# Appendix

## A. Additional Related Work

**Mamba Models.** The emergence of mamba [22, 105] has catalyzed a new wave of innovation in low-light image enhancement (LLIE), with new architectures often building upon established theoretical frameworks. A significant portion of initial models are retinex inspired, integrating Mamba as a powerful and efficient engine for image decomposition and restoration. Pioneering works like Retinex-Mamba [5] demonstrated mamba's viability by replacing computationally heavy transformer blocks to achieve greater efficiency. This was extended by LLEMamba [98], which embeds mamba within a deep unfolding network corresponding to an ADMM optimization algorithm, making the enhancement process both interpretable and context-aware. Other models like MambaLLIE [86] address the "local pixel forgetting" artifact of 1-D scanning by augmenting their state-space modules with local convolutions. Meanwhile, EffRetMamba [99] prioritizes speed by using jump-sampling to process a shorter sequence of image tokens, trading some information fidelity for a significant boost in inference time.

A second category of Mamba-based models moves beyond simple backbone replacement to propose novel interaction and learning paradigms. The most radical of these is BSMamba [97], which discards conventional spatial scanning entirely. Instead, it reorders image patches based on brightness and semantic content before the 1-D scan, allowing the model to establish long-range dependencies between functionally related but spatially distant pixels. Other models focus on alternative enhancement strategies or learning frameworks. ResVMUNetX [78], for example, opts for maximum efficiency by training a Mamba network to predict a simple additive residual map, enabling real-time video enhancement. Semi-LLIE [44] tackles the practical challenge of data scarcity, using a semi-supervised framework with a mamba backbone to effectively leverage vast amounts of unlabeled data. These diverse approaches highlight the versatility of the Mamba architecture and underscore a key trend: the most successful models are not naive, drop-in replacements, but are thoughtfully designed to synergize mamba's strengths with domain-specific knowledge and innovative learning strategies.

**Transformer Models.** Transformer-based methods have been widely used for low-light enhancement due to their ability to model long-range dependencies. LLFormer [80] uses global attention to handle illumination inconsistencies. IAT [13] adapts lightweight transformer blocks for exposure correction. IPT [9] employs multi-task pretraining across restoration tasks using a shared transformer backbone. Fourmer [104] integrates Fourier transforms for improved global modeling. LYT-Net [6] leverages the YUV color space for resource-efficient enhancement. MAXIM [73] introduces multi-axis gated MLPs for spatial and channel modeling. While effective, these models often incur high memory usage and quadratic runtime, making them less suited for real-time or mobile deployment.

Table 7. Descriptions of two variants of our model, **s**' and **l**', representing small and large model configurations.

| Model Type | Configuration | | | | Inference | Memory |
|---|---|---|---|---|---|---|
| | base channel | patch size | depth | params | speed | consumption |
| ExpoMamba$_s$ | 48 | 4 | 1 | 41 M | 36 ms | 2923 Mb |
| ExpoMamba$_l$ | 96 | 6 | 4 | 166 M | 95.6 ms | 5690 Mb |

**Diffusion Models.** Diffusion models have shown great potential in generating realistic and detailed images. The ExposureDiffusion model [82] integrates a diffusion process with a physics-based exposure model, enabling accurate noise modeling and enhanced performance in low-light conditions. Pyramid Diffusion [102] addresses computational inefficiencies by introducing a pyramid resolution approach, speeding enhancement without sacrificing quality. [65] handles image-to-image tasks using conditional diffusion processes. Models like [96] and deep non-equilibrium approaches [60] aim to reduce sampling steps for faster inference. However, starting from pure noise in conditional image restoration tasks remains a challenge for maintaining image quality while cutting down inference time [25].

**Hybrid Modeling.** Hybrid models include learning features in both spatial and frequency domains have been another popular area in image enhancement/restoration tasks. It has been predominantly explored in three subcategories: **(1)** Fourier Transform [93], Fourmer [104], FD-VisionMamba [101]; **(2)** Wavelet Transform [10, 70, 106]; and, **(3)** Homomorphic Filtering [4]. Such methods demonstrate that leveraging both spatial and frequency information can significantly improve enhancement performance. Recent hybrid models such as MAXIM [73] and PromptIR [61] have explored lightweight but flexible design spaces for image restoration and enhancement tasks.

