MambaRaw: Selective State Space Modeling for Efficient 4K Raw Image Reconstruction

Peize Li[1,2⋆], Fanhu Zeng[1⋆], Tongda Xu[1], Xingguo Xu[3], Xinjie Zhang[4], Xingtong Ge[5], Haotian Zhang[6], and Yan Wang[1†]

[1] Institute for AI Industry Research (AIR), Tsinghua University
[2] King's College London
[3] Dalian University of Technology
[4] Microsoft Research Asia
[5] Hong Kong University of Science and Technology
[6] School of Computer Science, Peking University

Abstract. In-camera JPEG previews are ubiquitous in raw image formats and provide an sRGB reference at negligible storage cost. Although existing metadata-based reconstruction frameworks can exploit this side information when recovering raw images, their context models often become computationally expensive especially at high resolution, *e.g.*, 4K raw image, given that attention mechanisms scale quadratically with feature maps, hindering its practical application. To address these limitations, we propose **MambaRaw**, a JPEG-conditioned metadata-based raw image reconstruction framework that uses State Space Models (SSMs) to estimate entropy parameters efficiently. Our key contribution comprises a Spatial-Energy Coupled Context Modeling mechanism with two lightweight modules: (1) TileMambaBlock, which performs Mamba-style selective scanning only on information-dense tiles to improve the efficiency; and (2) Energy-Aware Refinement (EAR), an identity-initialized residual module that enhance feature representation to match the long-tail energy distribution of raw signals. Extensive experiments on three camera datasets (Sony, Olympus, Samsung) show consistent improvements over strong metadata-based baselines and set a new state of the art for JPEG-guided raw reconstruction with great efficiency. Notably, at low metadata bitrates, MambaRaw increases PSNR by 1.2–1.4 dB and reduces end-to-end coding latency by about 9%. Code is released at `https://github.com/Peizeli1/MambaRaw`.

Keywords: Raw image reconstruction · Metadata-based processing · State space models · Efficient inference

1 Introduction

Raw images preserve scene-referred radiance with high bit depth and dynamic range. They provide a high-fidelity basis for computational photography. However, storing and transmitting high-resolution raw captures requires substantial

⋆ Equal Contribution. †Corresponding Author.

bandwidth. Standard codecs such as JPEG [42] and HEIF [20] often perform poorly on raw data. The spatial and channel statistics of sRGB differ from those of raw signals, so sRGB-oriented encoders are not well matched. Learned image compression (LIC) [2, 28] improves efficiency by learning context models, but extending LIC to raw data remains challenging. Raw signals have uneven and camera-dependent channel distributions, which are not well captured by uniform context models.

Many raw formats also store an aligned in-camera JPEG preview. Recent metadata-based reconstruction frameworks, such as SAM [32] and R2LCM [43], exploit this side information. They transmit a compact metadata bitstream and reconstruct raw signals with JPEG guidance. The main bottleneck is still context modeling at high resolution. As shown in Fig. 1a, convolutional context models have limited receptive fields and miss long-range spatial correlations. Self-attention is also costly because its compute and memory grow quadratically on 4K feature maps. In addition, prior methods apply these heavy context modules to all regions, which wastes computation. To this end, it is natural that raw reconstruction needs content-adaptive spatial modeling. Smooth regions require limited processing, while texture-rich regions benefit from accurate long-range reasoning. This motivates an efficient architecture that combines long-range sequence modeling with sparse computation.

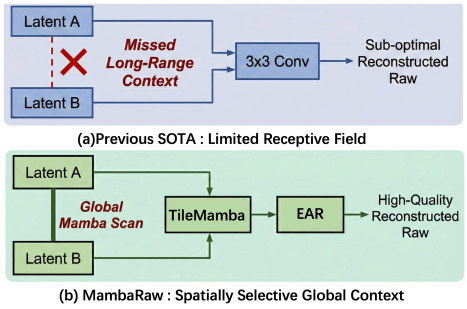


Fig. 1: Motivation and Comparison. (a) Convolution-based methods have limited receptive fields, which restricts long-range spatial modeling. (b) **MambaRaw** uses a spatial–energy coupled context model. It applies TileMambaBlock for selective scanning on information-dense tiles and uses EAR for energy-guided refinement.

Based on this observation, we propose **MambaRaw**, an efficient JPEG-conditioned metadata reconstruction framework that integrates state space models into entropy parameter estimation with Spatial-Energy Coupled Context Modeling. It includes two lightweight modules: (1) **TileMambaBlock**, which applies selective scanning to information-dense tiles only to reduce computation while preserving global context; (2) **Energy-Aware Refinement (EAR)**, an identity-initialized residual module that enhance features based on long-tail raw energy distribution.

Extensive experiments on three diverse camera datasets (Sony, Olympus, and Samsung) demonstrate that MambaRaw establishes a new state of the art for JPEG-guided raw reconstruction. By effectively addressing the context modeling bottleneck, our approach achieves consistent rate–distortion improvements over strong metadata-based baselines. Notably, at competitive metadata bitrates, MambaRaw yields substantial PSNR gains of 1.2–1.4 dB while simultaneously

reducing the end-to-end coding latency by approximately 9%, demonstrating a superior balance between high-fidelity reconstruction and practical computational efficiency. Our contributions are as follows:

- We propose a **spatial–energy coupled context modeling** paradigm for JPEG-guided raw reconstruction. It addresses high-resolution bottleneck by combining global spatial scanning with energy-guided entropy refinement.
- We introduce two lightweight modules: **TileMambaBlock** and **Energy-Aware Refinement (EAR)**, which enables content-adaptive selective scanning for efficient global modeling on information-dense tiles and enhances spatial entropy features for long-tail raw distributions.
- Extensive experiments show the effectiveness and efficiency of our method, with 1.2–1.4 dB PSNR improvement over strong baselines and about 9% lower end-to-end coding time.

2 Related Work

2.1 Learned Image Compression

Learned image compression advances rapidly since Ballé *et al.* [1] introduce end-to-end neural image compression. Many follow-up studies improve entropy modeling with hyperpriors [2], joint autoregressive and hierarchical priors [28], causal context prediction [14], and attention with mixture likelihoods [8, 10]. Other works focus on faster context designs, including checkerboard modeling [17] and efficient convolutional entropy modeling [22]. To better capture long-range dependencies, Transformer-based models are introduced for compression [23,51] and achieve improved rate–distortion performance. Uneven grouping and cross-channel context modeling [16, 26, 29, 33] further improve the balance between spatial and channel aggregation. Recent methods also explore window attention and mixed Transformer–CNN designs [23, 52]. Despite strong performance, these approaches can be expensive on high-resolution raw inputs because attention scales poorly and activation memory is large. Beyond images, learned compression principles extend to video, including probabilistic video rescaling [38], unsupervised video semantic compression [37, 40, 41], and low-bitrate coding frameworks for video understanding [39], all highlighting the importance of effective entropy modeling and content-adaptive context designs.

Efficient high-resolution processing is a common bottleneck in learned compression because context modeling operates on large feature maps. Prior works explore adaptive computation for efficient image restoration [49], C2SSM's cluster-centric scanning paradigm for ultra-high-definition image restoration [45], and saliency-driven bit allocation for perceptual compression [31]. Instead of applying heavy global reasoning to every location, we partition a 4K feature map into tiles with content-adaptive selective SSM processing and apply expensive context modeling only to regions that need it.

2.2 Metadata-based RAW Reconstruction

Several works study metadata-based raw reconstruction. In this setting, an sRGB image (or preview) is stored with a compact metadata bitstream that is sampled or learned from the raw capture. The raw image is then reconstructed when needed. This setting is related to, but different from InvISP [46], which learns an invertible mapping between rendered images and raw signals without explicitly allocating a metadata bitrate. Metadata-based reconstruction instead treats the JPEG/sRGB preview as an available reference and transmits only a compact side bitstream, making rate–distortion efficiency a central objective. Punnappurath and Brown [32] propose spatially aware metadata sampling and reconstruction. CAM [30] introduces content-adaptive metadata for sRGB-to-raw de-rendering and shows the benefit of online fine-tuning at test time. R2LCM [43] proposes learned compact metadata in feature space and improves entropy modeling for raw image compression; its journal extension Beyond-R2LCM [44] further enhances the entropy model and achieves stronger rate–distortion performance. These methods provide strong baselines, but their entropy and context models are largely convolution-based, which limits rate–distortion performance due to restricted receptive fields. While Transformer-based approaches can improve performance through long-range modeling, they suffer from low computational efficiency due to quadratic attention complexity and become costly for high-resolution inference.

