perceptual evaluation under the same cross-dataset protocol.

Method	LOL-v2-Real			LSRW		
	MUSIQ↑	MANIQA↑	CLIP-IQA↑	MUSIQ↑	MANIQA↑	CLIP-IQA↑
ASID	38.078	0.2155	0.1925	38.064	0.2155	0.1684
CATANet	41.756	0.2088	0.1801	39.027	0.2008	0.1489
HiT-SR	32.061	0.2014	0.1458	33.438	0.2336	0.1508
PFT-SR	28.927	0.1770	0.1896	31.787	0.2285	0.1741
DHGM	43.575	0.2554	0.1919	43.127	0.2247	0.1667
DeLang-SR	**55.826**	**0.2976**	**0.3755**	**54.865**	**0.2846**	**0.3434**

On LOL-v2-Real, DeLang-SR obtains the best PSNR, LPIPS, DISTS, exposure error, and color error among the evaluated methods, while its SSIM is within 0.0012 of the best result. It also achieves the highest scores on all three no-reference metrics. On LSRW, DeLang-SR performs best in PSNR, SSIM, LPIPS, DISTS, MUSIQ, MANIQA, and CLIP-IQA, whereas DHGM yields lower exposure and color errors. These results indicate that the fidelity–perception improvements transfer across the two constructed dark-SR protocols, while the illumination-related trade-off remains dataset dependent.

G.2 Update-Strength Bounds and Spatial Variation

We vary the output bounds at inference time using the same 5K checkpoint, without re-optimizing the model. Across the tested intervals in Table 10, PSNR varies by 0.106 dB, SSIM by 0.0013, and LPIPS by 0.0069. The default interval $[0.15, 0.95]$ gives the best PSNR and LPIPS, whereas alternative intervals yield small improvements in exposure or color error.

Table 10: Inference-time sensitivity to $(\sigma_{\min}, \sigma_{\max})$ on RELLISUR $\times 2$ at 5K steps.

Bounds	PSNR↑	SSIM↑	LPIPS↓	ExpErr↓	ColorErr↓
$[0.10, 0.90]$	22.910	0.7588	0.2536	0.0374	0.1148
$[0.15, 0.85]$	22.880	0.7596	0.2551	0.0341	0.1065
$[0.15, 0.95]$ (default)	**22.986**	0.7594	**0.2482**	0.0360	0.1120
$[0.20, 0.90]$	22.930	**0.7601**	0.2518	0.0349	0.1082

To isolate spatial variation, we use the same 5K checkpoint at inference time, without retraining, and replace each predicted field with its image-wise mean before broadcasting the resulting scalar to all latent locations. This control preserves the average update magnitude for each image while removing its spatial allocation. As shown in Table 11, the spatial field improves PSNR by 0.243 dB and SSIM by 0.0053, and reduces LPIPS by 0.0125. ExpErr and ColorErr change by only 0.0002 and 0.0015, respectively, indicating that the main benefit arises from spatially allocating the update rather than changing its global mean.

Table 11: Ablation of spatial variation in the update-strength field on RELLISUR $\times 2$ at 5K steps.

Update strength	PSNR↑	SSIM↑	LPIPS↓	ExpErr↓	ColorErr↓
Uniform image-wise mean	22.743	0.7541	0.2607	**0.0358**	**0.1105**
Spatial field	**22.986**	**0.7594**	**0.2482**	0.0360	0.1120

G.3 Additional Update-Strength Field Visualizations

We provide additional visualizations of the degradation-aware update-strength field $\mathbf{M}_\phi(y, \boldsymbol{\delta})$ in Fig. 9. For each case, the LQ-only restoration intent ℓ_d and continuous descriptor $\boldsymbol{\delta}$ are estimated from the input y without using the paired GT. The predicted update-strength fields are spatially non-uniform and provide interpretable local step-size maps for the one-step SD3 flow restoration. These examples further illustrate that DeLang-SR does not apply a uniform global update, but adapts the local update magnitude according to degradation severity and visual evidence.

G.4 Inference-Time Condition Source Analysis

Table 12: Inference-time condition source analysis on ***RELLISUR*** $\times 2$. Comparable results across generic, LQ-only, and oracle intents show that DeLang-SR does not rely on oracle degradation annotations.

