LIGHTWEIGHT BUT GLOBAL: EVENT-GUIDED BROAD-RANGE BRIGHTNESS RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Broad-range brightness restoration maps RGB frames captured under low, normal, or high illumination to a target exposure. Events provide complementary measurements when the frames are noisy or saturated. Balancing global spatial context, temporal information, and reconstruction detail within a compact model remains challenging. To this end, we present MambaSEE, a compact spatiotemporal architecture that assigns these functions to different operators. First, fixed-width additive fusion and attentive state-space modules model each frame–event pair with global context. Next, a weight-and-activation 8-bit quantization-aware ConvGRU aggregates three first-in, first-out pairs through backward-then-forward propagation, without observations later than the prediction time. Finally, a prompt-conditioned depthwise decoder reconstructs fine image details. On the SDE and SEE-600K benchmarks, MambaSEE achieves the highest reported average PSNR with the **lowest FP32-equivalent weight storage** among the compared methods. On SEE-600K, its PSNR values are 2.01, 1.39, and 5.32 dB higher than the published SEENet results for low-, high-, and normal-light inputs, respectively. MambaSEE uses mixed-precision weight storage equivalent to **0.48M** FP32 parameters, about one quarter of SEENet's.

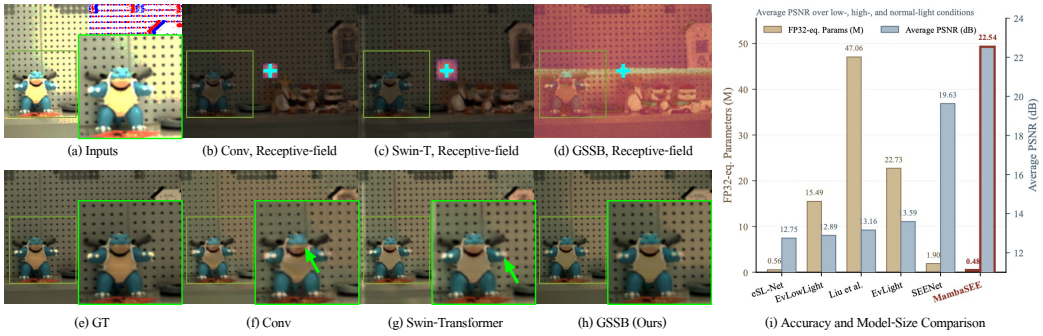


Figure 1: **Lightweight but global.** From left to right: schematic receptive-field illustration, qualitative restoration comparison, and accuracy–model-size comparison. The translucent red regions illustrate spatial context for the area marked by the green cross, not measured effective receptive fields. MambaSEE captures broader spatial context and recovers clearer structural details, whereas previous methods exhibit ghosting and discontinuities, as indicated by the green arrows. MambaSEE achieves the highest reported average PSNR with the lowest FP32-equivalent weight storage among the plotted methods.

1 INTRODUCTION

Natural illumination spans several orders of magnitude, whereas an RGB camera has a finite exposure and a limited digital range (Koshel, 2012; Bernacki, 2020). Underexposure hides weak signals beneath sensor noise, while overexposure clips strong responses. Both cases destroy texture and contrast, making normal-exposure restoration ill-posed. Existing frame-based methods address this problem during or after capture. Automatic exposure selects a global camera setting (Yuan & Sun,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2012). Multi-exposure and burst imaging combine observations (McCann & Rizzi, 2011; Hasinoff et al., 2016). Learning-based methods map degraded images to normal exposure (Wei et al., 2018; Zhang et al., 2019; Cai et al., 2023). However, these methods remain limited by frame measurements. Repeated exposures are sensitive to motion and often cause ghosting (Tursun et al., 2015), while single-frame methods must infer content hidden by noise or lost to saturation. These limitations have motivated the use of alternative sensing modalities.

Event cameras provide a complementary sensing pathway. They asynchronously record log-intensity changes with high temporal resolution and a dynamic range above 120 dB (Gallego et al., 2020). As a result, events can preserve motion boundaries and intensity changes when standard frames are degraded. Their benefits have been shown in low-light enhancement and high-dynamic-range reconstruction (Liu et al., 2023; Liang et al., 2023; 2024; Han et al., 2020; Yang et al., 2023; Cui et al., 2024). However, events do not record absolute intensity, color, or static texture. Therefore, event-guided restoration must use sparse intensity changes without allowing event noise to damage the dense image appearance.

SEENet (Lu et al., 2025) introduced events for broad-range brightness adjustment and presented the large-scale dataset. It treats events as a brightness dictionary and queries them using image features, enabling a unified solution for low-, normal-, and high-light inputs. However, its computational cost and frame-wise processing raise two challenges for compact spatiotemporal restoration. First, its pairwise attention requires high computation at image resolution, as shown in Fig. 1. Second, it processes each target observation independently, without explicitly aggregating preceding frame–event pairs. Simply reducing model capacity may weaken spatial modeling. Directly stacking temporal observations increases computation and may mix misaligned content. These trade-offs motivate our separation of global spatial modeling, bounded temporal aggregation, and lightweight reconstruction.

MambaSEE allocates computation according to the structure of each information source rather than uniformly shrinking a large fusion backbone. As shown in our framework, MambaSEE separates four functions. First, additive image–event fusion keeps the feature width fixed. Second, Global State-Space Blocks (GSSBs), inspired by MambaIRv2 (Guo et al., 2025a), group semantically related spatial tokens and model global context with linear sequence complexity. Third, a quantization-aware ConvGRU (Ballas et al., 2016) aggregates a three-pair FIFO history. Finally, a depthwise convolutional decoder focuses on local detail reconstruction. Spatial modeling is performed before temporal aggregation. Therefore, the temporal module processes frame-wise multimodal features, while the decoder receives both the aggregated feature and the target-exposure prompt.

To predict the image at time t, MambaSEE uses a three-pair FIFO window from $t-2$ to t. A backward pass supplies time-aligned context to a forward pass, whose newest fused feature feeds the decoder. Each window resets both recurrent states. Our contributions are threefold:

- **Lightweight Formulation:** We introduce a compact decomposition for event-guided brightness restoration that assigns multimodal fusion, global spatial reasoning, temporal aggregation, and local reconstruction to specialized lightweight operators. This decomposition supports online restoration with 0.48M FP32-equivalent weight storage.
- **Efficient Spatiotemporal Framework:** We combine fixed-width additive fusion and full-precision GSSBs with a backward-then-forward INT8 quantization-aware ConvGRU. Time-aligned reverse states condition the forward recurrence within a bounded FIFO history; a prompt-conditioned depthwise decoder reconstructs local details.
- **Strong Accuracy-Efficiency Trade-off:** In these benchmark comparisons, MambaSEE achieves the highest reported average PSNR on SDE (Liang et al., 2024) and the best reported PSNR, SSIM, and L_1 scores across all SEE-600K (Lu et al., 2025) illumination conditions.

2 Related Work

Frame-Based Methods: Frame-based brightness restoration covers both exposure correction across under- and overexposed inputs (Afifi et al., 2021) and low-light enhancement. Retinex-based methods separate reflectance and illumination (Wei et al., 2018; Zhang et al., 2019; Wu et al., 2022; Cai et al.,

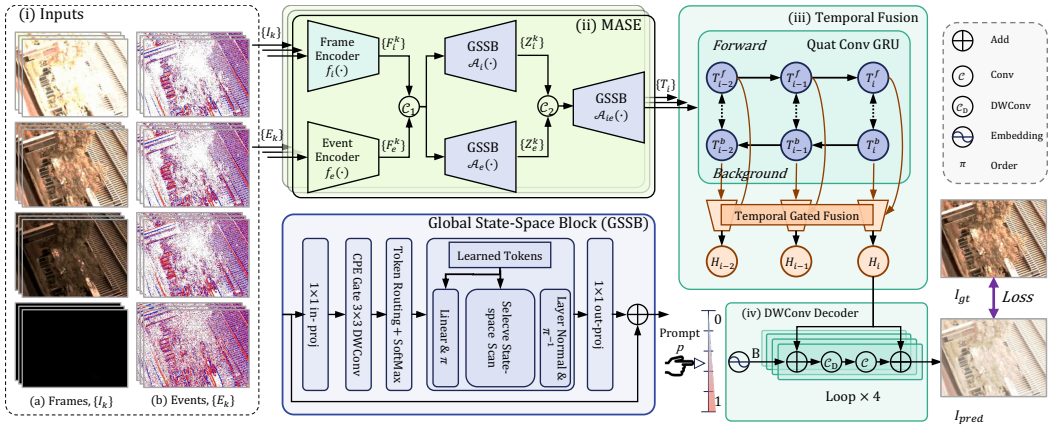


Figure 2: **Overview of MambaSEE.** (i) A FIFO sequence of frames and events is provided as input. (ii) MASE combines additive fusion and GSSBs for global spatial modeling. (iii) W8A8 ConvGRUs propagate backward then forward, followed by per-index gated fusion; only H_i feeds (iv) the prompt-conditioned DWConv decoder. The inset details the GSSB. $\oplus$, $\mathcal{C}$, $\mathcal{C}_D$, and π denote addition, convolution, depthwise convolution, and token ordering. For clarity, the diagram omits MASE's signed residual and activation paths (Eq. 5).

2023), while other methods learn enhancement curves, direct mappings, or conditional distributions (Guo et al., 2020; Xu et al., 2022; Wang et al., 2022a). Transformer restoration models enlarge the effective receptive field and capture long-range interactions (Liang et al., 2021; Wang et al., 2022b; Zamir et al., 2022; Wang et al., 2023). Despite their progress, these methods must infer missing content from the degraded frame alone.