## B. Extended Details on Phase Manipulation

This section provides a step-by-step derivation of how a uniform phase shift in the Fourier domain affects an image in the spatial domain.

We begin with the definition of the 2D Fourier Transform for an image $I(x, y)$:

$$\mathbf{F}(u, v) = \mathcal{F}\{I(x, y)\} = \iint_{-\infty}^{\infty} I(x, y) e^{-i2\pi(ux+vy)}\, dx\, dy \tag{10}$$

The transform can be represented in polar form using its amplitude $\mathbf{A}(u, v)$ and phase $\phi(u, v)$:

$$\mathbf{F}(u, v) = \mathbf{A}(u, v) e^{i\phi(u,v)} \tag{11}$$

The original image $I(x, y)$ is recovered via the Inverse Fourier Transform:

$$I(x, y) = \mathcal{F}^{-1}\{\mathbf{F}(u, v)\} = \iint_{-\infty}^{\infty} \mathbf{F}(u, v) e^{i2\pi(ux+vy)}\, du\, dv \tag{12}$$

Substituting the polar form into the inverse transform gives:

$$I(x, y) = \iint_{-\infty}^{\infty} \mathbf{A}(u, v) e^{i\phi(u,v)} e^{i2\pi(ux+vy)}\, du\, dv \tag{13}$$

Our goal is to analyze the effect of applying a uniform phase shift, $\Delta\phi$, to the phase component, resulting in a new phase $\phi'(u, v) = \phi(u, v) + \Delta\phi$. The modified Fourier spectrum is $\mathbf{F}'(u, v) = \mathbf{A}(u, v)e^{i(\phi(u,v)+\Delta\phi)}$. The corresponding image in the spatial domain, $I'(x, y)$, is derived as follows.

$$I'(x, y) = \iint \mathbf{A}(u, v) e^{i(\phi(u,v)+\Delta\phi)} e^{i2\pi(ux+vy)}\, du\, dv \tag{14}$$

$$= \iint \mathbf{A}(u, v) e^{i\phi(u,v)} e^{i\Delta\phi} e^{i2\pi(ux+vy)}\, du\, dv \tag{15}$$

$$= e^{i\Delta\phi} \iint \mathbf{A}(u, v) e^{i\phi(u,v)} e^{i2\pi(ux+vy)}\, du\, dv \tag{16}$$

$$= e^{i\Delta\phi} \cdot I(x, y) \tag{17}$$

$$= (\cos(\Delta\phi) + i \sin(\Delta\phi)) \cdot I(x, y) \tag{18}$$

$$= \cos(\Delta\phi) I(x, y) + i \sin(\Delta\phi) I(x, y) \tag{19}$$

In this derivation: Eq. (15) uses the exponential identity $e^{a+b} = e^a e^b$; Eq. (16) factors out the constant phase shift $\Delta\phi$ from the integral; Eq. (17) recognizes the inverse Fourier transform of the original image; Eq. (18) applies Euler's formula $e^{i\theta} = \cos\theta + i \sin\theta$. Finally, Eq. (19) shows that a uniform phase shift in the frequency domain results in a complex-valued image where the real and imaginary parts are scaled versions of the original. This final expression (19) demonstrates that a uniform phase shift in the frequency domain results in multiplying the original image $I(x, y)$ by a complex constant $e^{i\Delta\phi}$. The new image $I'(x, y)$ is a complex-valued signal where the real part is the original image scaled by $\cos(\Delta\phi)$, and the imaginary part is the original image scaled by $\sin(\Delta\phi)$. This rotation in the complex plane fundamentally alters the image's spatial characteristics, underscoring the critical role of phase information in defining structural content.

## C. Additional Human Evaluation Results

As an extension to the main paper's perceptual study, we present qualitative results and statistical rankings from our human evaluation conducted on the SICE test dataset. Participants rated multiple image enhancement models across seven visual attributes: *naturalness*, *brightness*, *colorfulness*, *reduced blur*, *more details*, *noise reduction*, and *contrast*. In the spider charts (Fig. 7), ExpoMamba consistently outperforms competing methods across all criteria, with es-

Figure 7. Additional results from the SICE test set for human evaluation (continued from Fig. 6), along with corresponding spider charts and average rank ratings.

