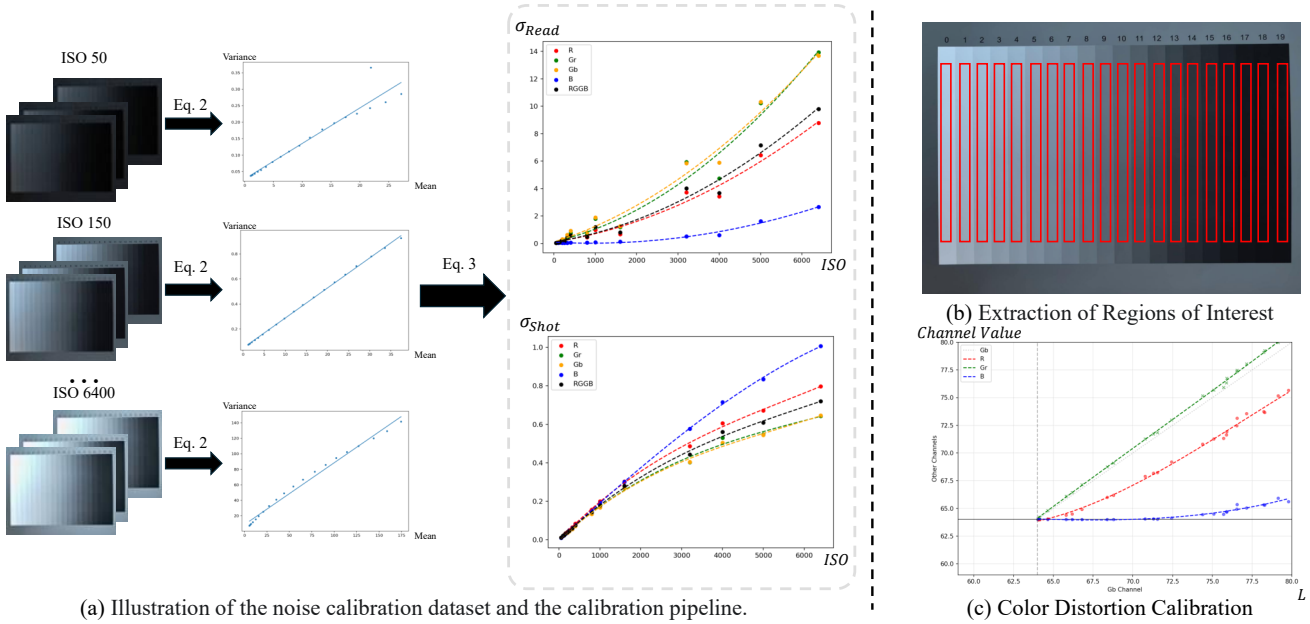
RawMetaDiff: Unlocking Extreme Darkness from Dual-Exposure RAW with Meta-Guided Diffusion ——Supplementary Material——

Panjun Liu[1] Jiyuan Xia[1] Yuanshen Guan[1] Yong Li[2] Zhiqiang Lang[2]
Ruikang Xu[2] Chang Chen[2] Dehua Song[2] Fenglong Song[2] Zhiwei Xiong[1]
[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab
{panjun_liu,jyxia,guanys}@mail.ustc.edu.cn,zwxiong@ustc.edu.cn
{liyong156,langzhiqiang,xuruikang1,chenchang25,dehua.song,songfenglong}@huawei.com

The supplementary material is organized as follows:

- Section 1 details the calibration dataset and presents the corresponding results.
- Section 2 provides an ablation analysis on traditional methods, revealing that single-frame input actually leads to better performance for such approaches.
- Section 3 presents additional qualitative results on both synthetic and real-world datasets.
- Section 4 provides the experimental settings and detailed training configurations.
- Section 5 provides further discussions and additional evaluations, including robustness to extreme motion, metadata specifics, scaling laws, and computational efficiency.



(a) Illustration of the noise calibration dataset and the calibration pipeline.

(b) Extraction of Regions of Interest

(c) Color Distortion Calibration

Figure 1. Overview of the sensor noise and color distortion calibration framework. **(a)** Illustration of the noise calibration pipeline. We capture raw sequences of a grayscale chart at a fixed exposure time (1/100s) across varying ISO settings. For each ISO, we estimate the corresponding shot noise and read noise parameters by fitting the mean-variance relationship (Eq. 2). We then model these parameters as functions of ISO (Eq. 3) to enable continuous noise synthesis. **(b)** Visualization of the Regions of Interest (ROIs) extracted from the grayscale step chart for statistical analysis. **(c)** Calibration of color distortion, which models the non-linear intensity response of each channel with respect to luminance (L).