Event-Guided Methods: Events support low-light reconstruction, HDR imaging, deblurring, and video restoration. Event-only methods offer high temporal resolution but lack color and absolute intensity (Zhang et al., 2020; Rebecq et al., 2019). Early frame–event fusion demonstrated high-quality image recovery (Wang et al., 2020), while event-guided low-light methods exploit structural and motion cues (Liu et al., 2023; Liang et al., 2023). EvLight and EvLight++ select reliable features through signal-to-noise guidance (Liang et al., 2024; Chen et al., 2024); recent real-event methods address extreme darkness (Wang et al., 2024), and RetinEV and ERetinex use events for illumination and structure recovery (Sun et al., 2025a; Guo et al., 2025b). Events also support multi-bracket and single-exposure HDR restoration across large exposure differences (Messikommer et al., 2022; Han et al., 2020; Yang et al., 2023; Cui et al., 2024; Weng et al., 2024b). SEE introduced broad-range brightness adjustment and the SEE-600K dataset (Lu et al., 2025), whose prompt-based setting and evaluation protocol are adopted in our work. MambaSEE further targets the accuracy–capacity trade-off and uses preceding observations for online restoration. EvLIR applies a ConvGRU to preserve event-bin order within one exposure interval (Zhou et al., 2026), whereas MambaSEE aggregates fused frame–event representations across frame times.

Efficient Global and Temporal Modeling: Self-attention provides global interaction but has quadratic complexity in the token count (Vaswani, 2017). Structured state-space models and selective SSMs provide linear-time sequence modeling (Gu et al., 2022; Gu & Dao, 2024); Vision Mamba and VMamba adapt this principle to bidirectional and two-dimensional visual scanning (Zhu et al., 2024; Liu et al., 2024). For restoration, MambaLLIE combines global and local state-space reasoning for low-light enhancement (Weng et al., 2024a), MambaIR augments selective scanning with local enhancement and channel attention (Guo et al., 2024), and MambaIRv2 groups semantically related tokens in the scan order (Guo et al., 2025a). EVDM further combines events and Mamba for cross-modal fusion and temporal propagation in video deblurring (Sun et al., 2025b). Temporal recurrence offers a complementary way to reuse observations across neighboring frames. ConvGRU preserves spatial structure during recurrent aggregation (Ballas et al., 2016), while bidirectional recurrent propagation is effective in video restoration (Chan et al., 2022). Efficiency also depends on arithmetic and data movement. Quantization-aware training exposes a network to low-bit weights and activations during optimization (Jacob et al., 2018; Esser et al., 2020; Shen et al., 2021), and BMTNet explores low precision around the sensitive selective recurrence of Mamba (Zhou et al., 2025). MambaSEE

instead retains full-precision global GSSBs and targets W8A8 QAT at the repeatedly evaluated temporal aggregator.

3 METHOD

Problem Formulation: Let $I_k \in \mathbb{R}^{3\times H\times W}$ denote the RGB image captured at time k, and let $E_k \in \mathbb{R}^{B_e\times H\times W}$ be its associated event slice within the current FIFO input, where $B_e = 64$. We use a FIFO window of $K = 3$ paired observations ending at the latest time i. Given a scalar target-brightness prompt p, MambaSEE f_{see} predicts

$$I_{\text{pred}} = f_{see}\big(\{I_k, E_k\}_{k=i-K+1}^{i}, p\big)\,, \tag{1}$$

where I_{gt} is the normally exposed target at the latest time. The model never consumes an observation later than i. Following the SEE task protocol (Lu et al., 2025), the target prompt is the global mean of the exposed reference:

$$p = \frac{1}{3HW}\sum_{c=1}^{3}\sum_{y=0}^{H-1}\sum_{x=0}^{W-1} I_{\text{gt}}(c, y, x). \tag{2}$$

This prompt descriptor specifies the desired global exposure but provides no spatial structure. Accordingly, our evaluation covers restoration to a known target exposure; automatic exposure selection remains outside its scope.

Overview: Single-image restoration backbones do not directly account for the different statistics of dense frames and sparse events, nor do they specify how preceding observations should be reused for online prediction. MambaSEE therefore separates modality adaptation, global spatial modeling, temporal aggregation, and reconstruction instead of concatenating every input into one widened backbone. As shown in Fig. 2, modality-specific heads first embed each RGB frame and event voxel in a shared feature space. A parameter-shared Mamba Additive Spatial Encoder (MASE) then performs fixed-width fusion and frame-wise global modeling, producing T_k. Next, a W8A8 quantization-aware FIFO ConvGRU aggregates $\{T_{i-2}, T_{i-1}, T_i\}$ into H_i, after which an exposure-conditioned depthwise decoder reconstructs I_{pred}. Thus, GSSBs model an HW-length spatial sequence per frame, while recurrence operates only on a short history of already fused features.

3.1 MAMBA ADDITIVE SPATIAL ENCODER

Modality-specific Embedding. We first construct a four-channel coordinate map

$$P_0(x, y) = [y \bmod 2 \quad x \bmod 2 \quad y/H \quad x/W]^\top\,, \tag{3}$$

where x, y are zero-based column and row indices, respectively. Two 1×1 convolutions transform P_0 into a learned positional embedding P. The same P is supplied to both modalities, providing a common spatial reference before their heterogeneous measurements are fused. For each time k, the image and event heads compute

$$F_{\text{i}}^k = f_{\text{i}}([I_k, P]), \qquad F_{\text{e}}^k = f_{\text{e}}([E_k, P]). \tag{4}$$

where $[\cdot,\cdot]$ denotes concatenation. f_{i} is a 1×1 projection, whereas f_{e} uses two 1×1 convolutions with sigmoid activations to adapt sparse events to the shared width. The separate heads prevent the magnitude and sparsity of event voxels from being treated as if they followed RGB intensity statistics. Meanwhile, the shared coordinate embedding provides both streams with an identical spatial reference. Its parity terms identify the local 2×2 sampling phase, while the normalized coordinates disambiguate spatially distant locations before semantic routing.

Fixed-width Additive Fusion. Addition avoids doubling the width of all subsequent global blocks. This is particularly important because every extra channel would be propagated through all token projections and state-space blocks. With 3×3 convolutions $\mathcal{C}_1, \mathcal{C}_2$ and H-Swish ϕ (Howard et al., 2019), MASE computes

$$\begin{aligned}
F^k &= \mathcal{C}_1(F_{\text{i}}^k + F_{\text{e}}^k), \qquad \widetilde{F}^k = \phi(F^k),\\
Z_{\text{i}}^k &= \mathcal{A}_{\text{i}}(\widetilde{F}^k), \qquad Z_{\text{e}}^k = \mathcal{A}_{\text{e}}(\widetilde{F}^k),\\
Z_{\text{ie}}^k &= \mathcal{C}_2(Z_{\text{i}}^k + Z_{\text{e}}^k),\\
T_k &= \phi\Big(\widetilde{F}^k - \mathcal{A}_{\text{ie}}(Z_{\text{ie}}^k) + F^k\Big)\,.
\end{aligned} \tag{5}$$

Both $\mathcal{A}_{\mathrm{i}}$ and $\mathcal{A}_{\mathrm{e}}$ receive the same fused feature $\widetilde{F}^k$ and differ only in their parameters; their subscripts identify the two parallel branches rather than image-only and event-only inputs. Independent transformations allow them to learn complementary global responses without duplicating the fused representation. The final residual retains low-level appearance, and the same MASE is applied independently to every pair before temporal fusion.

3.2 Global State-Space Modeling

Each $\mathcal{A}$ is a full-precision Global State-Space Block (GSSB) inspired by the Attentive State-Space Module of MambaIRv2 (Guo et al., 2025a). Given $U \in \mathbb{R}^{L\times C}$ with $L = HW$, a hard Gumbel-Softmax router (Jang et al., 2017) assigns visual tokens to learned semantic tokens and produces a state prompt ρ. The induced permutation π places semantically related locations nearby along the reordered sequence before selective scanning:

$$\begin{aligned} U_\pi &= \mathrm{Sort}_\pi(\Psi(U)), \\ V_\pi &= \mathcal{W}_o\,\mathrm{LN}[\mathrm{SSM}(U_\pi;\rho)], \\ \mathcal{A}(U) &= U + \mathrm{Unsort}_\pi(V_\pi), \end{aligned} \tag{6}$$

where Ψ denotes the input projection and conditional positional gate, while $\mathcal{W}_o$ is the output projection. The inverse permutation restores the image layout. Reordering allows distant regions with similar appearance or exposure to exchange information along the selective scan, while the scan remains linear in the number of spatial tokens. For the l-th reordered token $u_l = (U_\pi)_l$, the prompted selective state-space recurrence is written as

$$\begin{aligned} h_l &= \bar{A}_l h_{l-1} + \bar{B}_l u_l, \\ o_l &= (c_l + \rho_l)^\top h_l + D_{\mathrm{ssm}} u_l, \end{aligned} \tag{7}$$

where h_l and o_l denote the selective state and output feature, respectively. The input-dependent $\bar{A}_l$ and $\bar{B}_l$ are the discretized state-transition and input parameters, while c_l is the readout vector. The routing-derived prompt ρ_l modulates the state readout, and D_{ssm} denotes the learned direct feed-through parameter. The state therefore propagates global context without forming the dense pairwise affinity matrix required by self-attention. Unlike the single-image pipeline of MambaIRv2 (Guo et al., 2025a), MambaSEE places each GSSB after fixed-width image–event fusion and before temporal aggregation. Spatial scans therefore remain HW long, while the compact recurrent module handles history instead of extending the state-space sequence to KHW tokens. The novelty is thus the task-specific organization of global state-space reasoning with sparse events and bounded temporal evidence, rather than the attentive operator alone.