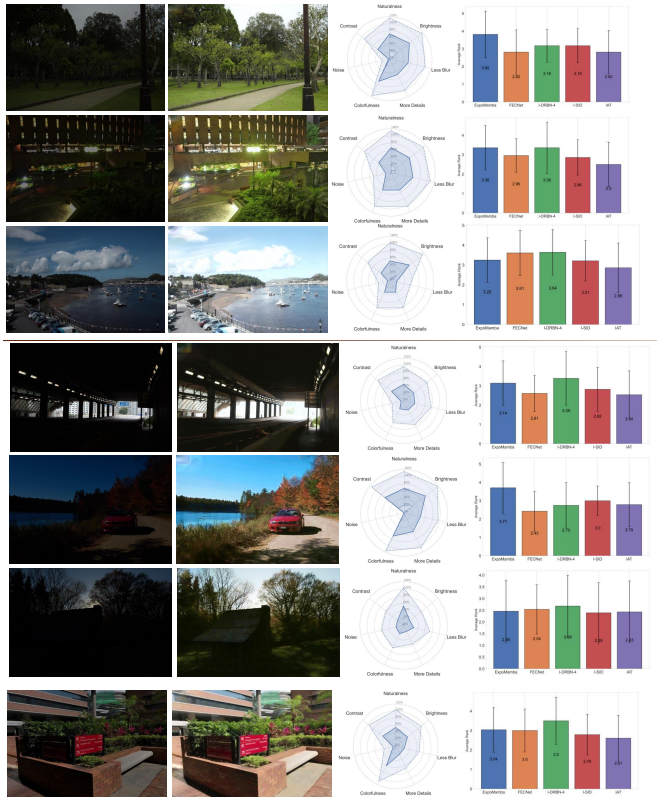
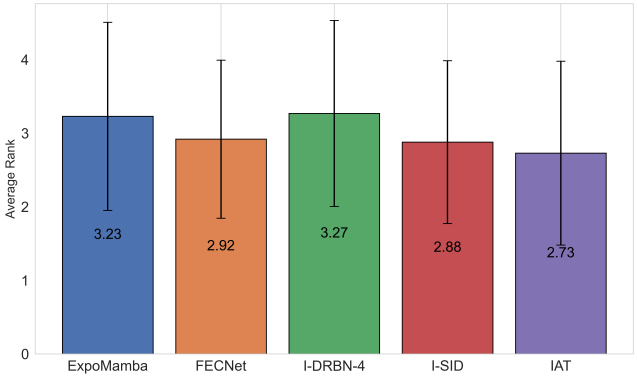


Figure 8. Aggregate rank comparison of ExpoMamba and baselines from human perceptual evaluation. Higher values indicate better subjective quality; error bars represent standard deviation.



pecially strong scores in the *noise reduction*, *naturalness*, and *detail retention* dimensions.

The aggregated average ranks across all images and participants are summarized in Fig. 8. ExpoMamba receives the highest overall perceptual ranking, demonstrating its ability to deliver visually balanced and subjectively preferred results. The small standard deviation further reflects strong inter-rater consistency, reinforcing the perceptual robustness of the proposed method.

## D. Frequency-Dependent Dynamic Adaptation.

Let $r(u,v) = \frac{\sqrt{u^2+v^2}}{r_{\max}}$ denote the normalized radial frequency. We parameterize frequency dependent state matrices via shallow gating functions $\alpha_A(r)$, $\alpha_B(r)$, $\alpha_C(r)$ produced by a 1 layer MLP with Softplus, shared across channels:

$$\alpha_\bullet(r) = \log\Big(1 + \exp\Big(\mathbf{w}_{\bullet 2}\ \mathrm{ReLU}\big(\mathbf{w}_{\bullet 1}\, r + b_{\bullet 1}\big) + b_{\bullet 2}\Big)\Big).$$

where, $\bullet \in \{A, B, C\}$. For each step $t$ and frequency bin $(u, v)$, we form

$$\begin{aligned} A_t(u,v) &= A_t \circ \big(1 + \alpha_A(r(u,v))\big), \\ B_t(u,v) &= B_t \circ \big(1 + \alpha_B(r(u,v))\big), \\ C_t(u,v) &= C_t \circ \big(1 + \alpha_C(r(u,v))\big). \end{aligned}$$