1. Calibration of Sensor Noise and Color Distortion

To bridge the domain gap between synthetic data and real-world captures, we construct a physics-grounded degradation model. Instead of applying generic Gaussian noise, we mathematically calibrate the sensor-specific characteristics from real raw data. Our calibration framework primarily focuses on two aspects: **Sensor Noise** and **Color Distortion**.

1.1. Sensor Noise Calibration

We model camera noise using a heteroscedastic Gaussian model, which serves as an effective approximation of the physical Poisson-Gaussian process, as shown in Fig. 1. This model accounts for both signal-dependent shot noise and signal-independent read noise. Specifically, a noisy pixel observation y' is formulated as:

$$y' = y + n, \quad \text{where} \quad n \sim \mathcal{N}(0, \sigma_r^2 + y \cdot \sigma_s), \tag{1}$$

where y represents the underlying clean linear intensity. The parameters σ_s and σ_r^2 represent the shot noise factor and read noise variance, respectively.

Data Acquisition and Parameter Estimation. To calibrate these parameters, we collected raw sequences of a standard grayscale step chart under uniform lighting across varying ISO settings. For a specific ISO, let $\mathcal{D} = \{y'^{(t)}\}_{t=1}^{T}$ be the captured temporal sequence of T frames. We utilize M distinct gray levels on the chart to extract Regions of Interest (ROIs). For the i-th gray level ROI, denoted as Ω_i, we compute the spatially averaged temporal mean μ_i and variance v_i to ensure statistical robustness:

$$\mu_i = \frac{1}{|\Omega_i|T} \sum_{t=1}^{T} \sum_{p \in \Omega_i} y_p'^{(t)}, \quad v_i = \frac{1}{|\Omega_i|} \sum_{p \in \Omega_i} \text{Var}_t(y_p'), \tag{2}$$

where $\text{Var}_t(\cdot)$ denotes the temporal variance of a single pixel p across the sequence. This process yields a set of reliable observation pairs $\{(\mu_i, v_i)\}_{i=1}^{M}$ covering the sensor's dynamic range. According to the noise model in Eq. (1), the variance is linearly related to the intensity ($v \approx \sigma_s \cdot \mu + \sigma_r^2$). Therefore, we estimate the optimal parameters for each channel by solving the following linear least-squares optimization:

$$\hat{\sigma}_s, \hat{\sigma}_r = \arg\min_{\sigma_s, \sigma_r} \sum_{i=1}^{M} \left\| v_i - (\sigma_s \cdot \mu_i + \sigma_r^2) \right\|^2. \tag{3}$$

Since these parameters are determined by the camera's ISO setting, we further fit polynomials to the calibrated points $(\text{ISO}, \hat{\sigma}_s)$ and $(\text{ISO}, \hat{\sigma}_r^2)$, enabling continuous noise synthesis during training.

1.2. Color Distortion Calibration

As discussed in the main paper, low-light conditions often introduce non-linear response shifts due to sensor-specific characteristics. For clarity, we restate the distortion model here. The relationship between the original linear RAW value C_k and the distorted value C_k' for each channel $k \in \{R, Gr, Gb, B\}$ is formulated as:

$$C_k' = f_k(L) \cdot C_k, \tag{4}$$

where $f_k(L)$ represents the sensitivity-dependent scaling factor.

Here, we provide specific details on the physical basis and implementation of this model. The observed non-linear response is primarily attributed to signal truncation near the black level. As the signal amplitude approaches the noise floor, the intensity distribution is asymmetrically clipped, causing the mean response to deviate from the ideal linearity. To parameterize this effect, as shown in Fig. 1, we approximate the local luminance L using the intensity of the Gb channel (i.e., $L \approx C_{Gb}$), which acts as the reference baseline. Consequently, the function $f_k(L)$ effectively models the response ratio of channel k relative to Gb. We derive these functions by applying a cubic spline fit to empirical calibration data, ensuring that the synthetic data faithfully reproduces the truncation-induced curvature and color biases observed in the real sensor.