3.3 Quantized FIFO Temporal Aggregation

FIFO Feature Fusion. MASE produces $\{T_{i-2}, T_{i-1}, T_i\}$, pooled to $X_k = \mathrm{AvgPool}_2(T_k)$. A reverse ConvGRU (Ballas et al., 2016) first traverses the window from i to $i-2$. Its aligned state T_k^b conditions the forward input $\widetilde{X}_k = X_k + \mathcal{P}_b(T_k^b)$ through a learned W8A8 1×1 projection. The forward ConvGRU then traverses $i-2$ to i. Upsampled states are fused with the full-resolution feature:

$$\begin{aligned} \Gamma_k &= [T_k, \mathrm{Up}(T_k^f), \mathrm{Up}(T_k^b)], \\ g_k &= G(\Gamma_k), \qquad \Delta_k = D(\Gamma_k), \\ H_k &= T_k + g_k \odot \Delta_k. \end{aligned} \tag{8}$$

The last convolution of the delta branch D is zero-initialized, so temporal fusion begins as an identity update.

Only H_i is decoded for the reported prediction. Although T_i^b starts from the current feature and zero state, every reverse state also enters the forward recurrence and can affect T_i^f. Both passes reset each window; no input later than i is used. Appendix B gives the recurrences and counting scope. Accuracy uses full-window recomputation; the conditional cost estimate assumes cached MASE features. Weights are shared across time, and $K = 3$ bounds recurrent work.

Quantization-aware Temporal Computation. We expose the temporal convolutions to INT8 discretization during training (Jacob et al., 2018; Esser et al., 2020; Shen et al., 2021). For tensor x with scale s, signed uniform fake quantization is

$$\begin{aligned} q(x;s) &= \mathrm{clip}\big(\mathrm{round}(x/s), -128, 127\big), \\ Q(x;s) &= s \cdot q(x;s). \end{aligned} \tag{9}$$

Weights use per-output-channel scales, activations use per-tensor scales, and a straight-through estimator handles the non-differentiable rounding operation. Fake quantization retains floating-point tensors during training while simulating the values representable by the deployment quantizer. W8A8 simulation covers the ConvGRUs, $\mathcal{P}_b$, and the convolutions in G and D; nonlinearities, interpolation, hidden states, GSSBs, and the decoder remain floating point. We therefore report bit-width-aware arithmetic estimates but make no measured latency or energy claim without an exported integer backend on target hardware.

3.4 Exposure-Conditioned Lightweight Reconstruction

The decoder expands p into a constant map and uses two 1×1 convolutions to produce the exposure embedding B shown in Fig. 2. Its input feature is initialized as $R_0 = \mathcal{C}_{\mathrm{in}}(H_i + B)$, and the same embedding is injected in every residual block. The decoder contains four residual groups with two blocks per group. Given feature R_l, a block computes

$$\begin{aligned} \widetilde{R}_l &= \mathcal{C}_D^l\big(\,\mathrm{GN}(R_l + B)\big), \\ R_{l+1} &= R_l + \mathcal{C}_{\mathrm{pw},2}^l\big(\phi(\mathcal{C}_{\mathrm{pw},1}^l(\widetilde{R}_l))\big), \end{aligned} \tag{10}$$

where $\mathcal{C}_D$ is a 5×5 depthwise convolution (Chollet, 2017), GN is group normalization (Wu & He, 2018), and the pointwise convolutions expand then restore the channel width. Depthwise filtering enlarges the local receptive field without the parameter cost of a dense spatial convolution, while group normalization does not rely on large-batch statistics. Four two-block residual groups use dilation rates 1 and 2. Repeatedly injecting B keeps the requested exposure available throughout reconstruction rather than only at the decoder input. A depthwise global tail, GELU (Hendrycks & Gimpel, 2016), and a 3×3 convolution produce I_{pred}.

Loss Function: We optimize a Charbonnier reconstruction term (Lai et al., 2018) and a gradient-magnitude term. Let $N = 3HW$, $\epsilon_n = I_{\mathrm{pred},n} - I_{\mathrm{gt},n}$, and $\delta_d = |\nabla_d I_{\mathrm{pred}}| - |\nabla_d I_{\mathrm{gt}}|$. The loss function is

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\epsilon_n^2 + 10^{-6}} + \sum_{d\in\{x,y\}} \|\delta_d\|_{1,\mathrm{mean}}. \tag{11}$$

The robust Charbonnier penalty limits the influence of large pixel residuals, whereas the gradient term discourages over-smoothed edges; $\|\cdot\|_{1,\mathrm{mean}}$ denotes the mean absolute value. Temporal evidence is learned in feature space.

4 Experiments

Experimental Setup: ***Datasets:*** We evaluate MambaSEE on SDE (Liang et al., 2024) and SEE-600K (Lu et al., 2025). SDE evaluates indoor and outdoor low-light enhancement, whereas SEE-600K covers low-, high-, and normal-light inputs. We use the official scene-level split of SEE-600K and the sequence-level split of SDE. Following the target-conditioned formulation of SEE (Lu et al., 2025), MambaSEE receives the reference-derived target-brightness prompt in Eq. 2 on both datasets. Each FIFO input contains three causal RGB frames ending at the prediction time and a 192-channel event tensor partitioned into three consecutive 64-channel slices. ***Evaluation metrics:*** We report PSNR and SSIM (Wang et al., 2004) on both datasets and the protocol-specific L_1 error on SEE-600K; higher PSNR/SSIM and lower L_1 are better. ***Complexity metrics:*** Table 2 reports operation-count estimates and FP32-equivalent weight storage. For MambaSEE, following EMQ (Dong et al., 2023) and USDN (Jeon et al., 2024), we compute

$$P_{\mathrm{eq32}} = \sum_l N_l \frac{b_l^w}{32}, \qquad \mathrm{BOPs} = \sum_l F_l b_l^w b_l^a. \tag{12}$$

Method	Low Light			High Light			Normal Light			Average		
	PSNR ↑	SSIM ↑	L_1 ↓	PSNR ↑	SSIM ↑	L_1 ↓	PSNR ↑	SSIM ↑	L_1 ↓	PSNR ↑	SSIM ↑	L_1 ↓
eSL-Net	11.95	0.3845	0.2421	12.84	0.4660	0.2076	13.45	0.5682	0.1957	12.75	0.4729	0.2151
EvLowLight	12.83	0.4511	0.2151	12.79	0.4696	0.2084	13.04	0.5531	0.2144	12.89	0.4913	0.2126
Liu et al.	13.48	0.5068	0.1946	12.30	0.4766	0.2221	13.70	0.5474	0.2151	13.16	0.5103	0.2106
EvLight	13.70	0.5150	0.1960	13.45	0.4918	0.1990	13.63	0.5924	0.2004	13.59	0.5331	0.1985
SEENet	<u>18.77</u>	<u>0.6303</u>	<u>0.0971</u>	<u>19.21</u>	<u>0.6675</u>	<u>0.0806</u>	<u>20.92</u>	<u>0.8002</u>	<u>0.0606</u>	<u>19.63</u>	<u>0.6993</u>	<u>0.0794</u>
MambaSEE (Ours)	**20.78**	**0.6961**	**0.0794**	**20.60**	**0.6981**	**0.0701**	**26.24**	**0.8761**	**0.0338**	**22.54**	**0.7568**	**0.0611**

Table 1: Benchmark comparison on SEE-600K. Baselines use the SEE-trained results of Lu et al. (2025); see Appendix A.3 for sources and input conditions. Average equally weights the three illumination conditions. Best and second-best reported scores are **bold** and <u>underlined</u>, respectively.

Method	FLOPs (G)	Eq. Storage (M)	Events	Indoor		Outdoor		Average	
				PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SNR	26.35	4.01	×	20.05	0.6302	22.18	0.6611	21.12	0.6457
UFormer	**12.00**	5.29	×	21.09	0.7524	22.32	0.7469	21.71	0.7497
LLFlow	409.50	39.91	×	20.92	0.6610	21.68	0.6467	21.30	0.6539
Retinexformer	15.57	1.61	×	21.30	0.6920	22.92	0.6834	22.11	0.6877
E2VID+	27.99	10.71	✓	15.19	0.5891	15.01	0.5765	15.10	0.5828
ELIE	440.32	33.36	✓	19.98	0.6168	20.69	0.6533	20.34	0.6350
HDRev	118.65	13.42	✓	21.13	0.6239	21.82	0.6824	21.47	0.6531
Wang et al.	170.32	7.38	✓	21.29	0.6786	22.08	0.7052	21.68	0.6919
eSL-Net	560.94	0.56	✓	21.25	0.7277	22.42	0.7187	21.84	0.7232
Liu et al.	44.71	47.06	✓	21.79	0.7051	23.35	0.6895	22.57	0.6973
EvLowlight	524.95	15.49	✓	20.57	0.6217	20.04	0.6485	20.31	0.6351
EvLight	180.90	22.73	✓	22.44	0.7697	23.21	0.7505	22.83	0.7601
SEENet	405.72	1.90	✓	22.54	0.7756	<u>24.60</u>	<u>0.7692</u>	<u>23.57</u>	0.7724
EvLight++	225.91	26.21	✓	<u>22.67</u>	<u>0.7791</u>	23.34	0.7676	23.01	<u>0.7733</u>
MambaSEE (Ours)	45.19†	**0.48**	✓	**25.01**	**0.7883**	**26.40**	**0.7798**	**25.71**	**0.7841**

Table 2: SDE benchmark comparison with published baselines. †MambaSEE: conditional steady-state estimate with cached spatial features; scores use full-window recomputation. MambaSEE's Average equally weights its two printed subset scores. Baseline FLOPs are quoted; MambaSEE counts one operation per MAC. Storage is FP32-equivalent with FP32 baseline weights. Sources and protocol details: Appendix A.3. Best and second-best reported scores are **bold** and <u>underlined</u>, respectively.

where N_l, F_l, b_l^w, and b_l^a denote weight count, operation count, and assigned weight/activation bit-widths. MambaSEE has $0.73M$ trainable parameters, including $0.32M$ W8 temporal weights, yielding $0.48M$ FP32-equivalent storage. Under steady-state feature reuse, the conditional estimate counts one new spatial pair, both complete recurrent passes, the inter-pass projections $\mathcal{P}_b$, all gated updates, and one decode. At 256×256, the reported estimate is $45.19G$ MACs. Within this scope, temporal W8A8 reduces nominal BOPs by 41.53% relative to all-FP32.