Condition Source	**Uses GT?**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**ExpErr↓**	**ColorErr↓**
Generic prompt	**No**	22.984	0.7593	0.2487	0.0361	0.1120
LQ-only degradation intent	**No**	22.985	0.7594	0.2482	0.0361	0.1120
Oracle degradation intent	**Yes**	22.985	0.7593	0.2479	0.0360	0.1119

We further report the numerical results of the inference-time condition source analysis in Table 12. The same full checkpoint is evaluated with three condition sources: a fixed generic prompt, the proposed LQ-only degradation intent, and an oracle degradation intent constructed from paired data. The comparable performance across these settings shows that DeLang-SR does not rely on oracle degradation annotations during deployment, and that the generated degradation language should be interpreted as a coarse restoration intent rather than a precise label. The generic prompt differs from the LQ-only intent by only 0.001 dB PSNR and 0.0005 LPIPS, indicating that detailed sentence composition has a limited effect on these aggregate metrics. This comparison changes only the language condition; all variants retain the continuous descriptor $\boldsymbol{\delta}$ and the visual-token pathway, so it does not remove the complete input-statistics block.

G.5 Component Ablation Visualization

We provide an additional ring-chart visualization of selected component ablations in Fig. 10. This figure complements the quantitative ablation table in the main paper by comparing the full DeLang-SR

