NEC-Diff: Noise-Robust Event–RAW Complementary Diffusion for Seeing Motion in Extreme Darkness

Supplementary Material

Summary

The supplementary material is organized as follows.

- Section 1 introduces the implementation details of the NEC-Diff framework and REAL dataset.
- Section 2 discusses more ablation studies of NEC-Diff.
- Section 3 shows more visual results on different datasets.

1. Implementation Details

1.1. Event Representation

The EV-Denoise network outputs a denoised event stream $\hat{E}(t) = \{e_i\}_{i=0}^{N-1}$. Before being fed into the event encoder, the events are converted into a voxel grid [1], represented as a tensor $V \in \mathbb{R}^{B\times H\times W}$ with B temporal bins, defined as

$$V(k) = \sum_i p_i \max(0, 1 - \left|k - \tfrac{t_i - t_0}{t_N - t_0}(B-1)\right|) \quad (1)$$

where N is the number of events, p_i and t_i represent the polarity and timestamp of the i-th event respectively, and $k \in [0, B-1]$. This method uniformly fills events into the two nearest bins using the fixed interpolation approach.

1.2. Diffusion Model

We adopt a diffusion-based conditional generative framework guided by the low-light raw image $\mathbf{y}$ and the corresponding low-light event frames $\mathbf{E}$. A multimodal fusion module produces the conditioning vector

$$\mathbf{c} = F(\mathbf{y}, \mathbf{E}), \quad (2)$$

which is supplied to the denoiser throughout the diffusion chain. The overall training and sampling procedures are summarized in Algorithm 1 and Algorithm 2, respectively.

Forward Diffusion. The forward process gradually perturbs a clean latent $\mathbf{x}_0$ according to a predefined noise schedule (see Algorithm 1). At each step, noise is injected through

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\,\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}\big), \quad (3)$$

where $\alpha_t = 1 - \beta_t$. During training, noisy states are drawn using the reparameterized form

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (4)$$

The denoiser predicts the noise residual $\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)$, and the diffusion branch is optimized using the noise prediction objective $\mathcal{L}_{\text{diff}} = \|\boldsymbol{\epsilon} - \hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)\|_2$.

Algorithm 1 Training

$\mathbf{c} = F(\mathbf{y}, \mathbf{E})$
while not converged **do**
  $t \sim \text{Uniform}(\{1, \dots, T\})$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\hat{\epsilon} = \epsilon_\theta(\sqrt{\bar{\alpha}_t}\,\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \mathbf{c}, t)$, $\mathcal{L}_{\text{diff}} = \|\epsilon - \hat{\epsilon}\|^2$
  if Joint Learning **then**
    Gradient descent on $\nabla_\theta(\mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{others}})$
  else
    Gradient descent on $\nabla_\theta \mathcal{L}_{\text{diff}}$
  end if
end while
return θ

Algorithm 2 Sampling

$\mathbf{c} = F(\mathbf{y}, \mathbf{E})$
$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

Reverse Denoising. During inference, we adopt deterministic DDIM sampling without additional noise injection. The reverse process recovers clean samples from noise according to the learned posterior, as detailed in Algorithm 2:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) = \mathcal{N}\big(\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t),\ \sigma_t^2 \mathbf{I}\big), \quad (5)$$

The mean is constructed from the predicted noise as

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t) = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{c}, t)\Big), \quad (6)$$

and the variance is given by

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\,\beta_t. \quad (7)$$

At inference, samples are recovered by iterating

$$\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t) + \sigma_t\,\mathbf{z}, \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \ \ (t > 1), \quad (8)$$

eventually yielding the clean estimate $\mathbf{x}_0$. The conditioning vector $\mathbf{c}$ is injected at each denoising step, providing multimodal guidance that helps restore scene-consistent structure and fine details.

1.3. Training and Test Details

Training Details. The proposed NEC-Diff framework adopts a two-stage training strategy. In the first stage, the image and event noise suppression modules are pre-trained independently. Specifically, the image denoising module is pre-trained on the ELD [2] dataset, while the event denoising module is pre-trained on the ED24 [3] dataset. This stage aims to reduce the training burden for the subsequent DDIM stage. In the second stage, the modules are jointly trained using the LLRVD-simu and REAL datasets, following the pre-training of each module in the first stage.