Implementation Details: The model uses 64-channel features, GSSBs with state dimension 16, semantic rank 32, and 64 tokens, and a 48-channel FIFO ConvGRU. We configure training schedules of 20 epochs on SEE-600K and 40 epochs on SDE, using AdamW (Loshchilov & Hutter, 2019) with respective initial learning rates of 2×10^{-4} and 4×10^{-4}. Both settings use weight decay 10^{-2}, cosine annealing (Loshchilov & Hutter, 2017), and batch size 2. We use 128×256 crops on SEE-600K and 256×256 crops on SDE. Mixed-precision training (Micikevicius et al., 2018) is enabled, the random seed is 9527, and the loss follows Eq. 11. These settings are shared across the corresponding ablation variants unless otherwise stated.

Compared Methods and Protocol: Following SEE's task formulation (Lu et al., 2025), we evaluate broad-range restoration on SEE-600K and low-light enhancement on SDE. Baseline results are quoted from SEE Table 2 (SDE) and the SEE-trained block of Table 3 (SEE-600K), with EvLight++ from Tables 2 and 7 of Chen et al. (2024). These are complete-method comparisons with model-specific inputs and conditioning, including MambaSEE's three-pair history and reference-derived exposure prompt. Sources and reporting conventions appear in Appendix A.3; our inference and scoring procedures are specified in Appendix B.

Case	Spatial Block	Decoder	Temporal	Low Light			High Light			Normal Light		
				PSNR ↑	SSIM ↑	L_1 ↓	PSNR ↑	SSIM ↑	L_1 ↓	PSNR ↑	SSIM ↑	L_1 ↓
1	Swin	R1	–	16.37	0.5971	0.1288	17.02	0.6044	0.1111	18.03	0.7212	0.1030
2	GSSB	R1	–	19.35	0.6394	0.0969	20.40	0.6778	0.0733	25.29	0.8552	0.0391
3	GSSB	DWConv	–	19.98	0.6650	0.0885	20.55	0.6952	0.0711	26.23	0.8736	0.0343
4	GSSB	DWConv	FP32 GRU	20.69	0.6922	0.0806	**20.67**	0.6975	**0.0697**	**26.40**	**0.8771**	**0.0334**
5	GSSB	DWConv	W8A8 GRU	**20.78**	**0.6961**	**0.0794**	20.60	**0.6981**	0.0701	26.24	0.8761	0.0338

Table 3: Progressive module-level ablation on SEE-600K. R1 is the original SEENet decoder (Lu et al., 2025), and DWConv is our depthwise-convolutional decoder. "–" disables temporal aggregation. Case 4 retains the FP32 reference scores; Case 5 is the backward-then-forward W8A8 model in Table 1. Their comparison describes reported quality, not an isolated quantization effect. Bold marks the best result in each column.

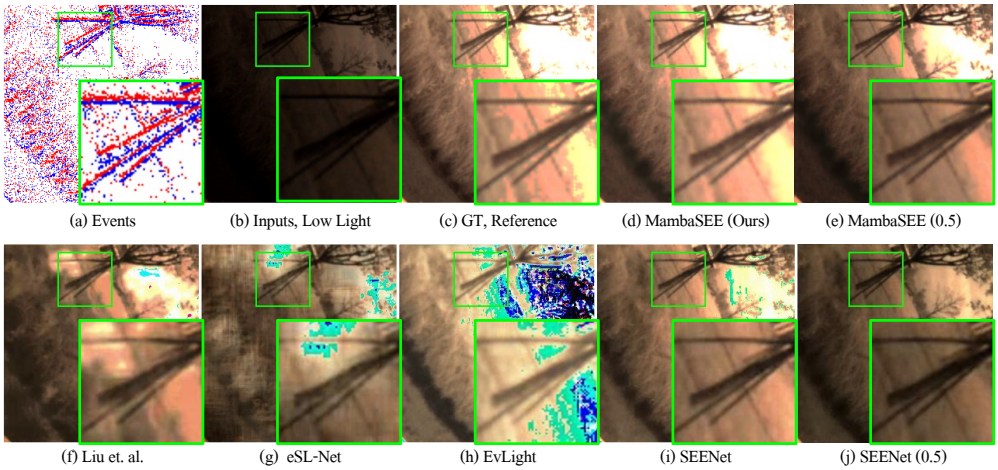


Figure 3: Comparison on SEE; "(0.5)" denotes a brightness prompt of $p = 0.5$.

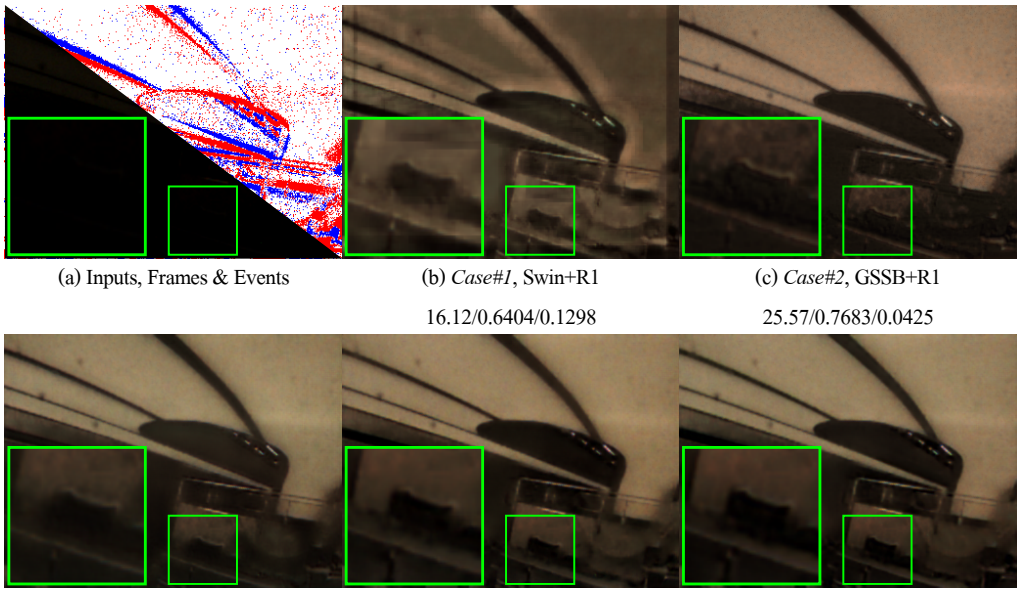
Comparison on SDE: Against the published SDE results in Table 2, MambaSEE has the highest tabulated average PSNR ($25.71dB$), exceeding SEENet by $2.14dB$ and Retinexformer, the highest-PSNR frame-only baseline, by $3.60dB$. Relative to SEENet, MambaSEE raises average SSIM from 0.7724 to 0.7841 while reducing FP32-equivalent storage by 74.7% ($1.90M$ to $0.48M$). Within the stated conditional streaming scope, temporal W8A8 reduces nominal BOPs by 41.53% relative to the all-FP32 variant. Cross-method BOP comparisons remain unavailable because the baselines do not report layer-wise precision.

Comparison on SEE-600K: Among the published SEE-trained benchmark results in Table 1, MambaSEE achieves the best reported PSNR, SSIM, and L_1 under all three illumination conditions. Relative to SEENet, PSNR gains are 2.01, 1.39, and 5.32 dB for low-, high-, and normal-light inputs, respectively. Improvements span both exposure extremes, with the largest PSNR margin under normal light.

4.1 Ablation Studies

Table 3 presents a progressive module-level ablation on SEE-600K. Each step evaluates a module replacement or addition in the current configuration, rather than isolating the effects of individual mechanisms. Fig. 4 illustrates the corresponding restoration results.

Spatial-block replacement: Replacing Swin with the full-precision GSSB while retaining the R1 decoder (Case 2) increases average PSNR and SSIM by 4.54 dB and 0.0832 and reduces L_1 by 0.0445. PSNR gains are 2.98, 3.38, and 7.26 dB under low, high, and normal light, respectively, with all three metrics improving in every condition. This is the largest observed increment along the tested sequence; the benefit of the GSSB replacement extends beyond low-light enhancement.



(a) Inputs, Frames & Events (b) *Case#1*, Swin+R1 16.12/0.6404/0.1298 (c) *Case#2*, GSSB+R1 25.57/0.7683/0.0425

(d) *Case#3*, GSSB+DWConv 26.17/0.7863/0.0382 (e) *Case#4*, GSSB+DWConv+FP32GRU 29.69/0.8518/0.0251 (f) *Case#5*, GSSB+DWConv+W8A8GRU 29.92/0.8551/0.0243

Figure 4: Qualitative comparison of the configurations in Table 3. Green boxes indicate matched regions and their 2× zoomed views. Values denote PSNR/SSIM/(L_1).

Decoder replacement: Replacing R1 with the proposed depthwise decoder (Case 3) improves average PSNR and SSIM by 0.57 dB and 0.0205 and reduces L_1 by 0.0051. All three metrics improve under each lighting condition, with the largest PSNR gain of 0.94 dB under normal light. Thus, within the GSSB-based configuration, the lightweight decoder improves reconstruction without sacrificing one illumination regime for another.

FIFO temporal module: Adding the FP32 FIFO ConvGRU (Case 4) improves average PSNR by 0.33 dB. The largest gain occurs in low light: PSNR rises by 0.71 dB, SSIM by 0.0272, and L_1 falls by 0.0079. High- and normal-light PSNR improve by 0.12 and 0.17 dB, respectively. The stronger low-light improvement is consistent with the motivation to supplement less reliable frame cues through temporal processing.