2.3 State Space Models for Vision

State Space Models (SSMs) are an efficient alternative to Transformers for long-sequence modeling. This line of work originates from the Structured State Space sequence (S4) model [12]. Mamba [11] extends SSMs with selective state spaces and input-dependent parameters, which enables linear-time sequence modeling. Recent work also studies theoretical connections between Transformers and SSMs [9]. In computer vision, Mamba is adapted into backbones such as Vision Mamba [50] and VMamba [24]. Progress on efficient 2D spatial modeling includes 2DMamba [48], which proposes a hardware-aware 2D selective scan for gigapixel whole-slide image classification. Other extensions include windowed scanning in LocalMamba [19], hybrid designs such as MambaVision [15], and domain-specific variants such as U-Mamba for biomedical segmentation [27]. SSM-based models are also applied to image restoration, including MambaIRv2 [13] with attentive state-space equations for non-causal modeling, Q-MambaIR [5], and VMambaIR [36].

In image compression and RAW processing, recent works [34, 35, 47] explore SSMs for efficient entropy or spatial modeling. RAWMamba [4] uses a Mamba framework for a unified sRGB-to-RAW de-rendering pipeline, while CMIC [6] proposes a content-adaptive Mamba architecture for learned image compression. These studies demonstrate the promise of SSMs, but they mainly redesign the reconstruction/backbone network or rely on complex token organizations. In contrast, our goal is JPEG-guided metadata-based RAW reconstruction under an

explicit metadata bitrate. MambaRaw therefore inserts SSMs into the entropy-parameter network and couples selective computation with a lightweight spatial energy map, so that high-resolution context modeling is concentrated on informative regions while preserving the existing metadata-reconstruction formulation.

3 Method

3.1 Problem Setup

Metadata-based Raw Reconstruction. We study JPEG-guided metadata-based raw reconstruction. Let $\mathbf{x}_{\text{raw}} \in [0,1]^{3\times H\times W}$ denote a scene-referred raw image (in a fixed raw color space) and $\mathbf{x}_{\text{jpg}} \in [0,1]^{3\times H\times W}$ its aligned in-camera JPEG preview. The encoder produces a compact metadata bitstream $\mathbf{s}$ from $(\mathbf{x}_{\text{raw}}, \mathbf{x}_{\text{jpg}})$, and the decoder reconstructs $\hat{\mathbf{x}}$ given $(\mathbf{s}, \mathbf{x}_{\text{jpg}})$. More details can be found in Appendix A.1.

State Space Models. Our method utilizes the Visual State Space (VSS) block from VMamba [24]. In discrete time, the state space model is defined as:

$$h_{t+1} = \mathbf{A}h_t + \mathbf{B}x_t, \quad y_t = \mathbf{C}h_t, \tag{1}$$

where x_t, y_t are the input and output at time t, h_t is the hidden state, and $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are learned input-dependent parameters. To apply 1D SSMs to 2D feature maps, we employ cross-scan with four scanning directions (left–right, right–left, top–bottom, bottom–top) and merge results:

$$\mathbf{y} = \text{CrossMerge}\big(\text{SS2D}(\text{CrossScan}(\mathbf{x}))\big), \tag{2}$$

where $\mathbf{x}, \mathbf{y}$ are the input and output 2D feature maps.

3.2 Motivation and Overview

Efficiently transmitting and recovering 4K raw images presents a unique challenge: balancing high-fidelity reconstruction with computational feasibility. We address this trade-off between efficiency and effectiveness by identifying two key lightweight modules, both of which can be unified through the lens of spatial feature **Energy**, defined as the squared magnitude of feature activations $E = f^2$.

Efficiency through Energy Selection. In high-resolution raw images, information is not uniformly distributed. High-energy regions typically correspond to detailed foregrounds (textures, edges), while low-energy regions correspond to smooth backgrounds, as shown in Figure 2b. Processing the entire 4K feature map with complex context models is redundant. Therefore, by using energy to distinguish foreground from background, we can selectively apply computationally intensive modeling only where it is most needed.

Effectiveness through Energy Distribution. Even within informative regions, the energy distribution is often long-tailed and uneven across spatial locations, as is illustrated in Figure 2a. A uniform context model may fail to capture

these subtle variations. To maximize effectiveness, we need a mechanism that can refine features adaptively based on their specific energy characteristics, enhancing the representation of complex signals.

Guided by these observations, we propose **MambaRaw**, which integrates two core modules: (1) **TileMambaBlock** for efficiency, using tile-wise energy to select and process only information-dense foreground regions; and (2) **Energy-Aware Refinement (EAR)** for effectiveness, utilizing an identity-initialized residual to align with the intrinsic long-tail energy distribution of raw signals. The overall mixed-scale inference used in our context model is summarized in Algorithm 1, including both the selective TileMambaBlock processing for efficiency and the EAR refinement (Sec. 3.5) for energy-calibrated features.

3.3 JPEG-Conditioned Reconstruction Backbone

As shown in Fig. 3, we build on a hyperprior-based framework with learned context modeling following the R2LCM line of work [43, 44]. The architecture consists of: (1) a JPEG-conditioned analysis transform g_a that maps $(\mathbf{x}_{\text{raw}}, \mathbf{x}_{\text{jpg}})$ to latents $\mathbf{y}$; (2) hyper analysis/synthesis transforms h_a, h_s that produce hyper-latents $\mathbf{z}$ and side information for entropy modeling; and (3) a JPEG-conditioned synthesis transform g_s that reconstructs $\hat{\mathbf{x}}$ from quantized latents.

We incorporate JPEG guidance via feature concatenation: at each resolution level l, we first resize the JPEG preview to match the spatial resolution of layer l:

$$\mathbf{x}_{\text{jpg}}^{(l)} = \phi_l(\mathbf{x}_{\text{jpg}}^{(l-1)}), \tag{3}$$

where $\phi_l(\cdot)$ denotes bilinear interpolation to the spatial resolution of layer l. We then concatenate the intermediate feature and the resized JPEG feature along channel dimension:

$$\tilde{\mathbf{F}}^{(l)} = \left[\mathcal{M}^{(l-1)}(\tilde{\mathbf{F}}^{(l-1)}), \mathbf{x}_{\text{jpg}}^{(l)}\right], \tag{4}$$

where $\mathcal{M}^{(l-1)}$ is the layer $l-1$ mapping function, $[\cdot,\cdot]$ denotes concatenation, and $\tilde{\mathbf{F}}^{(l)}$ is the JPEG-conditioned feature. The same conditioning applies to g_a, g_s, and the entropy-parameter branch, injecting JPEG guidance at every scale without cross-attention overhead. Resized JPEG features are concatenated before each mapping, providing running latents and aligned preview cues.