Test Details. We evaluate all methods on both the simulated and real datasets, namely the simulated LLRVD-simu and the real captured REAL dataset. LLRVD [4] is a low-light RAW video enhancement dataset, which also provides the camera's ISP parameters. This dataset is particularly suitable for our study because it provides paired RAW videos and ISP information, allowing both RAW-based and sRGB-based methods to be fairly evaluated under low-light conditions. Moreover, the dataset consists of continuous motion sequences, which enables the simulation of events using V2E [5] for event-based and hybrid methods. To this end, we generated LLRVD-simu based on LLRVD, where events were simulated with added noise to mimic the response characteristics of event sensors under low-light conditions. The REAL dataset was captured with low-light RAW videos, low-light events, and normal-light ground-truth frames, and also provides the ISP parameters corresponding to the low-light RAW inputs for RAW-to-sRGB conversion.

For both datasets, RAW-based methods directly process the low-light RAW inputs for denoising or enhancement, and the processed RAW images are converted to sRGB using the provided ISP parameters for quantitative evaluation. sRGB-based methods first convert the RAW videos to sRGB using the ISP parameters and then perform enhancement or denoising. For event-based or Event-sRGB methods, event inputs come from the simulated events in LLRVD-simu or the real low-light events in REAL. Event-sRGB methods use event information while their image branch follows the sRGB processing procedure to ensure consistency of input conditions. Overall, the testing procedure on REAL follows the same protocol as LLRVD-simu, except that the event inputs are obtained from real captures. To ensure fairness, all methods are fine-tuned on training data processed in the same way as the testing data.

1.4. REAL Dataset

Hardware Setup. We built a co-axial imaging system that integrates two frame cameras (FLIR BFS-U3-16S2) and one event camera (Prophesee EVK4). Light first passes through a beam splitter (Thorlabs BSW26R). One path records normally exposed reference frames, while the other is further split and directed to the event camera and a second frame camera for low-light acquisition. High-density ND filters (Thorlabs ND20A) are mounted on the low-light paths to ensure stable and physically consistent dark illumination. All sensors are synchronized using external hardware triggering, achieving microsecond-level temporal alignment across the different modalities. The event camera uses the default threshold settings provided by Prophesee for event detection.

Alignment. To achieve pixel-level spatial alignment across sensors, we first capture a standard checkerboard pattern with all cameras. The asynchronous events are then reconstructed into temporally aligned video frames using E2VID [10]. Using these reconstructed event frames together with the standard RAW images, we perform stereo calibration to estimate intrinsic and extrinsic parameters, followed by rectification of all cameras. Finally, considering the field of view of the event camera and the two frame-based cameras, the resolution of the captured images is adjusted to 816×672.

Downstream Task Annotations. To further support research in event-based vision and low-light enhancement, we enrich the REAL dataset by providing annotations for two downstream tasks: object detection and semantic segmentation. For object detection, we adopt the annotation scheme of the DSEC-Detection dataset [11], which includes the following object categories: pedestrian, rider, car, bus, truck, bicycle, motorcycle, and train. For semantic segmentation, we follow the guidelines of the DSEC-Semantic dataset [12], labeling 19 classes: road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, truck, bus, train, motorcycle, and bicycle. Example annotations for these tasks are shown in Fig. 1.

Dataset Comparison. A summary of existing low-light event-image datasets can be found in Tab. 1. LIE[6] is the first real-world event-frame low-light enhancement dataset, which only uses a single DAVIS 346 Color event camera to capture paired data by changing the brightness or adjusting the exposure time. SDE[7] collects data in daylight using a DAVIS 346 camera, and obtains paired data by attaching/detaching an ND8 filter for two separate shootings. EvLowLight[8] only contains low-light images and events, without providing normal-light ground truth for reference. RELED[9] is similar to our hardware system, but it cannot provide low-light data in RAW format. In contrast, our dataset can provide 47K paired high-resolution low-light events, low-light RAW images, and normal-light ground truth. To promote progress in this domain, we will addi-