Quantized serial temporal model: The final W8A8 model (Case 5) reaches 22.54 dB, 0.7568 SSIM, and 0.0611 L_1. Relative to the FP32 reference, average PSNR differs by -0.05 dB, SSIM by $+0.0012$, and L_1 by -0.0001; PSNR differences are $+0.09$, -0.07, and -0.16 dB under low, high, and normal light. These small differences leave the reported frame-level quality close to FP32, with lower FP32-equivalent weight storage and nominal BOPs under the stated accounting.

5 Conclusion

MambaSEE addresses event-guided broad-range brightness restoration through a compact division of spatial modeling, temporal aggregation, and image reconstruction. Fixed-width additive fusion and full-precision GSSBs capture global context, while an INT8 quantization-aware FIFO ConvGRU aggregates historical features and an exposure-conditioned depthwise decoder recovers local details. MambaSEE combines 0.48M FP32-equivalent weight storage with the highest reported average PSNR on SDE and the best reported PSNR, SSIM, and L_1 across SEE-600K illumination conditions among the compared methods. These results demonstrate a favorable accuracy–storage trade-off for event-guided restoration across diverse illumination conditions.

AI USE STATEMENT

Generative AI tools assisted with language editing and LaTeX formatting. The authors reviewed and verified all AI-assisted material and take responsibility for the final text, claims, and artifacts in this work.

REPRODUCIBILITY STATEMENT

The experimental setup in the main paper specifies the datasets, official splits, evaluation metrics, model configuration, optimization settings, random seed, and complexity accounting used in our experiments. The supplementary material provides additional dataset, preprocessing, input-construction, and notation details. Appendix A.3 documents the numerical sources and protocol boundaries of the quoted baseline comparisons. Appendix B specifies the temporal and decoder paths, quantizer scale computation, routing behavior, image evaluation settings, and the boundary of the cached complexity estimates.

REFERENCES

Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9157–9167, 2021.

Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *International Conference on Learning Representations*, 2016.

Jarosław Bernacki. Automatic exposure algorithms for digital photography. *Multimedia Tools and Applications*, 79(19):12751–12776, 2020.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12504–12513, 2023.

Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5972–5981, 2022.

Kanghao Chen, Guoqiang Liang, Hangyu Li, Yunfan Lu, and Lin Wang. Evlight++: Low-light video enhancement with an event camera: A large-scale real-world dataset, novel method, and more. *arXiv preprint arXiv:2408.16254*, 2024.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

Mengyao Cui, Zhigang Wang, Dong Wang, Bin Zhao, and Xuelong Li. Color event enhanced single-exposure hdr imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 2, pp. 1399–1407, 2024.

Peijie Dong, Lujun Li, Zimian Wei, Xin Niu, Zhiliang Tian, and Hengyue Pan. EMQ: Evolving training-free proxies for automated mixed precision quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17076–17086, 2023.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2020.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1780–1789, 2020.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-72649-1_13.

Hang Guo, Yong Guo, Yaohua Zha, Yulun Zhang, Wenbo Li, Tao Dai, Shu-Tao Xia, and Yawei Li. MambaIRv2: Attentive state space restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28124–28133, 2025a.

Xuejian Guo, Zhiqiang Tian, Yuehang Wang, Siqi Li, Yu Jiang, Shaoyi Du, and Yue Gao. ERetinex: Event camera meets retinex theory for low-light image enhancement. In *IEEE International Conference on Robotics and Automation*, pp. 3759–3765, 2025b. doi: 10.1109/ICRA55743.2025.11127340.

Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1730–1739, 2020.

Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6):1–12, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Ji-Ye Jeon, Xuan Truong Nguyen, Soojung Ryu, and Hyuk-Jae Lee. USDN: A unified sample-wise dynamic network with mixed-precision and early-exit. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 646–654, 2024.

R John Koshel. *Illumination Engineering: design with nonimaging optics*. John Wiley & Sons, 2012.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2599–2613, 2018.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23–33, 2024.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1833–1844, 2021.

Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10615–10625, 2023.

Lin Liu, Junfeng An, Jianzhuang Liu, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, Yan Feng Wang, and Qi Tian. Low-light video enhancement with synthetic event guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1692–1700, 2023.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yunfan Lu, Xiaogang Xu, Hao Lu, Yanlin Qian, Pengteng Li, Huizai Yao, Bin Yang, Junyi Li, Qianyi Cai, Weiyu Guo, and Hui Xiong. SEE: See everything every time—adaptive brightness adjustment for broad light range images via events. *arXiv preprint arXiv:2502.21120*, 2025.

John J McCann and Alessandro Rizzi. *The art and science of HDR imaging*. John Wiley & Sons, 2011.

Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. Multi-bracket high dynamic range imaging with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 547–557, 2022.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2019.

Mingzhu Shen, Feng Liang, Ruihao Gong, Yuhang Li, Chuming Li, Chen Lin, Fengwei Yu, Junjie Yan, and Wanli Ouyang. Once quantization-aware training: High performance extremely low-bit architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5340–5349, 2021.

Lei Sun, Yuhan Bao, Jiajun Zhai, Jingyun Liang, Yulun Zhang, Kaiwei Wang, Danda Pani Paudel, and Luc Van Gool. Low-light image enhancement using event-based illumination estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6667–6677, 2025a.

Zhijing Sun, Senyan Xu, Kean Liu, Runze Tian, Xueyang Fu, and Zheng-Jun Zha. EVDM: Event-based real-world video deblurring with mamba. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13793–13803, 2025b.

Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in hdr deghosting: A survey and evaluation. In *Computer Graphics Forum*, number 2, pp. 683–707. Wiley Online Library, 2015.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *Computer Vision–European Conference on Computer Vision 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 155–171. Springer, 2020.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2654–2662, 2023. doi: 10.1609/aaai.v37i3.25364.

Xicong Wang, Huiyuan Fu, Jiaxuan Wang, Xin Wang, Heng Zhang, and Huadong Ma. Exploring in extremely dark: Low-light video enhancement with real events. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4805–4813, 2024.

Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex C. Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2604–2612, 2022a.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17683–17693, 2022b.

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.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Jiangwei Weng, Zhiqiang Yan, Ying Tai, Jianjun Qian, Jian Yang, and Jun Li. MambaLLIE: Implicit retinex-aware low light enhancement with global-then-local state space. In *Advances in Neural Information Processing Systems*, volume 37, pp. 27440–27462, 2024a.

Jingchong Weng, Boyang Li, and Kai Huang. Event-based image enhancement under high dynamic range scenarios. In *Proceedings of the Asian Conference on Computer Vision*, pp. 2456–2470, 2024b.

Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2022.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision*, pp. 3–19, 2018.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17714–17724, 2022.

Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13924–13934, 2023.

Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pp. 771–785. Springer, 2012.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022.

Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *Computer Vision–European Conference on Computer Vision 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pp. 666–682. Springer, 2020.

Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1632–1640, 2019. doi: 10.1145/3343031.3350926.

Haoxian Zhou, Chuanzhi Xu, Langyi Chen, Pengfei Ye, Haodong Chen, Qiang Qu, Ali Anaissi, and Weidong Cai. EvLIR: Learning illumination residuals from ordered events for low-light image enhancement. *arXiv preprint arXiv:2606.29430*, 2026.

Shiyang Zhou, Haijin Zeng, Yunfan Lu, Tong Shao, Ke Tang, Yongyong Chen, Jie Liu, and Jingyong Su. Binarized mamba-transformer for lightweight quad bayer HybridEVS demosaicing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8817–8827, 2025.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62429–62442, 2024.

A Additional Dataset Details

This appendix complements the experimental setup in the main paper. We separate properties of the original benchmarks from the protocol used by MambaSEE so that the reported results can be reproduced without changing the official scene splits.

Dataset	Task	Sensor / resolution	Scale	Official split	MambaSEE target
SEE-600K	Broad-range brightness adjustment	Event camera; 346×260	610,126 images with corresponding events from 202 scenes	165 training and 37 test scenes	Normal-light RGB image at the newest FIFO time
SDE	Event-guided low-light enhancement	DAVIS346; 346×260	More than 30K paired samples from 91 sequences	76 training and 15 test sequences	Normal-light RGB image paired with the newest low-light input

Table 4: Summary of the two real-world benchmarks. Dataset scale and splits follow the official SEE-600K and SDE protocols.

A.1 SEE-600K

Dataset scope and acquisition. SEE-600K (Lu et al., 2025) was introduced for event-guided brightness adjustment beyond the conventional low-to-normal-light setting. It contains 610,126 images and corresponding event measurements from 202 real-world scenes. Each scene is recorded under approximately four illumination conditions on average, spanning more than a $1000\times$ change in illumination. The benchmark consequently contains low-, normal-, and high-light observations instead of treating every input as underexposed. Recordings from different illumination conditions are aligned using a programmable robotic arm and IMU-assisted registration. The reported registration accuracy is approximately 1 ms temporally and 0.3 pixels spatially.

Split and mappings. We use the official scene-level split. Thirty-seven scenario groups, comprising 15 indoor and 22 outdoor groups, are reserved for testing; the remaining 165 groups are used for training. No test scene is used for optimization. Training and evaluation use the three mappings reported in the main paper: low-to-normal, high-to-normal, and normal-to-normal. This protocol evaluates both correction of severe exposure errors and normal-exposure target reconstruction for normal-light inputs.

MambaSEE input construction. RGB frames are converted to floating point and normalized to $[0, 1]$. For an anchor time i, the loader forms the causal FIFO sequence $\{i-2, i-1, i\}$ from adjacent RGB frames. The event packets indexed $i-3, \ldots, i$ are concatenated before voxelization. Raw events are represented as tuples (t, x, y, σ), sorted by timestamp, and accumulated into equal-width, half-open temporal bins over the observed timestamp range. Event polarity is mapped to $\{-1, +1\}$ before accumulation, and each temporal bin is sign-binarized to $\{-1, 0, +1\}$. The model receives 192 ordered event channels and reshapes them into three consecutive 64-channel slices of this joint window representation, not three independently voxelized frame intervals. During testing, missing history at the beginning of a sequence is replaced by the current sample for both frame and event-packet indices. With zero-based file indexing, training anchors start at 3 and test anchors at 1; both advance by 50 and exclude the last file index. This anchor stride does not change the adjacent frame spacing within a FIFO window. Training uses 128×256 crops.