For the entropy model, the conditioned feature first enters an entropy stem ($\psi_{\text{ep}}^{\text{in}}$ mapping to the internal channel dimension C) and then passes through the proposed coupled context modules:

$$\mathbf{F}_{\text{in}} = \psi_{\text{ep}}^{\text{in}}(\tilde{\mathbf{F}}), \quad \mathbf{F}_c = \text{TileMambaBlock}(\mathbf{F}_{\text{in}}; T, \rho), \quad \mathbf{F}' = \text{EAR}(\mathbf{F}_c), \tag{5}$$

and hyperprior side information:

$$\mathbf{U} = h_s(\hat{\mathbf{z}}), \tag{6}$$

followed by the entropy head $\psi_{\text{ep}}^{\text{out}}$ that predicts independent single Gaussian distribution parameters for each latent element:

$$(\boldsymbol{\mu}, \log \boldsymbol{\sigma}) = \psi_{\text{ep}}^{\text{out}}(\mathbf{F}', \mathbf{U}). \tag{7}$$

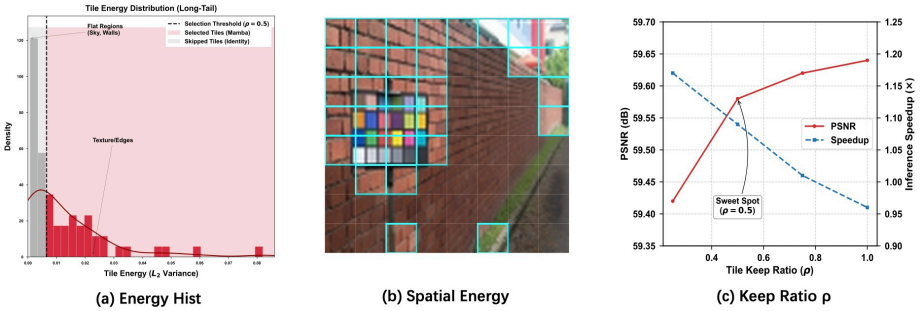


Fig. 2: Energy analysis and tile selection. **(a)**: Long-tail energy distribution motivating EAR. **(b)**: Spatial L2 energy map showing selected high-energy tiles (cyan) at $\rho = 0.5$. **(c)**: Impact of keep ratio ρ; $\rho = 0.5$ offers the optimal accuracy-speed trade-off.

Here, $\tilde{\mathbf{F}}$ is the JPEG-conditioned feature, $\mathbf{F}_{\text{in}}$ is the context input before SSM processing, $\mathbf{F}_c$ is the TileMambaBlock output, and $\mathbf{F}'$ is the final EAR output.

3.4 Efficiency: TileMambaBlock with Energy-Guided Selection

To address the efficiency bottleneck of 4K processing, we propose **TileMambaBlock**, which leverages the spatial distribution of energy to selectively apply long-range context modeling. As discussed in Sec. 3.2, high-energy regions ($E = f^2$) effectively distinguish detailed foregrounds from smooth backgrounds. Treating all regions equally with heavy context models incurs unnecessary computation. Instead, TileMambaBlock dynamically identifies and processes only information-dense tiles. This design is fundamentally aligned with the inherent sparsity of high-frequency information in natural images; by restricting advanced scanning mechanisms to structurally complex regions, we drastically reduce required float-point operations (FLOPs) without sacrificing the perceptual fidelity of raw imagery.

Tile partition. Given the context input $\mathbf{F}_{\text{in}} \in \mathbb{R}^{C\times H\times W}$ in Eq. 5, we partition $\mathbf{F}_{\text{in}}$ into non-overlapping tiles of size $T \times T$. The resulting set of tiles is denoted as $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_{N_t}\}$, where $N_t = \lceil H/T \rceil \times \lceil W/T \rceil$.

To facilitate Energy-Guided Selection, we quantify the information density of each tile using its L2 energy, serving as a computationally inexpensive proxy for entropy:

$$S_i = \frac{1}{CT^2} \sum_{c,h,w} \mathbf{t}_i[c,h,w]^2. \tag{8}$$

Based on these scores S_i, we identify a subset $\mathcal{S}$ containing the top-k most informative tiles, where $k = \lfloor \rho N_t \rfloor$ is controlled by a keep ratio $\rho \in (0, 1]$. The Mamba-based context modeling is then applied exclusively to these selected tiles:

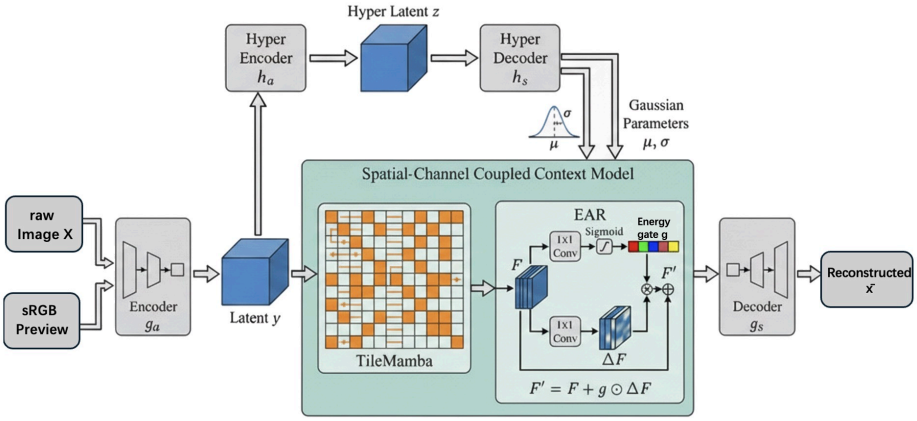


Fig. 3: The Overall Framework of MambaRaw. We adopt a two-level VAE architecture conditioned on the available JPEG preview. The core innovation lies in the Level-1 Context Model, where we replace standard separate spatial/channel contexts with a coupled design: **TileMambaBlock** for efficient long-range spatial modeling on selected information-dense tiles, and **EAR** for lightweight energy-guided refinement.

$$\mathbf{t}_i' = \begin{cases} \text{MambaBlock}(\mathbf{t}_i), & i \in \mathcal{S}, \\ \mathbf{t}_i, & \text{otherwise}, \end{cases} \tag{9}$$

where $\mathbf{t}_i'$ is the processed tile. This selection reserves expensive context modeling for complex regions. For small patches ($H, W \leq T$), it defaults to dense processing.

3.5 Effectiveness: Energy-Aware Refinement (EAR)

While TileMambaBlock ensures efficiency by selecting important spatial regions, we also need to ensure the effectiveness of context modeling within these regions. Raw image features exhibit a long-tail energy distribution (Figure 2a), where information is unevenly distributed across spatial locations and feature dimensions. A static or uniform context model may struggle to adapt to this variance.

To address this, we introduce the **Energy-Aware Refinement (EAR)**, a lightweight refurbishment module designed to enhance feature representation based on local energy statistics. EAR acts as a spatial-energy refinement step that dynamically adjusts feature responses according to their energy profile. Unlike SENets [18] which rely on global average pooling for channel recalibration, EAR explicitly preserves local spatial granularity rather than discarding it via uniform pooling. This introduces a spatially-varying inductive bias that prevents the over-smoothing of critical high-frequency details, allowing the entropy model to adapt robustly to nuanced local raw signal variations.

Equation 5 and Algorithm 1 outline the path $\tilde{\mathbf{F}} \to \mathbf{F}_{\text{in}} \to \mathbf{F}_c \to \mathbf{F}'$. From the context-enhanced feature $\mathbf{F}_c \in \mathbb{R}^{C \times H \times W}$, we compute spatial energy $\mathbf{e}$ along

Algorithm 1 Mixed-Scale Inference (blue: TileMambaBlock, red: EAR)

Require: JPEG-conditioned feature $\tilde{\mathbf{F}} \in \mathbb{R}^{C_0 \times H \times W}$, tile size T, keep ratio ρ
Ensure: Refined context features $\mathbf{F}'$

$\mathbf{F}_{\text{in}} \leftarrow \psi_{\text{ep}}^{\text{in}}(\tilde{\mathbf{F}})$ ▷ Context input projection
$N_h \leftarrow \lceil H/T \rceil,\ N_w \leftarrow \lceil W/T \rceil,\ N_t \leftarrow N_h N_w$
if $(H \le T \wedge W \le T)$ **or** $(\rho \ge 1)$ **then**
  $\mathbf{F}_c \leftarrow \text{MambaBlock}(\mathbf{F}_{\text{in}})$ ▷ Dense fallback
else
  Pad $\mathbf{F}_{\text{in}}$ to multiples of T: $\mathbf{F}_{\text{in,pad}}$
  Reshape $\mathbf{F}_{\text{in,pad}} \to \{\mathbf{t}_i\}_{i=1}^{N_t}$, $\mathbf{t}_i \in \mathbb{R}^{C \times T \times T}$
  Compute tile score: $S_i \leftarrow \frac{1}{CT^2} \sum \mathbf{t}_i^2$
  $k \leftarrow \max(1, \lfloor \rho N_t \rfloor),\ \mathcal{I}_{\text{top}} \leftarrow \text{TopKIndices}(S, k)$
  for $i \leftarrow 1$ to N_t **do**
    if $i \in \mathcal{I}_{\text{top}}$ **then**
      $\mathbf{t}_i' \leftarrow \text{MambaBlock}(\mathbf{t}_i)$
    else
      $\mathbf{t}_i' \leftarrow \mathbf{t}_i$ ▷ Skip smooth tiles
    end if
  end for
  Reshape $\{\mathbf{t}_i'\}$ and crop padding: $\mathbf{F}_c$
end if
$\mathbf{e} \leftarrow \frac{1}{C} \sum_{j=1}^{C} \mathbf{F}_{c,j}^2$
$\mathbf{g} \leftarrow \sigma(\text{Conv}_{1\times1}(\mathbf{e}))$
$\Delta\mathbf{F} \leftarrow \text{Conv}_{1\times1}(\text{ReLU}(\text{Conv}_{1\times1}(\mathbf{F}_c)))$
$\mathbf{F}' \leftarrow \mathbf{F}_c + \mathbf{g} \odot \Delta\mathbf{F}$ ▷ EAR residual
return $\mathbf{F}'$

