MRT: Learning Compact Representations with Mixed RWKV-Transformer for Extreme Image Compression

Han Liu[1], **Hengyu Man**[1,*], **Xingtao Wang**[1,2], **Wenrui Li**[1], **Debin Zhao**[1]

[1]Harbin Institute of Technology
[2]Harbin Institute of Technology Suzhou Research Institute
manhengyu@hotmail.com

Abstract

Recent advances in extreme image compression have revealed that mapping pixel data into highly compact latent representations can significantly improve coding efficiency. However, most existing methods compress images into 2-D latent spaces via convolutional neural networks (CNNs) or Swin Transformers, which tend to retain substantial spatial redundancy, thereby limiting overall compression performance. In this paper, we propose a novel Mixed RWKV-Transformer (MRT) architecture that encodes images into more compact 1-D latent representations by synergistically integrating the complementary strengths of linear-attention-based RWKV and self-attention-based Transformer models. Specifically, MRT partitions each image into fixed-size windows, utilizing RWKV modules to capture global dependencies across windows and Transformer blocks to model local redundancies within each window. The hierarchical attention mechanism enables more efficient and compact representation learning in the 1-D domain. To further enhance compression efficiency, we introduce a dedicated RWKV Compression Model (RCM) tailored to the structure characteristics of the intermediate 1-D latent features in MRT. Extensive experiments on standard image compression benchmarks validate the effectiveness of our approach. The proposed MRT framework consistently achieves superior reconstruction quality at bitrates below 0.02 bits per pixel (bpp). Quantitative results based on the DISTS metric show that MRT significantly outperforms the state-of-the-art 2-D architecture GLC, achieving bitrate savings of 43.75%, 30.59% on the Kodak and CLIC2020 test datasets, respectively.

Code — https://github.com/luke1453lh/MRT

Introduction

The explosive growth of visual data has intensified the demand for advanced image compression techniques capable of preserving high perceptual quality at extremely low bitrates. While traditional compression standards, e.g., H.266/VVC (Bross et al. 2021), and learning-based neural image compression methods (Li et al. 2024a; Jiang et al. 2023a; Man et al. 2023; Cheng et al. 2020; Ballé et al. 2018; Man et al. 2024) primarily optimize for distortion-oriented metrics such as PSNR and MS-SSIM, they often struggle to maintain semantic integrity and visual fidelity under severe bitrate constraints (Blau and Michaeli 2019). To address these limitations, recent compression approaches based on generative models have emerged, leveraging powerful generative priors to enable high-fidelity reconstruction even in extreme-low-bitrate scenarios (Mao et al. 2024; Muckley et al. 2023; Mentzer et al. 2020).

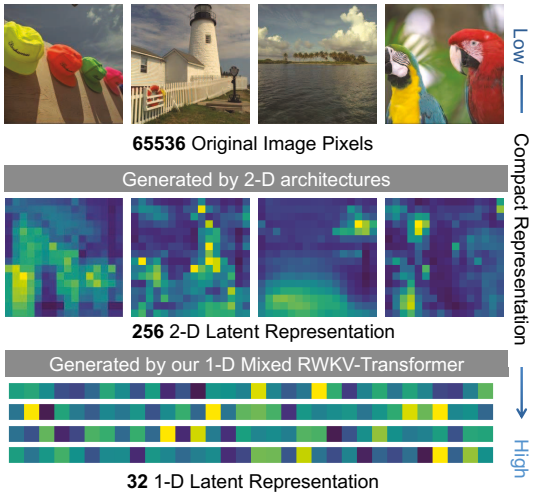

Figure 1: Visualization of original images and heatmaps of latent representations. 2-D architectures(Muckley et al. 2023) compress 65536 pixels into 256 latent features, while redundancy exists between these features. Our 1-D architectures generate more compact latent representations by producing 32 1-D latent features from 65536 pixels.