2. Ablation on Single-Frame Baseline

Conventional methods, such as Restormer, lack specific designs for dual-exposure fusion. Consequently, their performance on dual-frame inputs is inferior to that on single-frame inputs, as demonstrated in Table 1. Notably, perceptual metrics like MUSIQ and LPIPS exhibit high sensitivity to noise but are relatively insensitive to blur degradation. As a result, these methods yield scores slightly higher than the raw short-exposure input (we adopted). This observation is further corroborated by the visual comparisons in Figure 2.

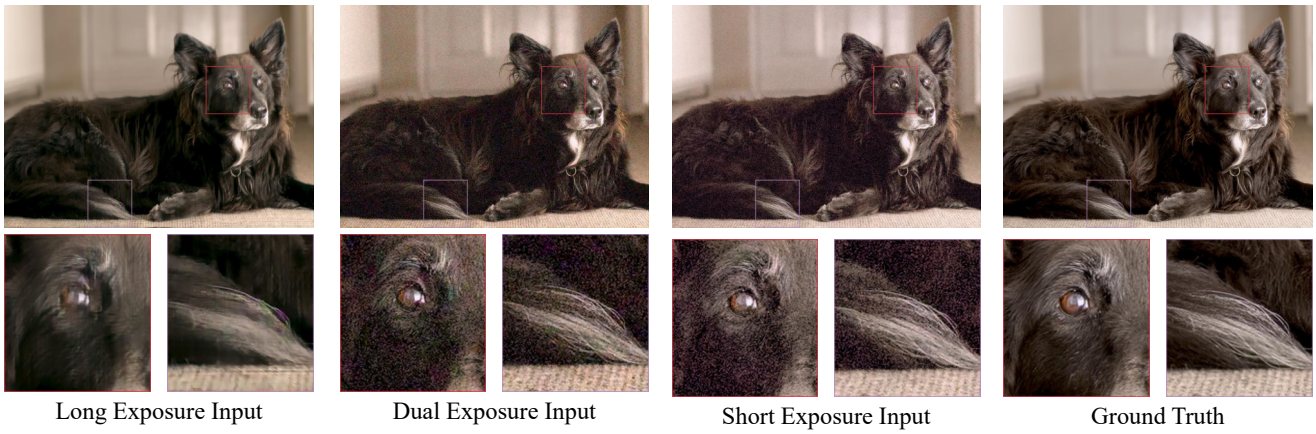


Figure 2. Ablation study on the input modalities for Restormer. We compare the results using different input configurations. As observed, the short-exposure input yields relatively superior performance, as it avoids the severe motion blur present in the long-exposure frame and the misalignment artifacts in the naive dual-exposure setting.

Table 1. Ablation study on the input modalities for Restormer. We compare the results using different input configurations. As observed, the short-exposure input yields relatively superior performance, as it avoids the severe motion blur present in the long-exposure frame and the misalignment artifacts in the dual-exposure setting. Consequently, we employ the short-exposure input for the single-frame baselines.

Methods	PSNR↑	MS-SSIM↑	LPIPS↓	Δ_E↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑	DeQA↑
Long Exposure Input	21.81	0.8569	0.3448	7.201	6.791	62.29	0.2644	0.3178	3.135
Dual Exposure Input	21.61	0.8458	0.4883	7.820	6.691	51.07	0.3277	0.6203	3.219
Short Exposure Input	23.69	0.8764	0.4108	6.207	6.519	55.57	0.3570	0.6802	3.353

3. Additional Qualitative Results

In Figure 3, we provide additional visualization results on the synthetic dataset. In Figure 4, we provide additional visualization results on the real-world dataset.

4. Training Hyperparameters and Details

Training Strategy. We employ a two-stage training strategy to optimize our framework.

In the first stage, we adaptively fine-tune the VAE using clean linear raw images $I \in \mathbb{R}^{3\times H\times W}$. The model is trained on 512×512 patches with a learning rate of 1×10^{-4}. The objective function comprises an adversarial loss ($\mathcal{L}_{GAN}$), a reconstruction loss ($\mathcal{L}_2$), and a KL divergence loss ($\mathcal{L}_{KL}$), all of which are computed in the linear raw domain. To align the restoration quality with human perception, we also calculate the perceptual loss ($\mathcal{L}_{LPIPS}$) in the sRGB domain via a simple differentiable ISP.