channels as an entropy proxy:

$$\mathbf{e} = \frac{1}{C} \sum_{j=1}^{C} (\mathbf{F}_{c,j})^2 \in \mathbb{R}^{1 \times H \times W}. \tag{10}$$

where $\mathbf{F}_{c,j}$ denotes the j-th channel slice of $\mathbf{F}_c$. We use $\mathbf{e}$ to gate an energy-guided residual for spatial enhancement:

$$\mathbf{g} = \sigma(\text{Conv}_{1\times1}(\mathbf{e})) \in \mathbb{R}^{C \times H \times W}, \tag{11}$$

$$\Delta\mathbf{F} = \text{Conv}_{1\times1}\big(\text{ReLU}(\text{Conv}_{1\times1}(\mathbf{F}_c))\big), \tag{12}$$

$$\mathbf{F}' = \mathbf{F}_c + \mathbf{g} \odot \Delta\mathbf{F}, \tag{13}$$

where σ is sigmoid, $\odot$ is element-wise multiplication, $\mathbf{g}$ is the gating tensor, and $\Delta\mathbf{F}$ is the residual variation. For stable early training, the last 1×1 convolution is zero-initialized, letting EAR start as an identity mapping.

3.6 Training Strategy

Loss function. We optimize the rate–distortion trade-off:

$$\mathcal{L} = R(\hat{\mathbf{y}}) + R(\hat{\mathbf{z}}) + \lambda \cdot D(\mathbf{x}, \hat{\mathbf{x}}), \tag{14}$$

where $R(\cdot)$ denotes estimated bitrate (metadata only) and $D(\cdot)$ is a distortion term. We train models across multiple λ values to cover different rate–distortion points.

Training protocol. We use the same backbone architecture as [44] and train from scratch. TileMambaBlock and EAR are inserted into the entropy-parameter network. EAR is identity-initialized to ensure stable early-stage training when the entropy model is not yet converged. Tile-wise selection is primarily beneficial for high-resolution inference; for patch-based training where feature maps are small, the module naturally reduces to dense processing.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on two standard benchmarks: NUS dataset [7] and AdobeFiveK dataset [3]. (1) **NUS dataset**: As a primary benchmark for raw reconstruction, this dataset covers diverse scenes captured by varying sensors. Following the protocol of CAM [30], we employ three representative subsets (Samsung NX2000, Olympus E-PL6, and Sony SLT-A57) and evaluate on the 4× downsampled version to ensure fair comparison with prior art. (2) **AdobeFiveK dataset:** We utilize AdobeFiveK to assess reconstruction fidelity under complex lighting and professional photographic conditions. We adopt the Software ISP evaluation setting defined in [44] with 4,500 training / 500 testing pairs, where sRGB targets are rendered via a software ISP at original resolution.

Implementation details. Measurements are performed in the raw-linear color space with inputs normalized to $[0, 1]$. We set the tile keep ratio $\rho = 0.5$ by default, which balances performance and efficiency. The TileMambaBlock and EAR modules are integrated into the entropy-parameter estimation network. We train the model from scratch using 256×256 patches, the Adam optimizer, and mixed precision. Distinct models are trained for each $\lambda \in \{0.02, 0.24, 0.8, 1.5, 2.0, 5.0, 10.0, 20.0\}$ for 1000 epochs.

Metrics and baselines. We report PSNR (dB), SSIM, and metadata bitrate (bpp). We compare against baselines including SAM [32], CAM [30], R2LCM [43], and Beyond-R2LCM [44]. Detailed definitions of the evaluation metrics are provided in Appendix A.

4.2 Main Results

Rate–distortion trade-off. Figure 4 and Table 1 present the rate–distortion comparison on the NUS dataset. Across all three camera subsets (Samsung, Olympus, Sony), MambaRaw consistently dominates the rate-distortion frontier. By effective tile-wise selective scanning and energy-guided refinement, our method delivers superior reconstruction quality (PSNR and SSIM) at comparable or lower metadata bitrates. In particular, our method outperforms robust metadata-based baselines such as R2LCM and Beyond-R2LCM by significant

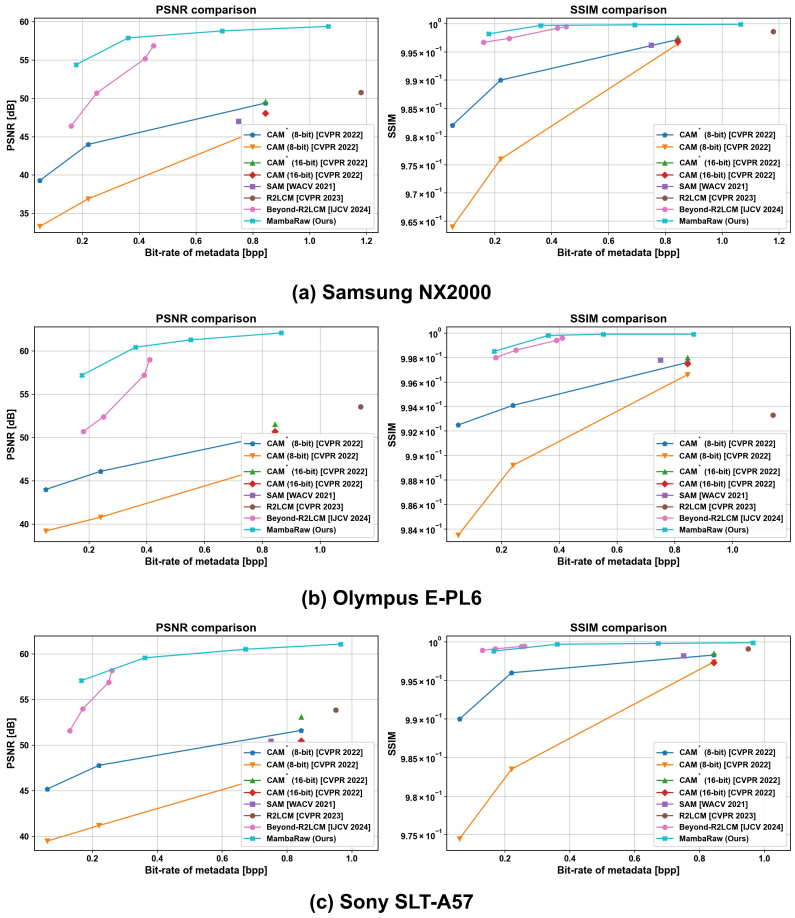


Fig. 4: RD curves over the NUS dataset (Samsung NX2000, Olympus E-PL6, Sony SLT-A57) following the setting of [30]. The left and right columns report PSNR and SSIM, respectively. For variable-rate models, a single model is trained for each curve and different operating points are obtained by changing the rate–distortion hyper-parameter of the trained model.

margins. Specifically, on the Samsung subset, MambaRaw achieves a **1.2 dB** PSNR gain, while on the Sony and Olympus subsets, the improvement reaches up to **1.4 dB**, particularly in detailed texture regions where conventional models struggle. This robust superiority across diverse sensor statistics and bitrates confirms that our spatial-energy coupled design effectively resolves the bottleneck of high-resolution context modeling.