Most existing image compression methods rely on 2-D neural network architectures (e.g., CNNs and Swin Transformers) that compress images into 2-D latent representations (e.g., mapping a 256×256 image to a 16×16 latent feature map) to reduce redundancy in the representations, which are significantly more compact than the original image space and achieve superior rate-distortion performance compared to pixel-space codecs (Bross et al. 2021; Feng et al. 2025). However, these 2-D representations still exhibit redundancy among neighboring features, leading to subop-

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 2: Qualitative comparison of reconstruction quality across different compression architectures. Our proposed MRT demonstrates superior visual fidelity compared to VTM (Bross et al. 2021) and existing 2-D architectures including LALIC (Feng et al. 2025), DiffEIC (Li et al. 2024b) and MS-ILLM (Muckley et al. 2023). MRT achieves better preservation of fine details even at lower bitrates, while conventional 2-D architectures exhibit significant degradation despite operating at higher bitrates.

timal rate-distortion performance at extremely low bitrates. As illustrated in Figure 1, the latent features produced by 2-D architectures tend to be highly correlated, limiting their efficiency in achieving truly compact representations (Yu et al. 2024).

In response to such limitation, recent studies have investigated 1-D tokenizers based on Vision Transformers (ViTs) (Yu et al. 2024; Bachmann et al. 2025), which transform images into a compact set of 1-D tokens for image generation (e.g., 32 tokens for a 256×256 image as shown in Figure 1). These 1-D architectures have demonstrated remarkable semantic compression capability, enabling compact yet expressive latent representations. However, ViT-based tokenizers are typically tailored for fixed-resolution image synthesis and require extensive fine-tuning to generalize across varying image sizes. In contrast, linear attention models such as RWKV (Duan et al. 2024; Yang et al. 2024; Feng et al. 2025) exhibit robust extrapolation to variable sequence lengths and excel at modeling long-range dependencies. These observations motivate a fundamental question: *How can we effectively combine the complementary strengths of ViTs and RWKV to achieve highly compact and flexible representations for extreme image compression?*

To tackle these challenges, we propose a novel Mixed RWKV-Transformer (MRT) architecture for extremely compact latent representation in image compression. In MRT, each image is partitioned into fixed-size windows. A linear-attention-based RWKV block is employed to capture global dependencies across windows, while ViT blocks are used to model local redundancies within each window. In contrast to previous codecs (Muckley et al. 2023; Mentzer et al. 2020; Jia et al. 2024) which typically compress each window into 2-D feature maps, MRT encodes each window into a set of highly compact 1-D tokens, which are then aggregated to form a global latent representation of the entire image. Considering that existing 2-D compression models are not directly applicable to 1-D tokens due to their distinct structural characteristics, we design a dedicated RWKV Compression Model (RCM), which incorporates multiple Bi-RWKV blocks with spatial and channel mixing modules to effectively eliminate redundancy across both dimensions, enabling efficient compression of the 1D representations and achieving superior rate-distortion performance.

In summary, the main contributions of this work are as follows:

- We propose a novel MRT architecture for extreme image compression, which combines the global modeling capability of RWKV with the local representation strength of ViTs, representing an image as highly compact 1-D latent features.
- We design a dedicated 1-D compression module, RCM, tailored for the intermediate 1-D features in MRT. By incorporating multiple Bi-RWKV blocks, RCM effectively estimates the entropy of latent representations and eliminates redundancy, further improving compression efficiency.
- Extensive experimental evaluations demonstrate that MRT achieves significant bitrate savings of 43.75%, 30.59% on the Kodak and CLIC2020 test datasets, respectively, while maintaining the same DISTS scores to state-of-the-art methods at extremely low bitrates.

Related Work

Generative Image Compression

Recent advances in generative models (e.g., tokenizers, GANs, diffusion models) have enabled superior perceptual quality in extreme image compression scenarios. HIFIC (Mentzer et al. 2020) first integrates conditional GANs into VAE frameworks, while MS-ILLM (Muckley et al. 2023) employs non-binary discriminators for improved statistical fidelity. VQGAN (Esser, Rombach, and