Supervision and evaluation. Only the output at the newest FIFO time is supervised and evaluated. The reference is the registered normal-light RGB frame at that time. As in the SEE protocol, the scalar exposure prompt is computed from the mean intensity of this reference and is therefore an oracle target-exposure descriptor. We report PSNR, SSIM, and mean L_1 error separately for low-, high-, and normal-light inputs.

A.2 SDE

Dataset scope and acquisition. SDE (Liang et al., 2024) is a real-world event-image benchmark for low-light enhancement. It contains more than 30K aligned low-/normal-light image-event pairs from 91 sequences, including 43 indoor and 48 outdoor sequences. The data are captured at 346×260 with a DAVIS346 event camera. A robotic arm repeats a prescribed nonlinear trajectory under low and normal illumination, enabling spatially matched recordings. The original benchmark reports sub-0.03 mm trajectory alignment and temporal matching errors below 0.01 s for 90% of the paired data.

Split and task. We follow the official split of 76 training and 15 test sequences. Indoor and outdoor test subsets are evaluated separately. Each input consists of a low-light RGB frame and the low-light event stream associated with its exposure interval; the paired normal-light RGB frame is the restoration target. We do not use the synthetic SDSD extension of SDE.

MambaSEE input construction. For each output time, three causal low-light frame-event pairs are selected. Each raw event interval is converted independently to a 64-channel signed voxel grid using the same chronological binning and polarity accumulation described above, after which the three grids are concatenated to 192 channels. The target is the normal-light frame corresponding to the newest input. Sequence-boundary samples retain their original output timestamps by replacing unavailable history with the current sample. RGB values are normalized to $[0, 1]$, and training uses 256×256 crops.

Supervision and evaluation. The target mean supplies the oracle exposure prompt used by the decoder. Following the official indoor/outdoor evaluation, we report PSNR and SSIM without introducing a new data split or a brightness-fitting post-process.

A.3 Baseline Comparison Protocol

Quantitative provenance. The baseline entries in the main SDE and SEE-600K tables are quoted from two published studies. We use arXiv:2502.21120v2 of Lu et al. (2025) and arXiv:2408.16254v1 of Chen et al. (2024). In Table 5, S2 denotes SEE Table 2 (SDE), S3 denotes the SEE-trained block of SEE Table 3 (SEE-600K), and E2/E7 denote the quality and complexity results in EvLight++ Tables 2/7. The SDE-trained block of SEE Table 3 is not used for the SEE-600K comparison. Retraining and reproduction described by Lu et al. (2025) were performed in that benchmark study, not as a unified baseline rerun for the quantitative tables in this work. Where the papers report different SDE-out PSNR values for a shared baseline (Liu et al. and EvLowLight), we consistently use S2.

Input and conditioning boundaries. MambaSEE uses a causal FIFO window of three frame–event pairs and evaluates only the newest-frame output. On both datasets, the brightness prompt is computed from the mean intensity of the normal-exposure reference image.

Complexity and storage. Baseline FLOPs and parameter counts are transcribed from S2 or E7, not reprofiled by us. MambaSEE's conditional steady-state estimate uses 256×256 inputs, counting one multiply–accumulate as one operation and assuming spatial-feature reuse across consecutive FIFO windows. In the FP32-equivalent storage column, baseline parameter counts are treated as FP32 weight-storage estimates, not measurements of packed checkpoints. MambaSEE instead uses its stated mixed-precision accounting: $0.73M$ trainable parameters, including $0.32M$ W8 weights, correspond to $0.48M$ FP32-equivalent storage.

B Implementation and Evaluation Specifications

The following settings specify the evaluated non-binary GSSB model with a W8A8 temporal path. They complement the main equations with the choices needed to reconstruct the computation and evaluation protocol.

Method	Source	Reported observations	Protocol note
SNR, UFormer, LLFlow, Retinexformer	S2	Image-only methods	LLFlow attribution follows S2.
E2VID+	S2	Events only	E2 identifies its evaluation as grayscale.
ELIE	S2	Uses events; exact history NR	No common history budget is inferred.
HDRev	S2	Event-guided HDR reconstruction	S2 reports adaptation to SDE.
Wang et al.	S2	Single-frame/event input in SEE	Exact history and prompt settings NR.
eSL-Net, Liu et al., EvLight	S2, S3	Single-frame/event input in SEE	Exact history and prompt settings NR.
EvLowLight	S2, S3	Multi-frame/event input in SEE	Not assumed to use MambaSEE's three-pair window.
SEENet	S2, S3	Frame–event pair with a brightness prompt	Released normal-target prediction uses the reference mean.
EvLight++	E2, E7	Video and events	Ordinary PSNR/SSIM from E2; Average is the mean of its reported indoor and outdoor scores. Costs at 256×256 from E7.
MambaSEE	This work	Three causal RGB frame–event pairs	Reference-mean prompt; newest output evaluated.

Table 5: Sources and reported input categories for the quantitative comparisons. NR means not fully specified in the benchmark reporting used here, not that a method lacks the corresponding capability. Source identifiers refer to tables in the cited papers.

B.1 Network Configuration

Embeddings and widths. The shared feature width is 64; each GSSB projects internally to 128 channels and uses selective-scan step-size rank 8. In both learned coordinate and exposure embeddings, the first 1×1 projection is concatenated with the original input before the second 1×1 projection. The exposure embedding additionally uses ReLU.

Temporal states and refinement. Each three-frame window starts with zero forward and reverse hidden states; states are not carried between windows. The ConvGRU uses 3×3 biased convolutions, 48 hidden channels, and reset gating before the candidate-state convolution. Its input is average-pooled with kernel and stride 2. Let $X_k = \mathrm{AvgPool}_2(T_k)$, and let $\mathcal{R}_f$ and $\mathcal{R}_b$ denote the independently parameterized forward and reverse cells. The reverse pass runs first, from i to $i - K + 1$. A bias-free W8A8 1×1 projection $\mathcal{P}_b$ maps its 48-channel states to 64 channels and conditions the corresponding forward inputs. The forward pass then runs from $i - K + 1$ to i:

$$\begin{aligned} T_k^b &= \mathcal{R}_b(X_k, T_{k+1}^b), \quad T_{i+1}^b = 0, \\ \widetilde{X}_k &= X_k + \mathcal{P}_b(T_k^b), \\ T_k^f &= \mathcal{R}_f(\widetilde{X}_k, T_{k-1}^f), \quad T_{i-K}^f = 0. \end{aligned} \tag{13}$$

States retain their original FIFO indices. Although $T_i^b = \mathcal{R}_b(X_i, 0)$, all reverse states, including the terminal T_{i-K+1}^b, feed the forward recurrence through $\mathcal{P}_b$ and can affect T_i^f. Thus the newest prediction receives reverse context from every window position, rather than only the direct T_i^b branch. The states are restored by bilinear interpolation with `align_corners=False`. The reported prediction reads the updated feature H_i at the newest FIFO index. The resulting context Γ_k has $64 + 48 + 48 = 160$ channels. G uses $160 \to 64 \to 64$ pointwise projections with GELU between them and a final sigmoid. D uses a $160 \to 64$ pointwise projection, a 3×3 depthwise convolution, and a $64 \to 64$ pointwise projection, with GELU after the first two convolutions. Only the final projections in G and D have biases. Both the weight and bias of the final projection in D are zero-initialized.

Decoder. The four residual groups each contain two 5×5 depthwise blocks with dilations 1 and 2, respectively. Each block uses one-group normalization and $64 \to 128 \to 64$ pointwise channel

mixing. Group and global tails both use a 3×3 depthwise convolution, GELU, and a 1×1 projection. Decoder residual additions have unit coefficients, without learned residual scales. The input and RGB output projections are 3×3 convolutions.

B.2 Quantization and Routing

Scale computation. The evaluated temporal quantizer uses the full signed range $[-128, 127]$. For each quantized tensor or output-channel slice, let $x_{\min}$ and $x_{\max}$ be its detached extrema. The zero-point-free scale is

$$s(x) = \max\left\{\frac{\max(x_{\max}, 0)}{127}, \frac{\max(-x_{\min}, 0)}{128}, \varepsilon_{\text{dtype}}\right\}, \tag{14}$$

where $\varepsilon_{\text{dtype}}$ is the machine epsilon of the tensor's floating-point type. Weight extrema are reduced over the input-channel and kernel dimensions; activation extrema are reduced over the entire input tensor, including the batch dimension. Scales are recomputed on every forward call in both training and evaluation, not learned, calibrated in a separate pass, or frozen. Rounding uses a straight-through estimator. Each selected convolution fake-quantizes its input and weight before a floating-point convolution. Biases, convolution outputs, stored hidden states, and nonlinear operations are not separately quantized. These rules also apply to the inter-pass projection $\mathcal{P}_b$.

Spatial routing. Hard Gumbel-Softmax with temperature 1 is used in both training and evaluation; evaluation does not replace it with deterministic argmax routing. Token identifiers are sorted with `stable=False`, and the inverse permutation restores the spatial layout. The state prompt is formed before token sorting and passed to the scan without an additional permutation. Random generators are seeded once at launch with seed 9527, rather than reset for each window.

B.3 Optimization and Evaluation

Optimization. Both dataset runs start from random initialization, without an FP32 checkpoint warm start; temporal fake quantization is active from the first optimization step. AdamW uses one parameter group and its default momentum coefficients $(0.9, 0.999)$ and numerical epsilon 10^{-8}. Cosine annealing is stepped once per epoch, with a minimum learning rate of 10^{-8} and periods of 20 epochs on SEE-600K and 40 on SDE. Crop origins are snapped down to multiples of four pixels, with the same crop applied to all frames, events, and targets. The Charbonnier and gradient-magnitude losses each have unit weight and use only the newest normal-target prediction. The Charbonnier square root uses an additive stabilizer of 10^{-6}. Gradient magnitudes use adjacent-pixel forward differences over valid horizontal and vertical locations, without padding. No temporal difference loss or sampled-exposure loss is enabled.