Quantitative comparison. We also extend performance evaluation to the AdobeFiveK dataset, which introduces greater variability in scene content and lighting. Table 2 summarizes the quantitative comparison against state-of-the-

Table 1: Quantitative results on the NUS dataset. PSNR (dB) and SSIM are evaluated on reconstructed raw images. Reported bpp are metadata-only bitrates, excluding the baseline JPEG preview. The * marks CAM with test-time online fine-tuning.

Method	Input Space	bpp ↓	Samsung NX2000		Olympus E-PL6		Sony SLT-A57	
			PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SAM [32]	16-bit	0.7500	47.03	0.9962	49.35	0.9978	50.44	0.9982
CAM [30]	8-bit	0.8438	46.24	0.9964	46.84	0.9966	47.66	0.9974
CAM [30]	16-bit	0.8438	48.08	0.9968	50.71	0.9975	50.49	0.9973
CAM* [30]	8-bit	0.8438	49.40	0.9972	50.37	0.9976	51.63	0.9983
CAM* [30]	16-bit	0.8438	49.57	0.9975	51.54	0.9980	53.11	0.9985
R2LCM [43]	8-bit	1.2250	50.78	0.9986	53.56	0.9993	53.87	0.9991
Beyond-R2LCM [44]	8-bit	0.3763	56.74	0.9996	59.04	0.9997	58.21	0.9996
MambaRaw (Ours)	8-bit	**0.3612**	**57.91**	**0.9997**	**60.45**	**0.9998**	**59.58**	**0.9997**

Table 2: Quantitative evaluation on AdobeFiveK dataset. The sRGB input is rendered using a software ISP and spatial resolution remains the same as the original raw image. The reported bits per pixel (bpp) are metadata-only bitrates.

Method	bpp	PSNR	SSIM
InvISP [46]	N/A	52.69	0.9994
SAM [32]	9.566e-4	49.61	0.9987
SAM [32]	9.521e-3	54.76	0.9995
CAM [30]	8.438e-1	56.72	0.9996
R2LCM [43] (w/o metadata)	N/A	53.03	0.9993
R2LCM [43]	4.901e-4	58.14	0.9997
Beyond-R2LCM [44]	3.760e-4	58.44	0.9997
Ours	**3.150e-4**	**58.55**	**0.9997**
R2LCM [43]	1.045e-2	59.02	0.9994
Beyond-R2LCM [44]	2.916e-3	59.09	0.9997
Ours	**2.450e-3**	**59.18**	**0.9998**

art methods. Notably, in the challenging low-bitrate regime ($<$ 5e-4 bpp), MambaRaw demonstrates superior efficiency, surpassing the PSNR of Beyond-R2LCM by **0.11** dB (58.55 vs. 58.44 dB) while requiring approximately **16% fewer bits** (3.150e-4 vs. 3.760e-4 bpp). This strict rate–distortion advantage is critical for applications where bandwidth is constrained.

Performance at Target Resolution (4K) and Efficiency Analysis. In addition to rate distortion results on 4× downsampled benchmarks, MambaRaw is designed for practical high resolution (4K) settings. To test whether the accuracy gains persist at the target resolution, we report full resolution performance (3840×2160) in Table 3. In true 4K, our method improves PSNR by **1.37 dB** over the baseline without increasing the parameter count. By performing selective scanning only on information dense tiles, our Spatial Energy Coupled Context Modeling avoids redundant global computation. As a result, MambaRaw reduces FLOPs by about **56%** and lowers end to end wall clock latency by 9% compared with Beyond R2LCM. At 4K, the CNN baseline exhibits substantial memory growth and approaches the device limit (22.8 GB), whereas MambaRaw remains within a consumer level memory budget (10.2 GB).

Table 3: Performance and efficiency on *true* 4K resolution inputs (3840×2160, λ=0.8). MambaRaw retains superior accuracy with high efficiency.

Method	FLOPs (G)↓	Mem (GB)↓	Time (ms)↓	PSNR↑	SSIM↑
Baseline [44]	5420.7	22.8	3125	55.21	0.9982
Ours	**2380.5**	**10.2**	**2859**	**56.58**	**0.9991**

Table 4: Progressive analysis of component effectiveness. This study validates the individual structural contribution of each proposed module.

Method	EAR	SSM	TileSelect	PSNR ↑	SSIM ↑	Time ↓
Baseline	-	-	-	58.21	0.9996	563
+ EAR	✓	-	-	58.55	0.9996	570
+ SSM (Dense)	✓	✓	-	59.61	0.9997	584
Ours	✓	✓	✓	**59.58**	**0.9997**	**515**

Table 5: Comparison of different tile selection metrics. L2 Energy provides the best balance of accuracy and speed.

Metric	PSNR ↑	SSIM ↑	Total Time (ms) ↓
Random	59.25	0.9996	515
Entropy	59.55	0.9997	542
Gradient	59.52	0.9997	528
L2 Energy (Ours)	**59.58**	**0.9997**	**515**

Table 6: Comparison of foundational context modeling blocks. SSM offers the best performance-speed trade-off.

Block Type	PSNR ↑	SSIM ↑	Total Time (ms) ↓
CNN (ResBlock)	58.45	0.9996	**480**
Transformer (Win-Attn)	59.52	0.9997	620
SSM (Ours)	**59.58**	**0.9997**	515

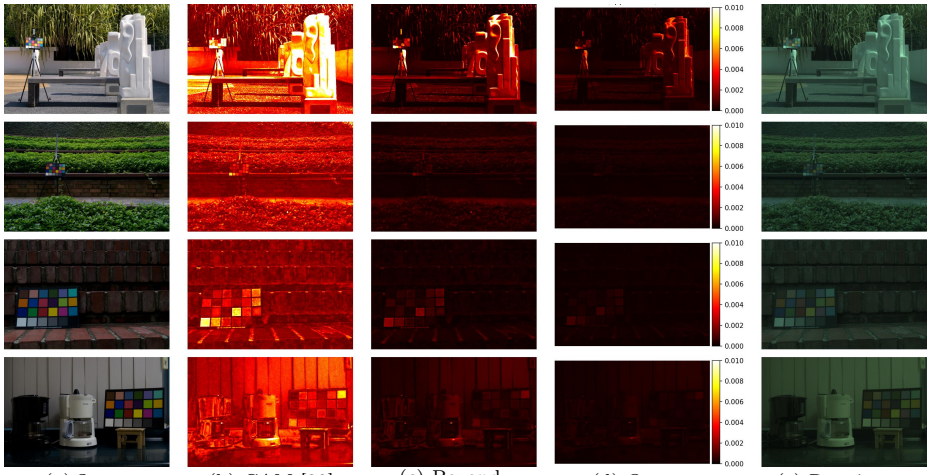
4.3 Ablation Study

We perform comprehensive ablation studies to validate our design choices. All experiments are conducted on Sony SLT-A57 at $\lambda = 0.8$ unless otherwise stated.

Effectiveness of Individual Components. We analyze the contribution of each component through progressive integration. Results in Table 4 show the step-by-step improvements. Specifically, the metadata-based baseline reaches 58.21 dB. Adding the Energy-Aware Refinement (EAR) brings a steady gain (0.34 dB) with negligible latency increase, ensuring robust spatial-energy adaptation. Integrating the dense SSM context model yields a massive performance promotion (1.06 dB) due to superior global spatial modeling, but slightly increases latency to 584 ms. Finally, employing the tile-wise selection mechanism (TileMambaBlock) maintains the high performance (59.58 dB) and reduces inference time by 12% (584 → 515 ms), proving that selective processing successfully prunes redundancy.

Effectiveness of Tile Selection Metric. We justify our choice of L2 Energy as the tile selection metric by comparing it with random selection, Entropy, and Gradient Magnitude. As shown in Table 5, Random selection leads to a noticeable performance drop (59.25 dB). While Entropy and Gradient metrics achieve competitive performance, they incur additional computational overhead. By contrast, **L2 Energy** achieves the highest efficiency (515 ms) with comparable SOTA performance. Figure 2b visualizes the spatial energy distribution: high-energy tiles concentrate on edges and textures while ignoring smooth background regions, validating the rationale for selective processing.