Here, $\circ$ denotes element wise multiplication. This keeps the SSM core intact while allowing smooth low to high frequency reweighting with three learnable gates. After separate processing through state-space models, the modified amplitude $\mathbf{A}''(\mathbf{u}, \mathbf{v})$ and phase $\mathbf{P}''(\mathbf{u}, \mathbf{v})$ are recombined and transformed back into the spatial domain to reconstruct the enhanced image: $\mathbf{I}'(\mathbf{x}, \mathbf{y}) = \mathbf{F}^{-1}(\mathbf{A}''(\mathbf{u}.\mathbf{v}) + i \cdot \mathbf{P}''(\mathbf{u}.\mathbf{v}))$; where, $\mathbf{F}^{-1}$ denotes the inverse Fourier Transform.

## E. Dynamic Patch Training

To improve robustness to variable input resolutions, ExpoMamba adopts a dynamic patch training strategy, as shown in Fig. 9. Specifically:

- In each training batch, a random patch size is selected from $\{128^2, 256^2, 324^2\}$.
- Each mini-batch contains images of the same resolution, but resolution varies across batches.
- The 2D scanning mechanism within FSSB is thereby trained to handle multi-scale representations efficiently.

This improves generalization to real-world conditions, especially on mobile devices and webcams that adapt resolution dynamically to conserve power or bandwidth.

## F. Dynamic Adjustment Approximation

The Dynamic Adjustment Approximation (DAA) module provides an unsupervised mechanism for exposure correction by leveraging intrinsic image statistics, eliminating the need for reference ground truth maps or pre-computed illumination priors. Unlike prior approaches such as KinD, LLFlow, or RetinexFormer, which rely heavily on ground-truth mean score as guidance signals derived from paired datasets, our method is entirely self-reliant and dynamically adapts to the brightness characteristics of each input. This approach helps in boosting the performance during inference.

Given an input image $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$, we compute two summary statistics across all pixels and channels: the mean $\mu$ and the median $m$. A normalized intensity value $\tau \in [0, 1]$ is selected as the desired luminance anchor (empirically $\tau = 0.345$ in our case). The goal is to shift the median pixel values toward this anchor in a manner weighted by their deviation from the image's current mean brightness.

The adjustment factor $\mathbf{F}$ is computed as:

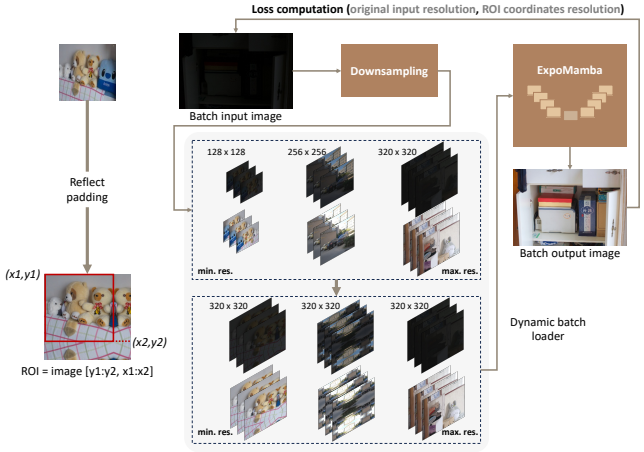
$$\mathbf{F} = \frac{m + \alpha \cdot (\tau - \mu)}{m} \tag{20}$$

Here, $\alpha$ is a tunable strength parameter that controls the degree of adjustment applied per image. The final adjusted image $\hat{\mathbf{I}}$ is computed element-wise as:

$$\hat{\mathbf{I}} = \mathbf{I} \times \mathbf{F} \tag{21}$$

This formulation ensures stability even under extreme low-light conditions by avoiding division by near-zero values and maintains exposure balance without introducing unnatural contrast. Because the adjustment is performed in a single-pass, the method is highly efficient and can be seamlessly integrated into real-time inference pipelines. Its effectiveness is especially pronounced in deployment scenarios where reference illumination statistics are unavailable, such as mobile or embedded imaging systems.