Table 1. Comparison with existing hybrid datasets

Dataset	Data Type	Release	Image Resolution	Numbers
LIE[6]	Event + sRGB	✓	256×256	2,231
SDE[7]	Event + sRGB	✓	256×256	31,477
EvLowLight[8]	Event + sRGB	×	–	–
RELED[9]	Event + sRGB	✓	1024×768	6,258
REAL	Event + RAW + sRGB	✓	816×672	47,834

Figure 1. Visualization examples from the REAL dataset. From left to right: linearly brightened RAW images, event frames, overlays of events and ground truth (GT), object detection annotations, and semantic segmentation annotations.

tionally open-source the low-light sRGB images obtained by performing ISP processing on the low-light RAW data, facilitating related research.

2. Ablation Study and Discussion

2.1. Why use Gaussian Blur for Preprocessing?

Since RAW images captured under extremely low light still contain noticeable noise, we apply a simple Gaussian blur to extract a coarse illumination prior. Although more advanced RAW denoising methods could provide cleaner priors, they introduce additional computational overhead. We further evaluate the impact of different RAW-denoising strategies on the final reconstruction quality, comparing LED [13], RID [14], and BRVE [15]. All methods are fine-tuned and evaluated on the REAL dataset for a fair comparison. The results are summarized in Tab. 2. It can be seen that even simple Gaussian smoothing provides a clear performance boost, while more sophisticated denoising methods offer no further gains. This is because the illumination prior serves only as a coarse guidance cue, whereas the final reconstruction quality is largely dictated by the consistency constraint.

2.2. Trailing Events at Night

Due to the temporal trail effect of events under extremely low-light conditions [16], reconstructed images may exhibit edge artifacts. To mitigate this issue, we explore two strategies. (1) Event Trail Suppression (ETS). We preprocess event streams using ETS [16] and use the corrected data for both training and testing. (2) Long event-window training. We extend the input event window to 40 ms centered at the image exposure time. This ensures that the intensity-consistency loss remains unaffected, as the accumulated event count—which encodes brightness change—remains valid as long as delayed trail events fall within the window. Trail suppression is then implicitly learned through Learnable Event Timestamp Calibration (LETC) [16] and

Table 2. Quantitative results of different denoising methods.

Methods	PSNR↑	SSIM↑	LPIPS↓
–	23.95	0.719	0.227
Gaussian Blur	24.51	0.742	0.201
LED [13]	24.50	0.743	0.199
BRVE [15]	24.52	0.742	0.201
RID [14]	24.53	0.743	0.198

Table 3. Quantitative comparison of trailing suppression strategies.