Impact of Tile Keep Ratio ρ. We analyze the trade-off between performance and efficiency by varying the tile keep ratio ρ. As shown in Figure 2c, increasing ρ scans more regions, slightly improving reconstruction quality with lower inference speedup. However, PSNR gains saturate beyond $\rho = 0.5$. We select $\rho = 0.5$ as a



(a) Input (8-bit sRGB image) (b) CAM [30] (bpp: 0.8438) (c) Beyond-R2LCM [44] (bpp: 0.3763) (d) Ours (bpp: 0.3612) (e) Raw image

Fig. 5: Qualitative comparison on Sony SLT-A57. Error maps show the per-pixel maximum absolute error over the three channels (after gamma correction for visibility); darker indicates smaller error.

widely applicable default, achieving a sweet spot that maintains state-of-the-art results with significant speedup.

Impact of Foundational Models. To verify the effectiveness of the proposed SSM-based design, we replace the core TileMambaBlock with CNN-based and Transformer-based (Swin Transformer [25]) alternatives. As shown in Table 6, the CNN variant suffers from limited long-range modeling with poor reconstruction performance. The Transformer variant matches our performance but incurs 20% higher latency due to quadratic attention complexity. By contrast, our SSM-based design achieves the best rate-distortion performance with optimal efficiency, strongly showing better global context modeling of the selective scanning in 4K image reconstruction.

4.4 Qualitative Visualization

To understand the superiority of MambaRaw, we present a qualitative comparison of the per-pixel absolute error in Figure 5. The sRGB preview (a) shows scenes with complex high-frequency textures (*e.g.*, the text on the book spine and the fabric pattern). The baseline Beyond-R2LCM (c) struggles to align with these fine details, resulting in residual errors in the error map (brighter regions). In contrast, MambaRaw (d) effectively suppresses these errors by refined spatial energy enhancement. Concretely, coupled spatial-channel context modeling allows the network to better predict complex signal variations, leading to a darker error map that indicates higher reconstruction fidelity. This aligns with

our quantitative results, confirming that selective SSM processing captures critical structural information more effectively than existing methods.

5 Conclusion

In this paper, we present **MambaRaw**, a JPEG-conditioned learned framework for efficient metadata-based 4K raw image reconstruction. We integrate state space models into entropy parameter estimation, and propose spatial-energy coupled context modeling with two lightweight modules: TileMambaBlock performs tile-wise selective context modeling to enable practical high-resolution inference and EAR uses energy-guided refinement to improve entropy enhancement and enhance feature representation to match the long-tail energy distribution of raw signals while maintaining stable training. Experiments on three camera datasets show consistent rate–distortion gains over strong metadata-based baselines, including up to 1.4 dB PSNR at similar bitrates, and reduce end-to-end coding latency by 9% on average.

Future work. This work focuses on single-frame raw reconstruction. A natural extension is raw video processing that exploits temporal redundancy. This direction can incorporate temporal SSM designs such as VideoMamba [21]. Another direction is hardware-aware optimization that maps selective SSM processing to mobile accelerators for real-time computational photography.

Acknowledgements

This work is supported by Wuxi Research Institute of Applied Technologies, Tsinghua University under Grant 20242001120.

References

1. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. arXiv preprint arXiv:1611.01704 (2016)
2. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436 (2018)
3. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR 2011. pp. 97–104. IEEE (2011)
4. Chen, H., Han, W., Zheng, H., Shen, J.: Rawmamba: Unified srgb-to-raw derendering with state space model. arXiv preprint arXiv:2411.11717 (2024)
5. Chen, Y., Qin, H., Zhang, Z., Magno, M., Benini, L., Li, Y.: Q-mambair: Accurate quantized mamba for efficient image restoration. arXiv preprint arXiv:2503.21970 (2025)
6. Chen, Y., Lyu, Z., He, B., Hu, H., Wang, Q., Tian, Y., Song, L., Zhang, W., Lu, G.: Cmic: Content-adaptive mamba for learned image compression. arXiv preprint arXiv:2508.02192 (2025)

7. Cheng, D., Prasad, D.K., Brown, M.S.: Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. Journal of the Optical Society of America A **31**(5), 1049–1058 (2014)
8. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7939–7948 (2020)
9. Dao, T., Gu, A.: Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. arXiv preprint arXiv:2405.21060 (2024)
10. Gao, G., You, P., Pan, R., Han, S., Zhang, Y., Dai, Y., Lee, H.: Neural image compression via attentional multi-scale back projection and frequency decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14677–14686 (2021)
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
12. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
13. Guo, H., Guo, Y., Zha, Y., Zhang, Y., Li, W., Dai, T., Xia, S.T., Li, Y.: Mambairv2: Attentive state space restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28124–28133 (2025)
14. Guo, Z., Zhang, Z., Feng, R., Chen, Z.: Causal contextual prediction for learned image compression. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 2329–2341 (2021)
15. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25261–25270 (2025)
16. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5718–5727 (2022)
17. He, D., Zheng, Y., Sun, B., Wang, Y., Qin, H.: Checkerboard context model for efficient learned image compression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14771–14780 (2021)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
19. Huang, T., Pei, X., You, S., Wang, F., Qian, C., Xu, C.: Localmamba: Visual state space model with windowed selective scan. In: European conference on computer vision. pp. 12–22. Springer (2024)
20. ISO/IEC: Information technology—high efficiency coding and media delivery in heterogeneous environments—part 3: 3d audio. Tech. Rep. ISO/IEC 23008-3:2015, International Organization for Standardization / International Electrotechnical Commission, Geneva, Switzerland (2015)
21. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: European conference on computer vision. pp. 237–255. Springer (2024)
22. Li, M., Ma, K., You, J., Zhang, D., Zuo, W.: Efficient and effective context-based convolutional entropy modeling for image compression. IEEE Transactions on Image Processing **29**, 5900–5911 (2020)

23. Liu, J., Sun, H., Katto, J.: Learned image compression with mixed transformer-cnn architectures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14388–14397 (2023)
24. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
26. Ma, C., Wang, Z., Liao, R., Ye, Y.: A cross channel context model for latents in deep image compression. arXiv preprint arXiv:2103.02884 (2021)
27. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
28. Minnen, D., Ballé, J., Toderici, G.D.: Joint autoregressive and hierarchical priors for learned image compression. Advances in neural information processing systems **31** (2018)
29. Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 3339–3343. IEEE (2020)
30. Nam, S., Punnappurath, A., Brubaker, M.A., Brown, M.S.: Learning srgb-to-raw-rgb de-rendering with content-aware metadata. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17704–17713 (2022)
31. Patel, Y., Appalaraju, S., Manmatha, R.: Saliency driven perceptual image compression. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 227–236 (2021)
32. Punnappurath, A., Brown, M.S.: Spatially aware metadata for raw reconstruction. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 218–226 (2021)
33. Qin, S., Lu, Y., Zhou, Y., Li, J., Ren, Y., Xue, Y., Xia, S.T., Chen, B.: Freqsic: Frequency-aware stereo image compression with bi-directional checkerboard context model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19393–19402 (2026)
34. Qin, S., Wang, J., Zhou, Y., Chen, B., Luo, T., An, B., Dai, T., Xia, S.T., Wang, Y.: Cassic: Towards content-adaptive state-space models for learned image compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15727–15736 (2025)
35. Qin, S., Zhang, X., Liu, Z., Wang, J., Chen, B., Li, J., Ren, Y., Xia, S.T., Zhang, J.: Mambasic: Mamba-based stereo image compression with bi-directional multi-reference entropy model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5306–5315 (2026)
36. Shi, Y., Xia, B., Jin, X., Wang, X., Zhao, T., Xia, X., Xiao, X., Yang, W.: Vmambair: Visual state space model for image restoration. IEEE Transactions on Circuits and Systems for Video Technology **35**(6), 5560–5574 (2025)
37. Tian, Y., Ling, X., Geng, C., Hu, Q., Lu, G., Zha, G.: Smc++: Masked learning of unsupervised video semantic compression. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
38. Tian, Y., Lu, G., Min, X., Che, Z., Zhai, G., Guo, G., Gao, Z.: Self-conditioned probabilistic learning of video rescaling. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4490–4499 (2021)