In the second stage, we freeze the decoder weights and focus on optimizing the remaining components. While maintaining the patch size at 512×512, we reduce the learning rate to 1×10^{-5}. During this phase, we enforce stronger perceptual and chromatic constraints in the sRGB domain, incorporating $\mathcal{L}_{LPIPS}$, an additional RGB-based $\mathcal{L}_2$ loss, and a color consistency loss ($\mathcal{L}_{color}$) calculated as the Euclidean distance in the CIELAB space. These are optimized alongside the standard raw-domain losses ($\mathcal{L}_2$, $\mathcal{L}_{GAN}$).

5. Additional Discussions and Evaluations

In this section, we provide further clarifications and additional experimental results to comprehensively evaluate our proposed method.

5.1. Robustness to Extreme Motion

We evaluated the extreme motion scenario. Although minor hallucinations occur, our method constrains outputs within the natural image manifold. Consequently, it sidesteps both the over-smoothing typical of regression baselines and artifacts caused by misalignment (see Figure 5).

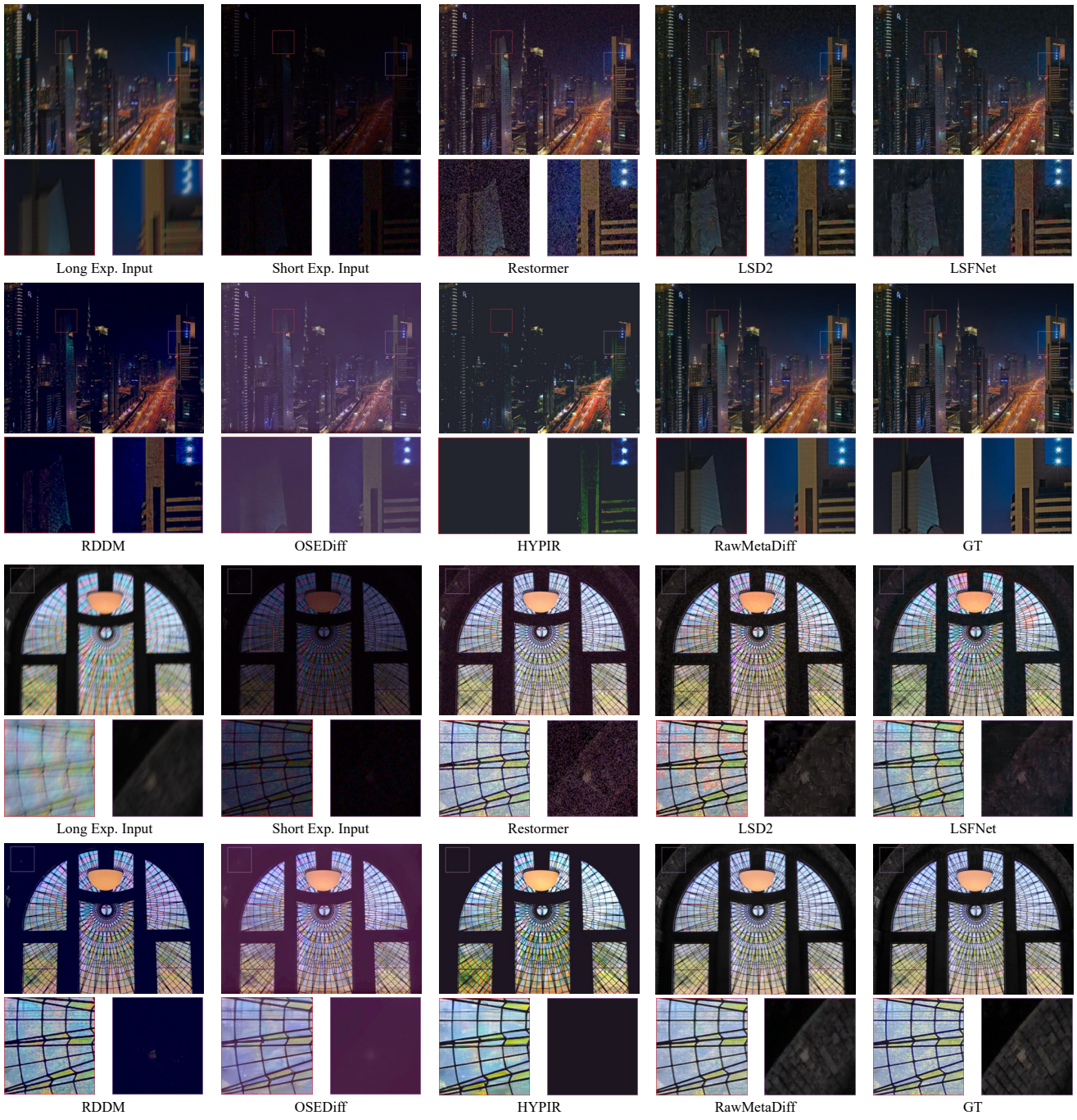


Figure 3. Additional visualization results on the synthetic dataset.