Methods	PSNR↑	SSIM↑	LPIPS↓
ETS	24.59	0.748	0.197
DDIM	24.51	0.742	0.201
LETC + DDIM	24.49	0.741	0.201

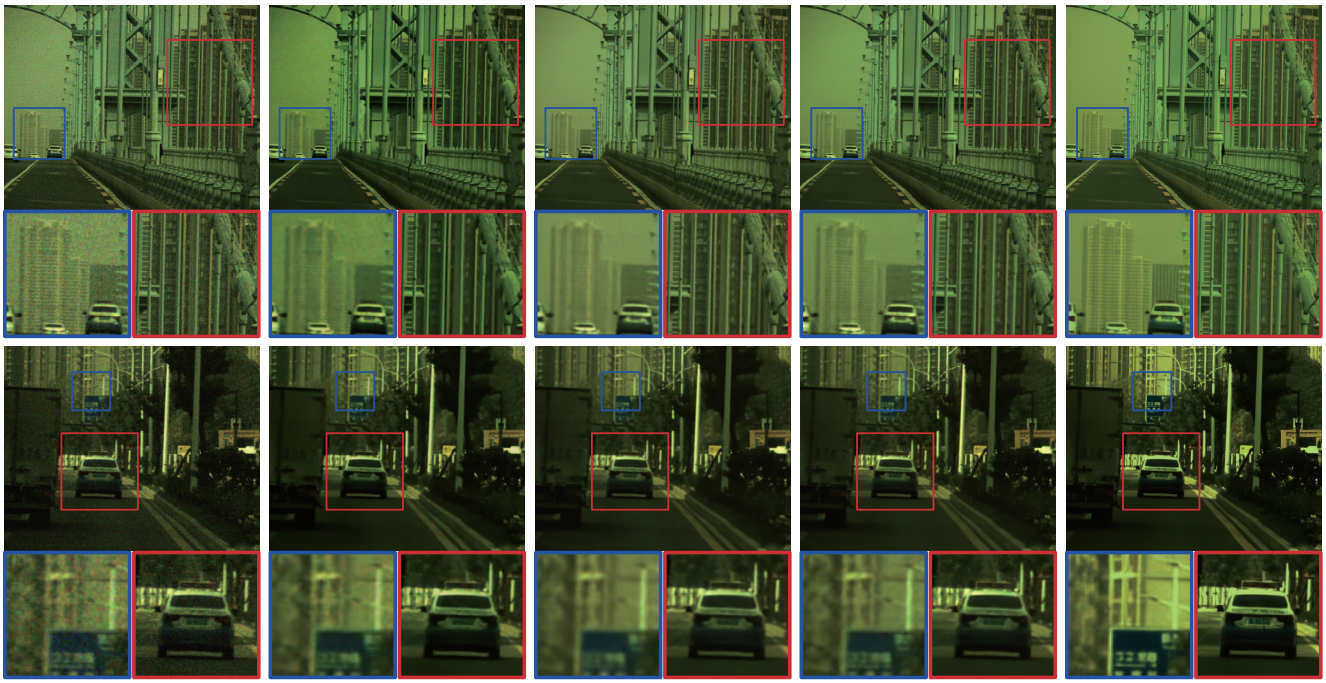


Figure 2. Visual comparison between NEC-Diff and existing RAW denoising methods.

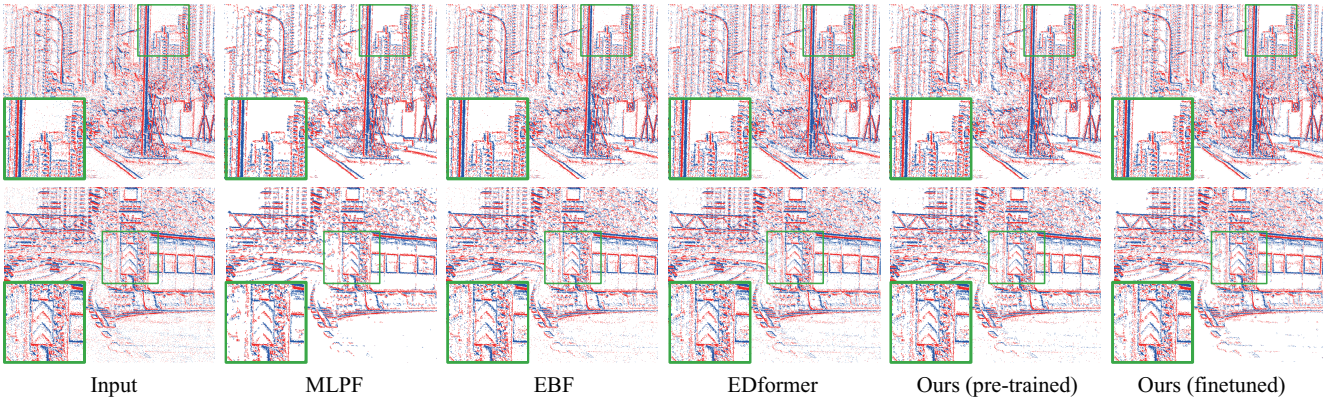


Figure 3. Visual comparison between NEC-Diff and existing event denoising methods.

the diffusion model. Tab. 3 compares both strategies. ETS preprocessing achieves superior reconstruction quality, but at the cost of requiring additional preprocessing for all input data, whereas the implicit-learning strategy is more flexible at inference time.

2.3. Comparison of RAW Denoising Results

Comparison Methods. To validate the effectiveness of the denoising strategy in NEC-Diff, we compare the output of the pretrained RAW-Denoise module and the output after end-to-end training with the intensity-consistency loss against three SOTA RAW denoising methods. LED [13] and RID [14] are image-based approaches, while BRVE [15] operates on video sequences. All methods are evaluated on the REAL dataset. Since ground-truth clean images are unavailable, we adopt the no-reference metrics NIQE [17] and BRISQUE [18] to assess denoising performance. In addition, a user study (US) is conducted to quantify perceptual quality, where 30 participants rate the visual results on a scale from 1 to 10, with higher scores indicating better denoising quality.