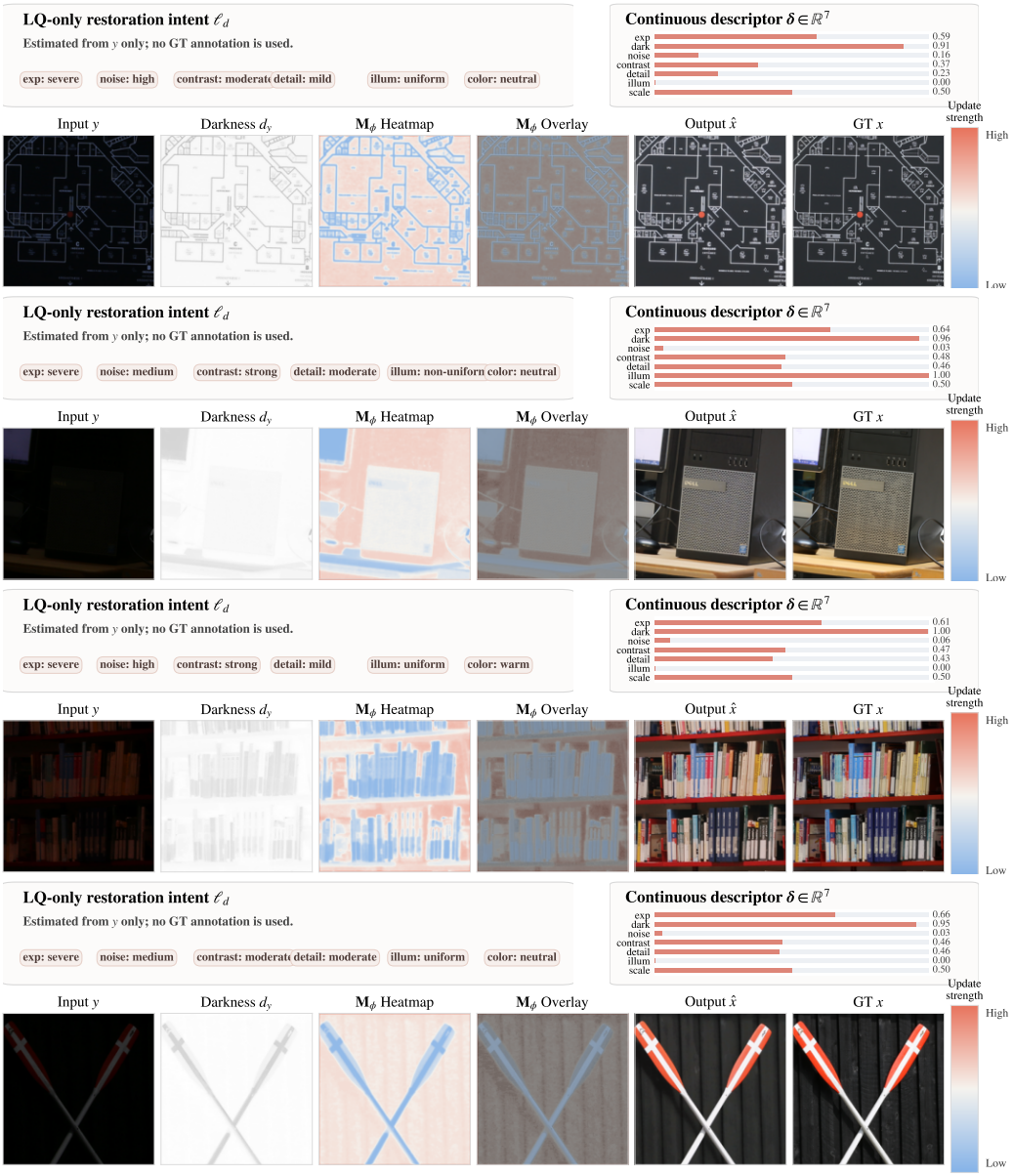

Figure 9: Additional visualizations of the update-strength field $\mathbf{M}_\phi(y, \boldsymbol{\delta})$. For each example, ℓ_d and $\boldsymbol{\delta}$ are estimated from the input y without GT annotations. The predicted maps provide spatially non-uniform step-size fields for one-step SD3 flow restoration.

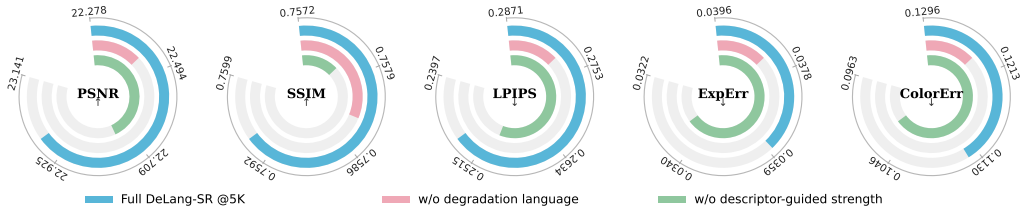

Figure 10: Additional ring-chart visualization of selected component ablations on ***RELLISUR*** $\times 2$. The chart compares Full DeLang-SR, w/o degradation language, and w/o descriptor conditioning in $\mathbf{M}_\phi$ across PSNR, SSIM, LPIPS, ExpErr, and ColorErr. The full model achieves a better overall trade-off, supporting the complementarity of direction steering and magnitude control.

model with two representative variants: removing degradation language and removing descriptor conditioning from the update-strength predictor. Across fidelity, perceptual, exposure, and color-related metrics, the full model generally provides a better overall trade-off. This supports the complementarity between degradation-language direction steering and descriptor-guided magnitude control.

G.6 Additional Visual Results on RELLISUR

We provide additional qualitative comparisons on RELLISUR under both $\times 2$ and $\times 4$ settings in Figs. 11 and 12. For each example, we show the restored full image with the selected ROI marked in green and the corresponding enlarged crop. The $\times 2$ results examine local structure and contrast restoration under dark degradation, while the $\times 4$ results further test fine-detail recovery from more limited spatial observations. Across both scaling factors, DeLang-SR preserves clearer local structures, repetitive textures, and natural contrast, whereas competing methods often suffer from over-smoothing, color shifts, or unstable texture reconstruction under severe darkness.