Figure 9. Illustration of dynamic patch training: multiple resolutions are randomly batched with padding, enabling the model to generalize across scales.



## G. Additional Benchmark Visualizations

Fig. 10 illustrates additional visual comparisons across 16 baseline methods and the proposed ExpoMamba model. While methods such as Zero-DCE++, RUAS, and LIME tend to either over-enhance or introduce unnatural hues, ExpoMamba preserves both local structure and global illumination. Compared to LLFormer and Restormer, which occasionally oversmooth textures or lose highlight fidelity, ExpoMamba maintains sharper edges and natural luminance transitions. These results further validate the model's ability to generalize well across diverse lighting conditions present in the LOLv1 dataset.

## H. Discussion: Inference Time vs. FLOPs

In this work, we prioritize reporting real inference time over theoretical FLOPs, as the former provides a more accurate reflection of practical deployment efficiency. While FLOPs serve as a platform-agnostic metric for algorithmic complexity, they fail to capture critical system-level considerations such as memory bandwidth, caching behavior, and inter-layer communication, all of which play a substantial role in runtime performance on real hardware. In contrast, actual inference time directly reflects the influence of design choices such as data flow optimization, activation reuse, and parallel execution scheduling.

This distinction becomes especially important in latency-sensitive applications like augmented reality, autonomous navigation, or mobile imaging, where wall-clock latency, not theoretical compute, determines responsiveness and usability. For ExpoMamba, we report inference time measured on an NVIDIA A10G GPU, which accounts for the full end-to-end processing pipeline, including frequency decomposition, dual VSSM inference, and reconstruction. This emphasis on timing enables a fairer and more relevant comparison with baseline models in the context of deployment scenarios.

Optimizing for real-time execution rather than abstract operation counts ensures that ExpoMamba is not only computationally efficient but also pragmatically viable for edge environments where latency, memory, and energy constraints coexist.

## I. Extended Comparative Efficiency and Model Scalability

**Comparative Efficiency Analysis.** ExpoMamba achieves a strong balance between visual quality and practical deployment metrics. As reported in Tab. 1, it processes a $400 \times 600$ image in 36 ms with a 2923 MB memory footprint—substantially faster than DiffLL [34] (158 ms, 8249 MB) and LLFormer [80] (1956 ms, ¿6 GB). This performance highlights the advantage of using linear-time operations and frequency-state modeling over transformer-based alternatives for real-time applications.

**Balancing Speed and Effectiveness.** Although ExpoMamba is not the smallest model in terms of parameter count, it outperforms many smaller baselines (e.g., IAT: 0.09M, FECNet+ERL: 0.15M) in both perceptual enhancement and downstream task accuracy (Tab. 4). Its modular architecture supports scalable deployment: ExpoMamba$_s$ offers a lightweight configuration for edge devices, while ExpoMamba$_l$ can scale up for cloud or desktop inference. This adaptability—combined with low latency and generalization across tasks—makes it a strong candidate for real-world LLIE pipelines in mobile, surveillance, and automotive domains.

## J. Expanded Algorithm for ExpoMamba

**Algorithm 2** ExpoMamba Training with Frequency State Space Block (FSSB)

**Input:** Dataset $\mathcal{D}$, Training epochs $E$, Components: FSSB, $\text{VSSM}_\text{A}$, $\text{VSSM}_\text{P}$, HDR, ComplexConv
**Output:** Optimized model parameters $\theta$
**// Frequency Decomposition**
**for each image** $I \in \mathcal{D}$ **do**
  Compute Fourier Transform $\mathcal{F}(u, v) = \mathcal{F}[I(x, y)]$
  Decompose: $\mathcal{F}(u, v) \rightarrow A(u, v), P(u, v)$
**end for**
**// Frequency-State Modeling (FSSB)**
**for each component** $(u, v)$ **do**
  Update VSSMs:
  $\mathbf{h}[t+1] = \mathbf{A}[t] \cdot \mathbf{h}[t] + \mathbf{B}[t] \cdot \mathbf{x}[t], \quad \mathbf{y}[t] = \mathbf{C}[t] \cdot \mathbf{h}[t]$
  Generate modulated outputs $A''(u, v), P''(u, v)$
**end for**
**// Inverse Transform & Reconstruction**
Combine frequency outputs: $\hat{\mathcal{F}}(u, v) = A''(u, v) + i \cdot P''(u, v)$
Apply Inverse Fourier Transform: $\hat{I}(x, y) = \mathcal{F}^{-1}[\hat{\mathcal{F}}(u, v)]$
**// Model Training**
**for** $e = 1$ to $E$ **do**
  **for each batch** $\mathcal{B} \subset \mathcal{D}$ **do**
    Pass $\mathcal{B}$ through FSSB $\rightarrow$ HDR $\rightarrow$ ComplexConv
    Compute loss $\mathcal{L}$; Backpropagate $\nabla_\theta \mathcal{L}$
  **end for**
**end for**
**Return:** Trained parameters $\theta$