Qualitative and Quantitative Results. Fig. 2 and Tab. 4 present the qualitative and quantitative comparisons on the REAL low-light benchmark. The RAW-based method LED [13] leaves noticeable residual noise, while RID [14]

Table 4. Quantitative comparison of different denoising methods.

Methods	NIQE ↓	BRISQUE ↓	US ↑
LED [13]	6.299	50.07	6.63
BRVE [15]	5.643	47.84	6.85
RID [14]	5.386	46.53	6.91
Ours (pre-trained)	5.253	44.68	6.88
Ours (finetuned)	**4.188**	**37.83**	**7.94**

and the video-based BRVE [15] introduce blurring in fine textures. Our RAW-only pretrained model also suffers from texture smoothing. In contrast, after incorporating event information and performing end-to-end finetuning with the intensity-consistency loss, the RAW-Denoise branch effectively preserves fine textures while maintaining strong denoising performance.

2.4. Comparison of Event Denoising Results

Comparison Methods. We further evaluate the contribution of the NEC-Diff denoising strategy to event denoising. Specifically, we compare the outputs of the pretrained EV-Denoise module and the end-to-end finetuned version with three SOTA event denoising methods: MLPF [19], EBF [20], and EDformer [3]. All methods are evaluated on the REAL dataset. Due to the significant modality gap between events and images, existing no-reference image quality metrics cannot reliably assess event denoising performance; therefore, we focus on visual comparison only.

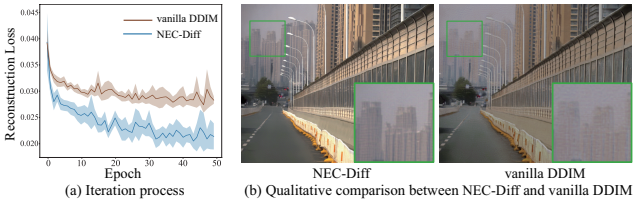
Qualitative Results. Fig. 3 presents the qualitative comparison. MLPF [19] effectively suppresses noise but tends to over-smooth textures in relatively uniform regions. EBF [20], EDformer [3], and our pretrained EV-Denoise model better preserve fine structures, yet often leave noticeable residual noise. In contrast, benefiting from illumination priors and the intensity-consistency constraint, our method achieves a more balanced denoising strength, simultaneously suppressing noise and retaining subtle textures.

2.5. Effectiveness of the Consistency Loss

The ablation results for the consistency loss are partially presented in Fig. 7(b) of the main paper. For completeness, we additionally report the corresponding quantitative results in Tab. 5. Here, ER denotes cross-modal refinement, where the other modality is used as an auxiliary input for denoising. Specifically, RAW guides event denoising, while events guide RAW denoising. We evaluate the effects of ER and the consistency loss $\mathcal{L}_{\text{cons}}$ on the REAL dataset. As shown in the table, introducing either ER or $\mathcal{L}_{\text{cons}}$ alone yields only limited improvements. By contrast, their combination leads to clear and consistent gains across all metrics, improving PSNR from 22.35 dB to 24.51 dB and SSIM from 0.685 to 0.742, while reducing LPIPS from 0.212 to 0.201. This

Table 5. Ablation study of the refinement strategy (ER) and the consistency loss $\mathcal{L}_{\text{cons}}$ on the REAL dataset.

ER	$\mathcal{L}_{\text{cons}}$	PSNR↑	SSIM↑	LPIPS↓
		22.35	0.685	0.212
✓		22.61	0.692	0.209
	✓	23.12	0.701	0.202
✓	✓	**24.51**	**0.742**	**0.201**



(a) Iteration process (b) Qualitative comparison between NEC-Diff and vanilla DDIM

Figure 4. Comparison between NEC-Diff and vanilla DDIM.

demonstrates that the consistency loss works synergistically with ER, helping stabilize multimodal optimization and improve the final reconstruction quality.