5.2. Metadata Specifics and Ablation

Specifically, we incorporate AWB, ISO, Exposure Time (Exp), and CCM as our metadata. We apply **logarithmic** normalization to ISO and Exp to effectively compress their wide dynamic ranges, while adopting a linear normalization for CCM and AWB to preserve crucial **negative values**. These features are concatenated and then mapped to P_m via an MLP with LayerNorm, projecting the dimension from 16 to 128 channels.

To clarify the contributions of different metadata, we present representative combinations in Table 2. ISO and Exp focus on **detail injection**, while AWB and CCM handle **color transfer**.

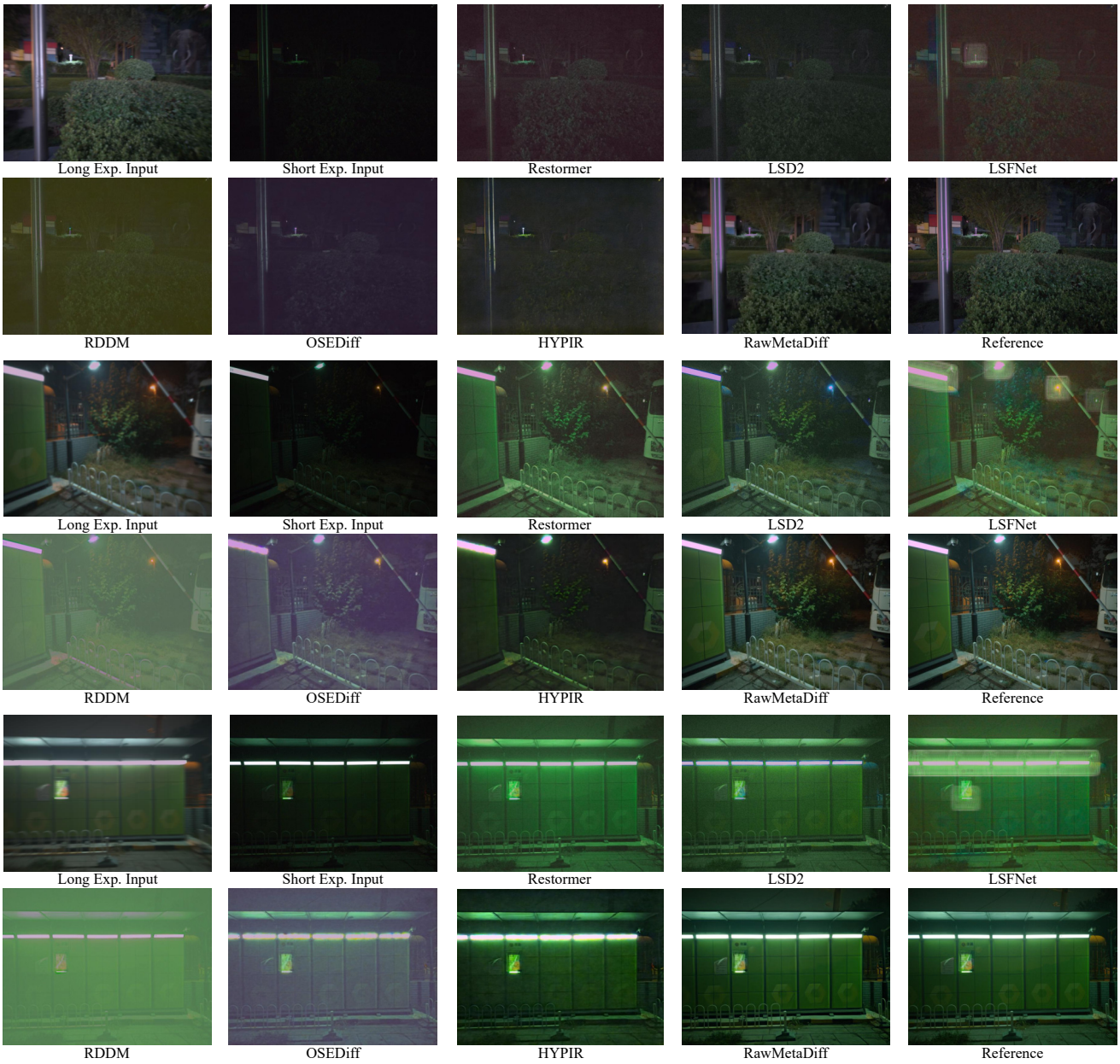


Figure 4. Additional visualization results on the real-world dataset.