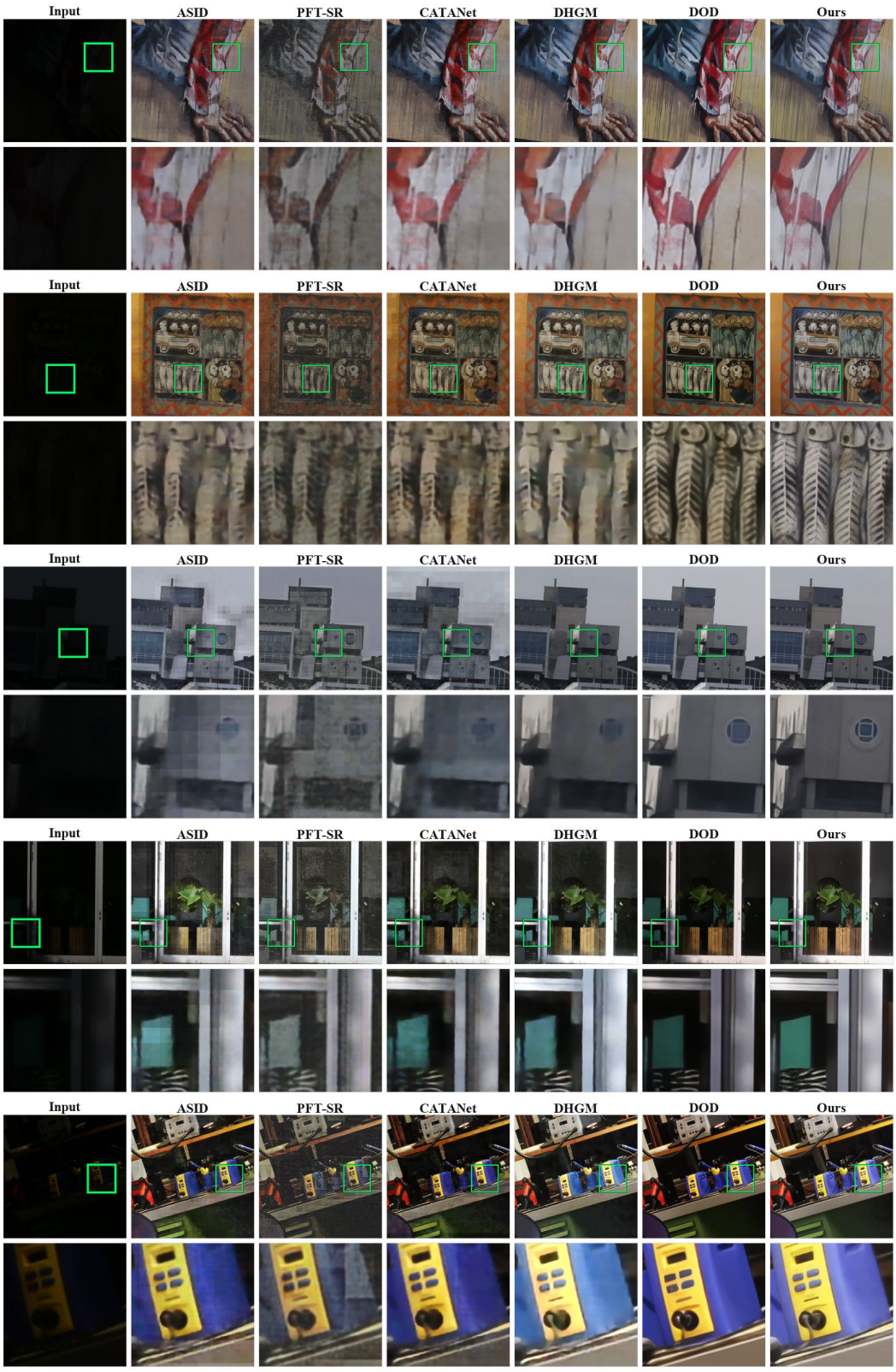

Figure 11: Additional visual comparison on ***RELLISUR*** $\times 2$. Each row shows the restored full image with the selected ROI marked in green and the corresponding enlarged crop. DeLang-SR preserves clearer local structures and more faithful contrast under challenging dark $\times 2$ super-resolution.