2.6. Role of the SNR Map

The SNR map is introduced to indicate more reliable regions for guiding event–frame fusion, rather than providing additional scene information. In other words, it acts as a reliability prior that helps the model adaptively balance the contributions of the low-light frame and the event stream under severe noise. To further clarify this design choice, we compare NEC-Diff with a vanilla DDIM that directly concatenates the image and event features without SNR-based weighting. The comparison is shown in Fig. 4. As can be observed, vanilla DDIM exhibits unstable optimization behavior and is difficult to converge during training, which leads to inferior visual quality in the final reconstruction. In contrast, the proposed design provides more stable optimization and better restoration results, demonstrating that the SNR-guided fusion strategy is essential for effective multimodal diffusion under extreme low-light conditions.

3. Additional Results

To complement the results presented in the main paper, we provide additional qualitative evaluations on the REAL and LLRVD-simu datasets, as shown in Fig. 5. These supplementary examples include more scenes that were not shown in the main text. As illustrated in the following figures, the proposed method consistently outperforms existing approaches by preserving fine structures and suppressing noise artifacts. The extended comparisons further demonstrate the strong robustness and generalization ability of NEC-Diff framework across both real-captured and simulated low-light scenarios.

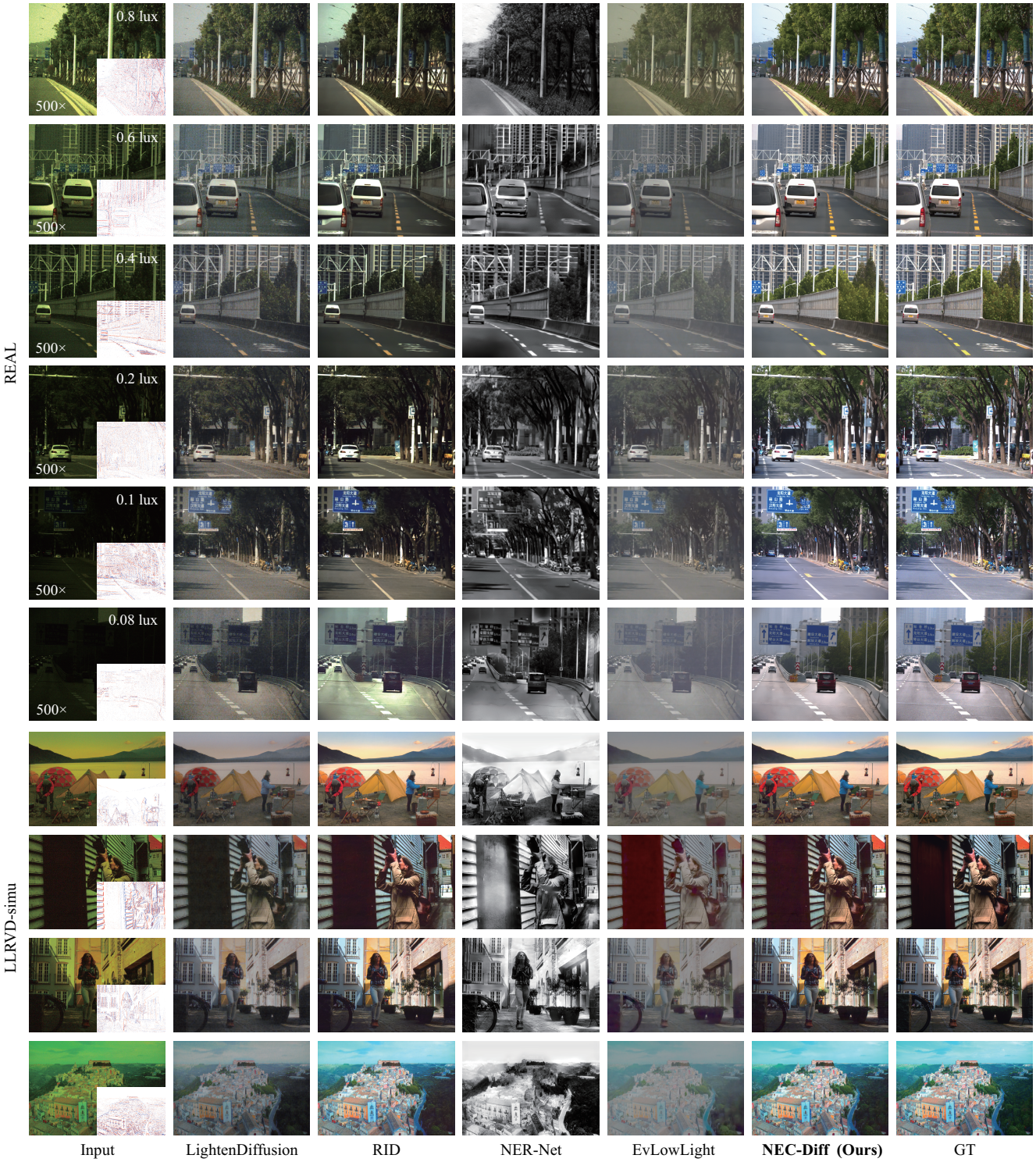


Figure 5. Additional qualitative comparisons under low-light conditions. The first six rows correspond to examples from the REAL dataset, while the last four rows present results from the LLRVD-simu dataset.

References

[1] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 989–997, 2019. 1

[2] Yue Cao, Ming Liu, Shuai Liu, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Physics-guided iso-dependent sensor noise modeling for extreme low-light photography. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5744–5753, 2023. 2

[3] Bin Jiang, Bo Xiong, Bohan Qu, M Salman Asif, You Zhou, and Zhan Ma. Edformer: Transformer-based event denoising across varied noise levels. In *Eur. Conf. Comput. Vis.*, pages 200–216, 2024. 2, 5

[4] Ying Fu, Zichun Wang, Tao Zhang, and Jun Zhang. Low-light raw video denoising with a high-quality realistic motion dataset. *IEEE Transactions on Multimedia*, 25:8119–8131, 2022. 2