39. Tian, Y., Lu, G., Yan, Y., Zhai, G., Chen, L., Gao, Z.: A coding framework and benchmark towards low-bitrate video understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5852–5872 (2024)
40. Tian, Y., Lu, G., Zhai, G.: Free-vsc: Free semantics from visual foundation models for unsupervised video semantic compression. In: European Conference on Computer Vision. pp. 163–183. Springer (2024)
41. Tian, Y., Lu, G., Zhai, G., Gao, Z.: Non-semantics suppressed mask learning for unsupervised video semantic compression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13610–13622 (2023)
42. Wallace, G.K.: The jpeg still picture compression standard. Communications of the ACM **34**(4), 30–44 (1991)
43. Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.P., Kot, A.C., Wen, B.: Raw image reconstruction with learned compact metadata. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18206–18215 (2023)
44. Wang, Y., Yu, Y., Yang, W., Guo, L., Chau, L.P., Kot, A.C., Wen, B.: Beyond learned metadata-based raw image reconstruction. International Journal of Computer Vision **132**(12), 5514–5533 (2024)
45. Wu, C., Wang, L., Zheng, Z., Cui, Y., Yang, Z., Chen, X., Zhang, Y., Jiang, W., Xia, J.: Scan clusters, not pixels: A cluster-centric paradigm for efficient ultra-high-definition image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15528–15537 (2026)
46. Xing, Y., Qian, Z., Chen, Q.: Invertible image signal processing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6287–6296 (2021)
47. Zeng, F., Tang, H., Shao, Y., Chen, S., Shao, L., Wang, Y.: Mambaic: State space models for high-performance learned image compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18041–18050 (2025)
48. Zhang, J., Nguyen, A.T., Han, X., Trinh, V.Q.H., Qin, H., Samaras, D., Hosseini, M.S.: 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3583–3592 (2025)
49. Zhou, Y., Zhou, P., Ng, T.K.: Efficient cascaded multiscale adaptive network for image restoration. In: European Conference on Computer Vision. pp. 92–110. Springer (2024)
50. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)
51. Zhu, Y., Yang, Y., Cohen, T.: Transformer-based transform coding. In: International conference on learning representations (2022)
52. Zou, R., Song, C., Zhang, Z.: The devil is in the details: Window-based attention for image compression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17492–17501 (2022)

A More Details

A.1 Network Architecture

Our MambaRaw framework directly adopts the two-level JPEG-conditioned learned-context backbone of Beyond-R2LCM [44]. It consists of two cascaded analysis transforms and their corresponding synthesis transforms. At both levels, latent elements are progressively coded using learned spatial sampling masks and JPEG-conditioned context prediction. We preserve the original backbone and replace only the Level-1 entropy-parameter network with an input projection, TileMambaBlock, EAR, and an output projection. Detailed architectural settings are summarized in Table A1.

Table A1: Key hyperparameters used in our implementation.

Component	**Setting**
Backbone channel width	$N = 192$
Latent channel reduction factor	8
Reconstruction levels	2
Learned spatial sampling rounds	4
Context prediction	Masked deconvolution
Entropy distribution	Independent Gaussian (μ, σ)
Tile size (latent)	$T = 64$
Tile keep ratio	$\rho = 0.5$ (default)
SSM block	VSS (VMamba) with expansion factor 2
EAR initialization	last 1×1 conv zero-initialized

Entropy model integration. The core innovation lies in the Level-1 entropy-parameter network. We replace its original convolutional parameter estimator with an input projection, **TileMambaBlock**, **Energy-Aware Refinement (EAR)**, and an output projection. The same entropy-parameter network is used during training, encoding, and decoding.

JPEG-conditioned feature injection. For the Level-1 entropy-parameter network, we concatenate the feature propagated from the deeper level, the progressively predicted context feature, and the cumulative sampling mask. The JPEG preview is bilinearly resized to the corresponding latent resolution and concatenated before the input projection. The resulting feature follows the chain in Eq. 5, i.e., $\tilde{\mathbf{F}} \rightarrow \mathbf{F}_{\text{in}} \rightarrow \mathbf{F}_c \rightarrow \mathbf{F}'$. The aligned JPEG feature is additionally concatenated in the output projection, which predicts the Gaussian mean and scale (μ, σ).

Context Modules. (1) TileMambaBlock. Given an intermediate feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, we partition it into non-overlapping $T \times T$ tiles (on the latent feature resolution). We score each tile by its L2 energy (Eq. 8) and apply the SSM-based context block only to the top-k tiles, where $k = \lfloor \rho N_t \rfloor$. Unless otherwise specified, we use $T = 64$ and $\rho = 0.5$.

(2) State space block. The internal SSM uses the Visual State Space (VSS) block from VMamba [24] with a state expansion factor of 2 and 2D cross-scan (four directions). This provides long-range spatial aggregation with linear-time complexity.The VSS and selective-scan implementation follows MambaIC [47], while the entropy-coding structure follows Beyond-R2LCM [44].
(3) EAR. EAR refines entropy features based on local energy statistics while preserving spatial granularity. Given features $\mathbf{F}$, we compute an energy map (Eq. 10) and predict a gating tensor (Eq. 11) to modulate a lightweight residual branch (Eqs. 12–13).
(4) Learned spatial sampling. Following Beyond-R2LCM [44], latent elements are progressively coded in four spatial sampling rounds. In each round, a JPEG-conditioned learned mask selects the current spatial positions. The context-prediction network uses the previously reconstructed positions, the cumulative sampling mask, and the aligned JPEG preview to estimate the Gaussian mean and scale. The same sampling order is used during training, encoding, and decoding.
What we mean by 4K. Throughout the paper, "4K" refers to *4K-class* high-resolution RAW captures (*i.e.*, images whose long side is on the order of ∼4K pixels). Note that some benchmarks adopt downsampled evaluation protocols for fair comparison: for NUS, we report RD results on the 4× downsampled setting (Sec. 4.1). Our efficiency design targets the entropy model bottleneck on large feature maps and is therefore most beneficial at higher resolutions; the relative speedup from tile-wise selection typically increases with input size.

A.2 Training Strategy

We optimize all models using Adam with hyperparameters $(\beta_1, \beta_2) = (0.9, 0.999)$. We set the initial learning rate to 1×10^{-4} for all parameters. We use a cosine annealing schedule to decay the learning rate to 1×10^{-6} over 1000 epochs, which improves stability in the later stage of training. We train all models from scratch with a total batch size of 8 on NVIDIA RTX A30 GPUs. We adopt Automatic Mixed Precision (AMP) to reduce memory usage and improve training throughput while maintaining reconstruction quality. During training, we apply data augmentation on the fly to 256×256 patches, including random horizontal flips, random vertical flips, and random 90° rotations. This augmentation reduces overfitting and improves generalization of our spatial energy coupled context modeling across diverse raw image sequences.

A.3 Efficiency Measurement Details

Table A2 breaks down the runtime of TileMambaBlock. The selection overhead from L2 scoring, Top-K, and padding/reshaping is only 27 ms (5.2%), showing that the speedup mainly comes from avoiding dense SSM scanning on low-information tiles.

Table A2: Tile-selection overhead on Sony SLT-A57. Runtime is measured for the entropy-context branch at $\lambda = 0.8$.

Item	Time (ms)	Share
Dense SSM w/o selection	584	–
TileMamba w/ selection	515	100%
L2 score computation	10	1.9%
Top-K selection	6	1.2%
Padding/reshape	11	2.1%
Selection overhead total	27	5.2%
Selected MambaBlock scan	305	59.2%
Other codec modules	183	35.5%

Table A3: NUS subsets used in this work and their spatial resolution protocol.

Camera Subset	Native RAW Resolution	4× Evaluation Resolution
Samsung NX2000	5472×3648	1368×912
Olympus E-PL6	4608×3456	1152×864
Sony SLT-A57	4912×3264	1228×816

A.4 Evaluation Metrics

We use two standard full reference image quality metrics to evaluate reconstruction fidelity: Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). PSNR quantifies the pixel level difference between the reconstructed raw image and the ground truth, and it is reported in decibels (dB). Higher PSNR indicates lower distortion. SSIM measures similarity in structural information, luminance, and contrast, and it ranges from 0 to 1. A value of 1 indicates perfect structural agreement. We compute both metrics on raw linear RGB images normalized to the $[0, 1]$ range.