## K. Model Configuration.

Model configuration (see Table 7 for details) provides a detailed comparison between the two variants of ExpoMamba, highlighting their configurations and performance metrics. Notably, despite an increase of 125 million parameters, the memory of the larger ExpoMamba$_l$ variant is 5690 Mb, which is a modest increase compared to transformer-based models and ExpoMamba$_s$.

Figure 10. Qualitative comparison of ExpoMamba and baselines on the LOLv1 dataset. Results demonstrate structural fidelity and color balance under mixed lighting conditions.

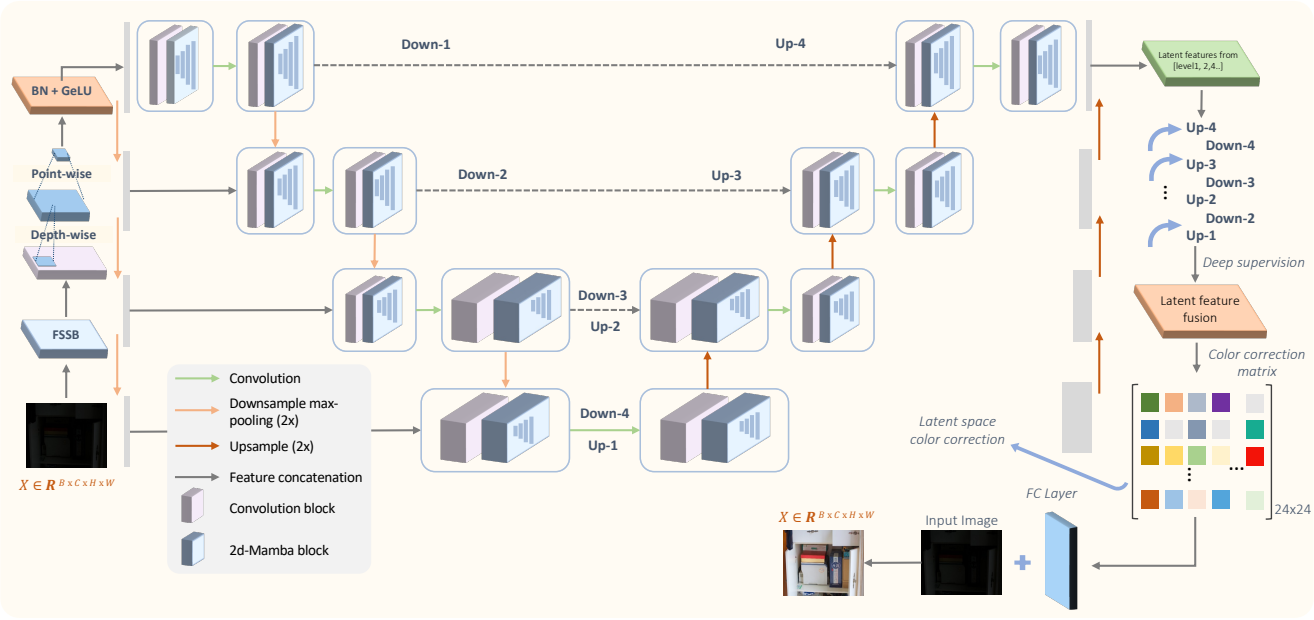




Figure 11. Overview of the ExpoMamba Architecture. The diagram illustrates the information flow through the *ExpoMamba* model.