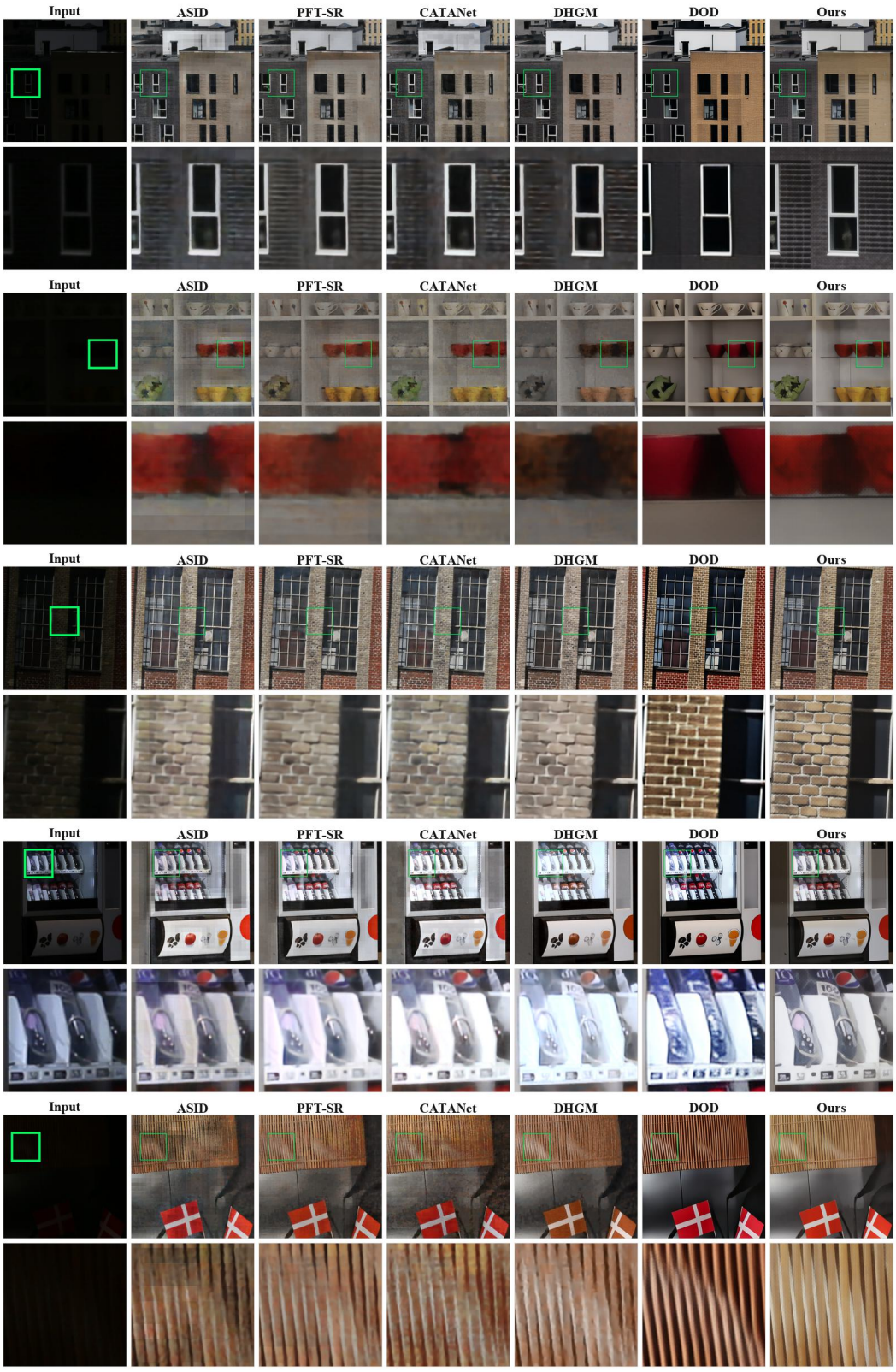

Figure 12: Additional visual comparison on ***RELLISUR*** $\times 4$. Each row shows the full restored image and the enlarged ROI crop. DeLang-SR preserves clearer local structures and more faithful contrast under challenging dark $\times 4$ super-resolution.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the scope of DeLang-SR as direction–magnitude decoupled one-step flow restoration for dark image super-resolution. These claims are supported by the method in Section 3 and the evaluations in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations related to extremely severe dark degradation and LQ-only degradation estimation. These points are summarized in the conclusion, together with possible future directions for uncertainty-aware and confidence-calibrated control.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include formal theoretical results, theorems, or proofs. The mathematical formulations are used to define the proposed one-step flow update and training objectives.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the datasets, evaluation metrics, baselines, model components, and implementation protocol needed to reproduce the main results. Additional implementation details, including degradation construction and module architectures, are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper uses publicly available datasets, and we provide code in the supplemental material to reproduce the main results. The instructions include data preparation, model training, inference, and evaluation protocols.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the dataset splits, evaluation metrics, baseline protocols, model components, and training/inference settings in Section 4 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports quantitative comparisons, per-image score distributions, ablation studies, and checkpoint sensitivity analyses, but does not include formal error bars, confidence intervals, or statistical significance tests. We omit them mainly due to the high computational cost of repeatedly training SD3-based restoration models.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments were conducted mainly on NVIDIA RTX 4090 and RTX A6000 GPUs, and the paper reports inference runtime and memory usage in the efficiency analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The work focuses on image restoration and super-resolution using publicly available datasets, and does not involve human-subject studies, private data collection, or high-risk decision-making applications.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work may benefit low-light photography and visual restoration by improving visibility and perceptual quality in dark scenes. Potential negative impacts include hallucinated details or misuse in surveillance/forensic contexts, so the method should not be used as evidence in safety-critical or identity-sensitive decisions without human verification.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a scraped dataset or a general-purpose generative model with high misuse risk. The proposed model is designed for low-light image restoration and super-resolution on publicly available benchmarks, and potential misuse concerns are discussed in the broader impact checklist item.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources of the datasets, pretrained model, and baseline methods used in the experiments, including RELLISUR, DarkFace, SD3, and prior restoration/SR models. Existing assets are used for research purposes following their released terms where available, and the paper itself is submitted under the selected CC BY 4.0 open-access license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release new datasets, benchmarks, or model assets at submission time. The work uses existing public datasets and pretrained models, and the proposed method is documented through the method and implementation sections.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing public image restoration benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. Therefore, IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method, experiments, and theoretical analysis do not use LLMs as an important, original, or non-standard component. LLMs were used only for language editing and formatting assistance, which did not affect the scientific methodology, experimental results, or theoretical claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.