A.5 Dataset Protocol Details (NUS)

To make the NUS evaluation protocol explicit, we summarize the exact setup used in this paper and in the compared metadata-based baselines (see Table A3):

- **Subset cameras.** Samsung NX2000, Olympus E-PL6, and Sony SLT-A57.
- **Input-target pair.** Aligned in-camera sRGB preview (input) and corresponding RAW image (target).
- **Resolution protocol.** Following Nam *et al.* [30], we evaluate on the 4× downsampled release to ensure fair comparison with prior work.
- **Split protocol.** We follow the official processed split adopted in Beyond-R2LCM [44] and do not re-split the data.
- **Color space and normalization.** All measurements are performed in raw-linear space with values normalized to $[0, 1]$.

Table A4: Tile-size and Top-K/keep-ratio sensitivity on Sony SLT-A57.

Study	Setting	PSNR	SSIM	Time (ms)
Tile size	$T = 16$	59.60	0.9997	544
	$T = 32$	59.59	0.9997	528
	$T = 64$	59.58	0.9997	515
	$T = 128$	59.50	0.9996	507
Keep ratio	$\rho = 0.25$	59.42	0.9996	482
	$\rho = 0.50$	59.58	0.9997	515
	$\rho = 0.75$	59.60	0.9997	556
	$\rho = 1.00$	59.61	0.9997	584

B More Experimental Results

Before the camera-wise ablations, Table A4 reports the sensitivity to tile size and keep ratio. Tables A5–A7 then provide detailed ablations on three camera subsets at representative rate–distortion operating points ($\lambda \in \{0.02, 0.8, 5.0, 20.0\}$).

B.1 Hyperparameter Sensitivity

Table A4 analyzes the sensitivity of the two key hyperparameters in TileMambaBlock: the tile size T and the keep ratio ρ. For tile size, smaller tiles ($T = 16, 32$) give slightly higher PSNR but incur more overhead, while $T = 128$ is faster but less accurate due to overly coarse energy-based selection; thus, $T = 64$ offers a better balance. For keep ratio, dense processing ($\rho = 1.0$) only improves PSNR over $\rho = 0.5$ by 0.03 dB but costs much more latency, whereas $\rho = 0.25$ is faster but loses informative tiles, so we use $T = 64$ and $\rho = 0.5$ as the default near-dense-accuracy and low-latency configuration.

B.2 Detailed Ablations Across Camera Subsets

Across all three camera subsets, we observe consistent trends: (i) dense SSM + EAR yields strong RD gains but increases latency; (ii) enabling tile-wise selection retains most of the RD improvements while substantially reducing total coding time; and (iii) the relative improvements remain stable from low-rate to high-rate regimes, indicating that the proposed modules are not tuned to a single operating point.

C More Visualization Results

We provide additional qualitative comparisons to further demonstrate the reconstruction fidelity of our proposed MambaRaw framework. Figure A1 presents more visual examples from the Sony SLT-A57 subset. Consistent with the observations in the main text, our method effectively preserves high-frequency details

Table A5: Detailed ablation on Sony SLT-A57 across all λ values.

λ	Config	PSNR (dB)	SSIM	bpp	Enc. (ms)	Dec. (ms)	Total (ms)
0.02	Baseline	56.45	0.9986	0.172	167	321	488
0.02	+ SSM & EAR	57.12	0.9989	0.168	182	340	522
0.02	MambaRaw	57.08	0.9988	0.165	155	302	457
0.8	Baseline	58.21	0.9996	0.376	174	335	509
0.8	+ SSM & EAR	59.61	0.9997	0.365	188	358	546
0.8	MambaRaw	59.58	0.9997	0.361	158	309	467
5.0	Baseline	59.42	0.9997	0.705	191	367	558
5.0	+ SSM & EAR	60.55	0.9998	0.685	205	392	597
5.0	MambaRaw	60.52	0.9998	0.672	175	338	513
20.0	Baseline	59.85	0.9998	1.052	205	394	599
20.0	+ SSM & EAR	61.12	0.9999	0.985	221	420	641
20.0	MambaRaw	61.08	0.9999	0.965	192	365	557

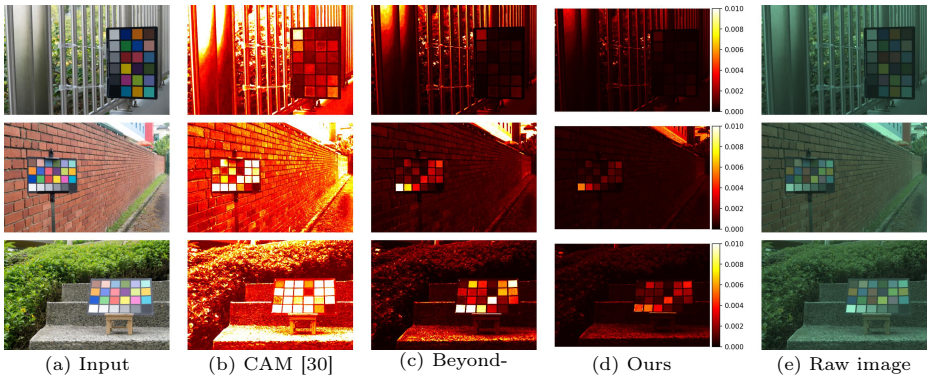
Table A6: Detailed ablation on Samsung NX2000 across all λ values.

λ	Config	PSNR (dB)	SSIM	bpp	Enc. (ms)	Dec. (ms)	Total (ms)
0.02	Baseline	53.85	0.9981	0.188	215	402	617
0.02	+ SSM & EAR	54.45	0.9986	0.182	232	428	660
0.02	MambaRaw	54.42	0.9982	0.178	195	365	560
0.8	Baseline	56.74	0.9996	0.376	228	424	652
0.8	+ SSM & EAR	57.95	0.9997	0.372	245	452	697
0.8	MambaRaw	57.91	0.9997	0.361	208	385	593
5.0	Baseline	57.82	0.9997	0.725	248	465	713
5.0	+ SSM & EAR	58.85	0.9998	0.705	265	495	760
5.0	MambaRaw	58.82	0.9998	0.692	225	422	647
20.0	Baseline	58.25	0.9998	1.125	268	507	775
20.0	+ SSM & EAR	59.45	0.9999	1.085	285	538	823
20.0	MambaRaw	59.41	0.9999	1.065	242	460	702

Table A7: Detailed ablation on Olympus E-PL6 across all λ values.

λ	Config	PSNR (dB)	SSIM	bpp	Enc. (ms)	Dec. (ms)	Total (ms)
0.02	Baseline	56.15	0.9984	0.185	173	334	507
0.02	+ SSM & EAR	57.25	0.9988	0.178	189	356	545
0.02	MambaRaw	57.21	0.9985	0.175	162	315	477
0.8	Baseline	59.04	0.9997	0.376	181	348	529
0.8	+ SSM & EAR	60.52	0.9998	0.368	195	370	565
0.8	MambaRaw	60.45	0.9998	0.361	168	325	493
5.0	Baseline	60.18	0.9998	0.585	195	376	571
5.0	+ SSM & EAR	61.35	0.9999	0.565	210	398	608
5.0	MambaRaw	61.30	0.9999	0.552	180	350	530
20.0	Baseline	60.65	0.9999	0.925	208	401	609
20.0	+ SSM & EAR	62.15	0.9999	0.885	225	428	653
20.0	MambaRaw	62.10	0.9999	0.865	192	375	567

and complex textures, resulting in visibly lower residual errors compared to the baseline methods. This further validates that the spatial-energy coupled context modeling in MambaRaw can robustly capture fine structural information across diverse scenes.



(a) Input (8-bit sRGB image) (b) CAM [30] (bpp: 0.8438) (c) Beyond-R2LCM [44] (bpp: 0.3763) (d) Ours (bpp: 0.3612) (e) Raw image

Fig. A1: Additional qualitative results on Sony SLT-A57 in the same setting as Figure 5. Error maps visualize the per-pixel maximum absolute error over the three channels (after gamma correction for visibility); darker indicates smaller error. We keep the same error scale across methods to enable direct comparison.