Inference and image scoring. Inference uses batch size 1 with mixed precision enabled. SEE-600K is evaluated at its native 260×346 size; SDE uses the top-left 256×256 region, without resizing. The target-brightness prompt is computed from the reference region supplied to the model. Predictions and references are multiplied by 255, clipped to $[0, 255]$, cast to unsigned 8-bit integers, and saved as PNG images. The scoring scripts reload these images as RGB values divided by 255, without brightness fitting or additional border cropping. PSNR is computed per image from RGB mean squared error with an additive 10^{-10} stabilizer. SSIM uses an 11×11 Gaussian window with standard deviation 1.5, zero-padded local statistics, and constants 0.01^2 and 0.03^2; channel and spatial values are averaged. SEE-600K L_1 is the mean absolute RGB error.

Aggregation. On SEE-600K, image scores are averaged within each lighting condition, and the table's Average column equally averages the three condition scores. For MambaSEE on SDE, image scores are averaged within each subset; Average is the arithmetic mean of the displayed indoor and outdoor scores.

B.4 Complexity Accounting

At 256×256, complexity is estimated under steady-state reuse of historical spatial features, covering one new frame–event encoding, the complete temporal module, and one output decode. MACs count convolutions and linear projections, excluding routing products, selective-scan recurrence, and other

non-convolutional operations. BOPs use the assigned weight and activation bit widths. Accuracy is evaluated with full-window recomputation.

C Notation and Variable Definitions

The following tables collect the notation used throughout the Method section. Superscript k and subscript k both index a FIFO time; their placement follows the notation of the corresponding module. Unless stated otherwise, feature maps have spatial size $H \times W$ and channel width $C = 64$. The sequence/block index l is local to its subsection: it denotes a reordered token position in the GSSB recurrence and a residual block in the decoder.

Table 6: Input, output, and indexing notation.

Symbol	Shape / value	Definition
H, W	Positive integers	Spatial height and width of an RGB frame or event voxel grid.
i, k	Integer indices	i is the newest output time; $k \in \{i-K+1, \ldots, i\}$ indexes an observation in the FIFO window.
K	3	Number of causal frame-event pairs in one FIFO input.
I_k	$\mathbb{R}^{3\times H\times W}$	RGB input frame captured at time k.
E_k	$\mathbb{R}^{B_e\times H\times W}$	Event voxel slice associated with time k.
B_e	64	Number of event channels per FIFO time; the complete event input has $KB_e = 192$ channels.
I_{gt}	$\mathbb{R}^{3\times H\times W}$	Normally exposed reference image at the newest time i.
I_{pred}	$\mathbb{R}^{3\times H\times W}$	Image reconstructed by MambaSEE.
p	Scalar	Oracle target-brightness prompt, equal to the global mean of I_{gt}.
f_{see}	Mapping	Complete MambaSEE function from the FIFO observations and p to I_{pred}.
c, x, y	Indices	c indexes RGB channels; x, y are zero-based column and row indices, respectively.
$[\cdot,\cdot]$	Operator	Concatenation along the channel or token-feature dimension.

Table 7: Notation for modality embedding and the Mamba Additive Spatial Encoder (MASE).

Symbol	Shape / value	Definition
$P_0(x, y)$	$\mathbb{R}^4$ per pixel	Row-first coordinate descriptor containing Bayer parity $(y \bmod 2, x \bmod 2)$ and normalized position $(y/H, x/W)$.
P	$\mathbb{R}^{C\times H\times W}$	Learned positional embedding shared by the image and event heads.
$f_{\mathrm{i}}, f_{\mathrm{e}}$	Learned mappings	Modality-specific image and event heads, respectively.
$F_{\mathrm{i}}^k, F_{\mathrm{e}}^k$	$\mathbb{R}^{C\times H\times W}$	Position-aware image and event features at time k.
$\mathcal{C}_1, \mathcal{C}_2$	3×3 convolutions	Fixed-width fusion convolutions before and after the first pair of GSSBs.
ϕ	Activation	H-Swish in MASE; the decoder channel mixers use GELU.
$F^k, \widetilde{F}^k$	$\mathbb{R}^{C\times H\times W}$	Additively fused feature and its activated version.
$\mathcal{A}_{\mathrm{i}}, \mathcal{A}_{\mathrm{e}}$	GSSB mappings	Parallel GSSBs with independent parameters. Both receive the same fused feature; the subscripts name branches rather than single-modality inputs.
$Z_{\mathrm{i}}^k, Z_{\mathrm{e}}^k$	$\mathbb{R}^{C\times H\times W}$	Outputs of the two parallel GSSB branches.
Z_{ie}^k	$\mathbb{R}^{C\times H\times W}$	Mid-level additive fusion of the two branch outputs.
$\mathcal{A}_{\mathrm{ie}}$	GSSB mapping	Final GSSB applied after mid-level fusion.
T_k	$\mathbb{R}^{C\times H\times W}$	MASE output passed to the temporal module.
U	$\mathbb{R}^{L\times C}$	Flattened visual token sequence entering one GSSB.
L	HW	Number of spatial tokens in a frame.
C	64	Shared feature width used by MASE and the decoder.

Table 8: Notation inside a Global State-Space Block (GSSB).

Symbol	Shape / value	Definition
Ψ	Learned mapping	GSSB input projection together with its conditional positional gate.
$\rho = \{\rho_l\}_{l=1}^{L}$	State-prompt sequence	Semantic prompts produced by routing visual tokens to learned semantic tokens.
ρ_l	State-prompt vector	Routing-derived prompt that modulates the state readout at reordered position l.
π	Permutation	Token order induced by the hard Gumbel-Softmax semantic router.
Sort_π, Unsort_π	Operators	Apply the semantic permutation and restore the original spatial order.
U_π	$\mathbb{R}^{L\times 2C}$	Projected tokens arranged in semantic scan order; the internal width is $2C = 128$.
V_π	$\mathbb{R}^{L\times C}$	Prompted SSM output after normalization and output projection.
$\mathcal{W}_o$	Learned projection	Output projection of the selective state-space path.
LN	Operator	Layer normalization.
$u_l = (U_\pi)_l$	Token vector	Input token at reordered sequence position l.
h_l	State vector	Selective SSM hidden state after processing u_l.
$\bar{A}_l, \bar{B}_l$	SSM parameters	Input-dependent discretized state-transition and input projections.
c_l	SSM parameter	Input-dependent readout vector at sequence position l; it is unrelated to the RGB-channel index c.
D_{ssm}	SSM parameter	Direct skip coefficient in the state-space readout.
o_l	Token vector	Output token produced by the prompted selective state-space update.
$\mathrm{SSM}(U_\pi;\rho)$	$\mathbb{R}^{L\times 2C}$	Ordered output sequence before the projection back to width C.
$\mathcal{A}(U)$	$\mathbb{R}^{L\times C}$	Residual GSSB output after inverse permutation.

Table 9: Notation for FIFO temporal aggregation and quantization-aware training. The ConvGRU hidden states and nonlinearities remain floating point; W8A8 applies to the selected temporal convolutions.

Symbol	Shape / value	Definition
X_k, $\widetilde{X}_k$	$\mathbb{R}^{C\times\lfloor H/2\rfloor\times\lfloor W/2\rfloor}$	Pooled spatial feature and backward-conditioned forward input.
$\mathcal{P}_b$	$48\to 64$, 1×1	Bias-free W8A8 projection from reverse state to forward input.
T_k^f, T_k^b	Recurrent feature maps	Time-aligned states of the serial reverse-then-forward passes in Eq. 13.
Up	Operator	Bilinear interpolation that restores a recurrent state to $H \times W$.
Γ_k	Concatenated feature	Context formed from T_k, $\mathrm{Up}(T_k^f)$, and $\mathrm{Up}(T_k^b)$.
G, D	Learned mappings	Temporal modulation-gate and residual-update branches. D here is unrelated to D_{ssm}.
g_k	$\mathbb{R}^{C\times H\times W}$	Learned modulation gate generated from Γ_k.
Δ_k	$\mathbb{R}^{C\times H\times W}$	Candidate temporal residual generated by the zero-initialized delta branch.
H_k	$\mathbb{R}^{C\times H\times W}$	Temporally refined feature; only H_i is decoded for the reported output.
$\odot$	Operator	Element-wise multiplication.
x	Tensor	Floating-point weights or activations presented to fake quantization.
s	Positive scalar/tensor	Quantization scale; per-output-channel for weights and per-tensor for activations.
$q(x;s)$	Integer-valued tensor	Rounded and clipped signed 8-bit code.
$Q(x;s)$	Floating-point tensor	Dequantized fake-quantization output used during QAT.
$[-128, 127]$	Integer range	Full signed INT8 range used by the evaluated temporal quantizer.

Table 10: Notation for exposure-conditioned reconstruction and the training objective.

Symbol	Shape / value	Definition
B	$\mathbb{R}^{C\times H\times W}$	Exposure embedding generated from the scalar prompt p; it is unrelated to the event width B_e.