[5] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1312–1321, 2021. 2

[6] Yu Jiang, Yuehang Wang, Siqi Li, Yongji Zhang, Minghao Zhao, and Yue Gao. Event-based low-illumination image enhancement. *IEEE Trans. Multimedia*, 26:1920–1931, 2023. 2

[7] Guoqiang Liang, Kanghao Chen, Hangyu Li, Yunfan Lu, and Lin Wang. Towards robust event-guided low-light image enhancement: a large-scale real-world event-image dataset and novel approach. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23–33, 2024. 2

[8] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Int. Conf. Comput. Vis.*, pages 10615–10625, 2023. 2

[9] Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *Eur. Conf. Comput. Vis.*, pages 433–451, 2024. 2

[10] Henri Rebecq, Rene Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3857–3866, 2019. 2

[11] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. 2

[12] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *Eur. Conf. Comput. Vis.*, pages 341–357. Springer, 2022. 2

[13] Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Ruixun Zhang, Xialei Liu, and Chongyi Li. Lighting every darkness in two pairs: A calibration-free pipeline for raw denoising. In *Int. Conf. Comput. Vis.*, pages 13275–13284, 2023. 3, 4, 5

[14] Feiran Li, Haiyang Jiang, and Daisuke Iso. Noise modeling in one hour: Minimizing preparation efforts for self-supervised low-light raw image denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5699–5708, 2025. 3, 4, 5

[15] Gengchen Zhang, Yulun Zhang, Xin Yuan, and Ying Fu. Binarized low-light raw video enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 25753–25762, 2024. 3, 4, 5

[16] Haoyue Liu, Shihan Peng, Lin Zhu, Yi Chang, Hanyu Zhou, and Luxin Yan. Seeing motion at nighttime with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25648–25658, 2024. 3

[17] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Sign. Process. Letters*, 20(3):209–212, 2012. 4

[18] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 4

[19] Shasha Guo and Tobi Delbruck. Low cost and latency event camera background activity denoising. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):785–795, 2022. 5

[20] Shasha Guo, Chenyang Shi, Lei Wang, Jing Jin, and Yuliang Lu. Ebf: An event-based bilateral filter for effective neuromorphic vision sensor denoising. *IEEE Trans. Circuits Syst. Video Technol.*, 2025. 5