Figure 5. Visual comparison under extreme motion scenarios.

Table 2. Additional ablation of metadata combinations.

Methods	PSNR↑	$\Delta E\downarrow$	MUSIQ↑	DeQA↑	CLIP-IQA↑
None	22.43	6.9561	70.58	4.081	0.6745
ISO + Exp	23.13	6.4526	71.24	4.113	0.6934
AWB + CCM	22.97	6.1385	70.84	4.098	0.6847
Ours	23.59	5.5077	71.45	4.115	0.7008

Table 3. Comparisons on real RAW datasets.

Methods	PSNR↑	LPIPS↓	$\Delta E\downarrow$	MUSIQ↑	DeQA↑	CLIP-IQA↑
Restormer	**28.92**	0.4425	6.215	52.72	3.194	0.5243
LSD2	27.43	0.3324	7.143	54.13	3.117	0.4717
HYPIR	26.57	0.3122	7.430	54.23	3.319	0.4627
Ours	28.75	**0.2303**	**5.382**	**58.43**	**3.417**	**0.5723**

5.3. Evaluation on Real-World RAW Data

Evaluating paired metrics on public RAW datasets is currently limited because a proper dataset requires three components: noisy low-exposure, blurry high-exposure, and reference images, which are currently unavailable. Furthermore, paired metrics are unsuitable for our DERaw dataset. In real-world scenes, spatial misalignments are often unavoidable and can compromise the validity of PSNR/SSIM. Therefore, we prioritize no-reference metrics to provide a more objective and faithful reflection of perceptual quality.

Nevertheless, we provide comparisons on real RAW data in Table 3. Our model delivers superior perceptual quality compared to the baselines. Meanwhile, its PSNR performance remains comparable with Restormer.

5.4. Training Data and Scaling Law

We use simulated raw data for training. The core novelty of our pipeline is the usage of real-world data to calibrate simulation parameters for enhanced fidelity. While calibration is fundamental, we have integrated several existing degradation models to build a robust framework.

Performance relies on both data quality and scale. While Real RAW data outperforms simulated data at the same size (5k) under our degradation pipeline, its scarcity presents a bottleneck. In contrast, larger-scale simulated datasets yield superior performance. As shown in Figure 6, performance steadily improves with dataset size, following a clear scaling law.

5.5. Computational Overhead

As shown in Table 4, our model outperforms generative baselines at comparable latency. Regarding regression methods, we match the state-of-the-art Restormer in speed while delivering superior performance. Notably, our latency is on par with Restormer despite higher nominal FLOPs. This is because Restormer incurs heavy I/O overhead from high-resolution attention. In contrast, our backbone operates in a compact latent space with cache-friendly VAE. This design also results in excellent VRAM efficiency.

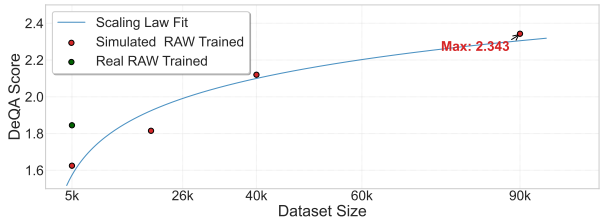


Figure 6. Performance scaling with training dataset size.

Table 4. Computational overhead comparison.

Methods	LSD2	LSFNet	Restormer	RDDM	OSEDiff	HYPIR	Ours
CLIP-IQA	0.6069	0.5887	0.6787	0.6163	0.5939	0.5883	0.7008
Latency / ms	38	46	900	4750	894	894	898
VRAM / GB	2.5	1.6	9.2	9.5	7.8	7.8	7.9
FLOPs / T	1.6	1.1	4.5	30.8	21.1	21.1	21.3