Table 10 (continued)

Symbol	Shape / value	Definition
$\mathcal{C}_{\text{in}}$	3×3 convolution	Decoder input projection applied to $H_i + B$.
R_0	$\mathbb{R}^{C \times H \times W}$	Initial decoder feature after input projection.
$R_l, \widetilde{R}_l$	$\mathbb{R}^{C \times H \times W}$	Input to decoder block l and its normalized depthwise-filtered feature.
$\mathcal{C}_D^l$	5×5 depthwise convolution	Local spatial operator in decoder block l; it is distinct from the temporal branch D.
$\mathcal{C}_{\text{pw},1}^l, \mathcal{C}_{\text{pw},2}^l$	1×1 convolutions	Channel expansion and channel restoration in decoder block l.
GN	Operator	Group normalization with one group in the implementation.
l	Block index	Index of a residual decoder block.
$\mathcal{L}$	Scalar	Total training objective.
N	$3HW$	Number of scalar RGB values in one output image.
ϵ_n	Scalar	Prediction residual at flattened RGB index n.
δ_d	Difference tensor	Difference between prediction and target gradient magnitudes along direction d.
d	x or y	Horizontal or vertical finite-difference direction.
∇_d	Operator	Image gradient along direction d.
$\|\cdot\|_{1,\text{mean}}$	Scalar operator	Mean absolute value over all elements.

D Additional Qualitative Results on SEE-600K

Figures 5 and 6 complement the direct method comparison in the main paper with scene-diverse results on challenging SEE-600K inputs. The eight examples come from different test scenes and span low-to-normal and high-to-normal correction in indoor and outdoor environments. They also cover varied content, including fine text, resolution patterns, nighttime structures, and repetitive geometry. Normal-exposure target reconstruction for normal-light inputs is evaluated quantitatively in the main paper.

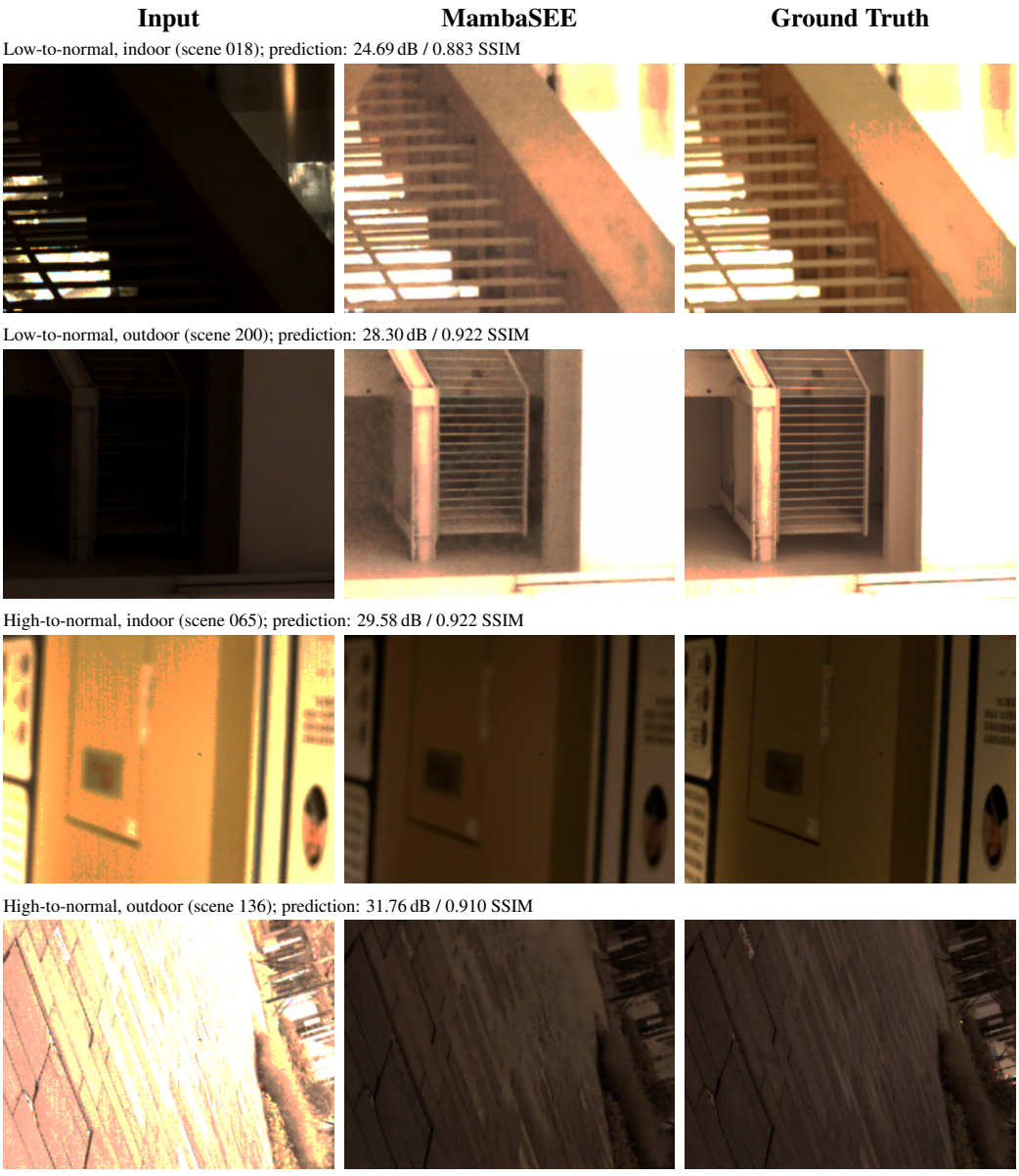


Figure 5: Additional qualitative results on SEE-600K. Each row shows the input, MambaSEE prediction, and normally exposed ground truth. The rows cover low-to-normal indoor, low-to-normal outdoor, high-to-normal indoor, and high-to-normal outdoor cases, respectively. Prediction PSNR/SSIM values are given in the row labels.

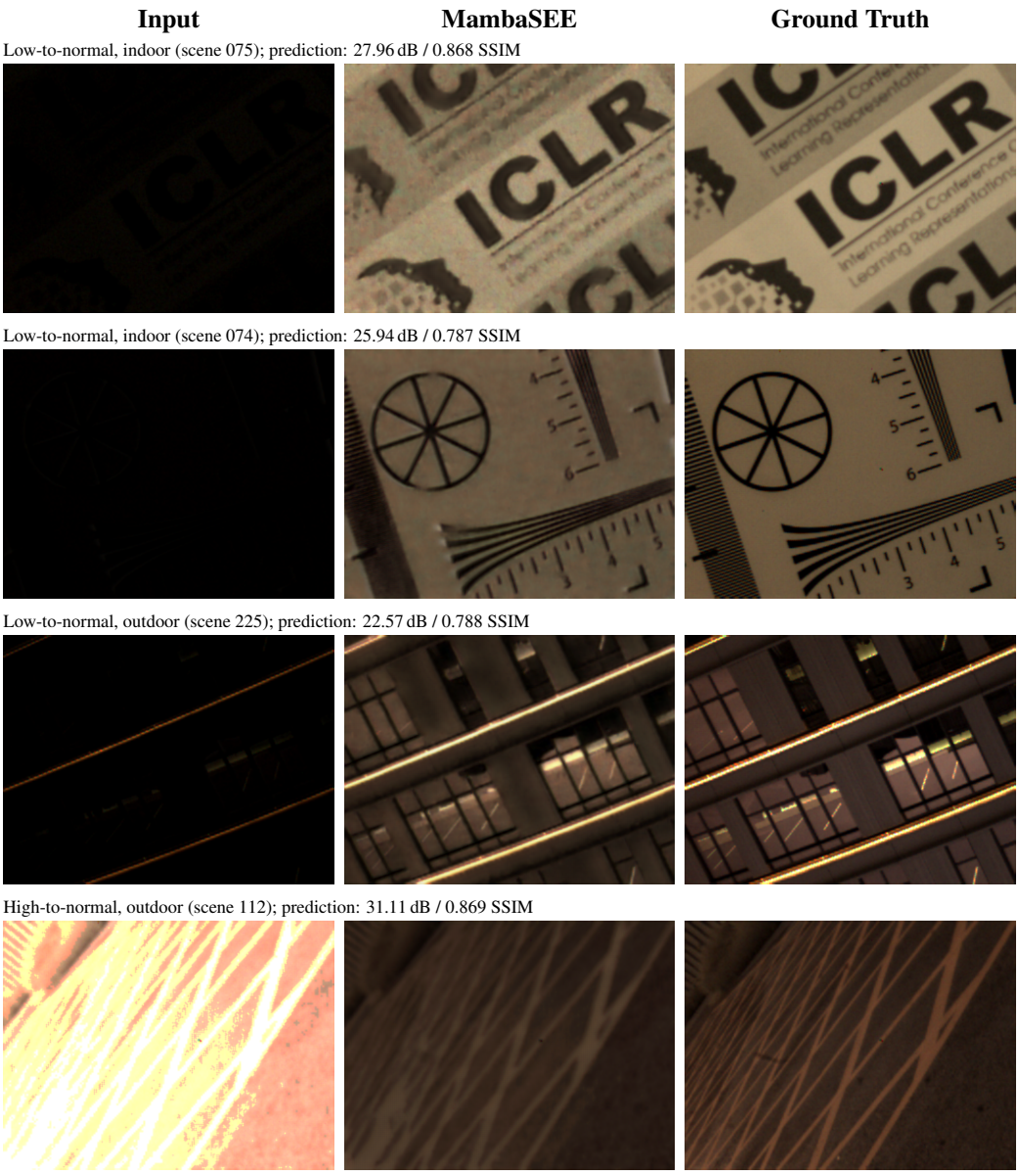


Figure 6: Further qualitative results from four additional SEE-600K test scenes. The examples emphasize recovery of printed text, resolution patterns, nighttime outdoor structures, and repetitive geometry. The layout follows Fig. 5; all eight displayed examples are drawn from different test scenes.

For severely underexposed inputs, MambaSEE recovers the principal scene layout and structural boundaries from observations in which little content is directly visible. For overexposed inputs, it suppresses saturation and moves the global brightness and surface appearance toward the normal-light reference. The results illustrate that the same compact model handles two opposite exposure corrections across indoor and outdoor scenes without post-processing. Remaining differences are mainly local smoothing and small color deviations in strongly saturated or texture-rich regions.