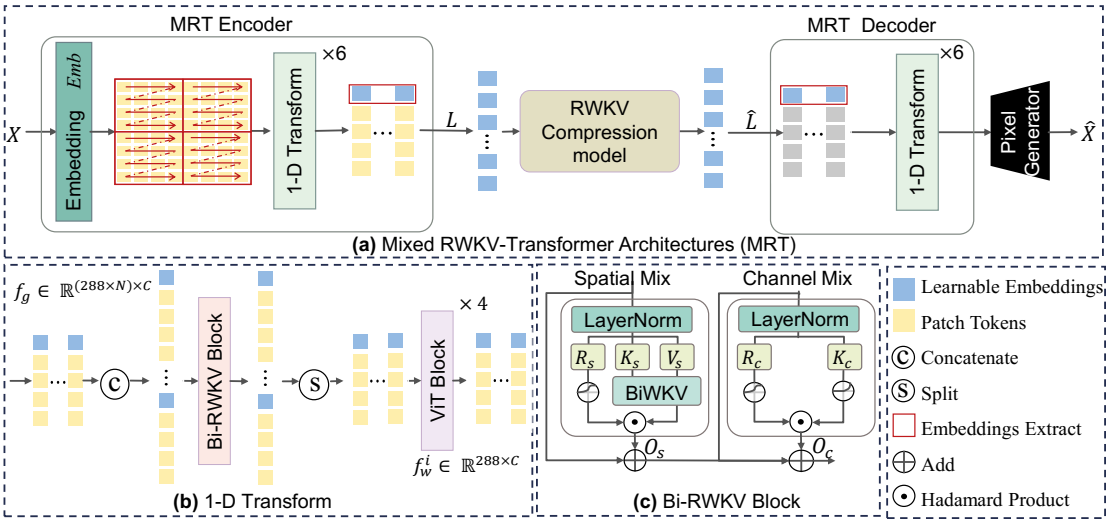

Figure 3: (a) Overview of the proposed MRT, which consists of a MRT encoder, RCM, a MRT decoder and a pixel generator. MRT encoder applies a stack of 1-D transformation layers to extract compact 1-D latent representations, which are then compressed. The decoder mirrors this process to reconstruct the image. (b) The 1-D transform alternates between Bi-RWKV and windowed ViT blocks to capture both global and local dependencies. (c) The Bi-RWKV block comprises spatial and channel mix modules, each utilizing layer normalization, linear projections, and BiWKV attention for efficient sequence modeling.

Ommer 2021) demonstrates strong alignment between latent representations and human perceptual characteristics. Recent works have integrated VQVAE tokenizers into compression frameworks, with VQ-means (Mao et al. 2024) introducing K-means clustering and UIGC (Xue et al. 2024) achieving better rate-perception performance through transformer-based token regeneration. GLC (Jia et al. 2024) further reduces redundancy through latent space compression. Diffusion-based approaches (Li et al. 2025; Yang and Mandt 2024; Körber et al. 2024; Jiang et al. 2023b) have also shown promise in generative compression. However, existing methods rely on 2-D architectures that limit compression efficiency due to spatial redundancy. To address this limitation, we propose a 1-D architecture that encodes images into compact 1-D latent representations, enabling more effective extreme image compression.

Linear Attention

Among emerging linear attention models, Receptance Weighted Key Value (RWKV) (Duan et al. 2024; Yuan et al. 2024) and Mamba (Gu and Dao 2023; Zhu et al. 2024) have demonstrated significant potential for capturing long-range dependencies. This work focuses on RWKV as the foundational component of our MRT architecture. Originally developed for natural language processing (Peng et al. 2023), RWKV employs a novel WKV attention mechanism that excels at modeling long-range dependencies across extended sequences while maintaining linear computational complexity. Recent work has shown RWKV's effectiveness in computer vision tasks, particularly in scenarios requiring global context understanding. Vision RWKV (VRWKV) (Duan et al. 2024) achieves comparable performance to Vision Transformers in image synthesis by effectively capturing long-range spatial relationships. Restore-RWKV (Yang et al. 2024) has established state-of-the-art benchmarks in image restoration by leveraging RWKV's ability to model dependencies across distant image regions, and LALIC (Feng et al. 2025) has successfully integrated RWKV into image compression frameworks, demonstrating superior compression performance through enhanced long-range dependency modeling.

Method

Mixed RWKV-Transformers Architecture

In this paper, we propose a novel framework, MRT, which transforms images into compact 1-D latent representations for more efficient image compression. MRT is composed of three key modules: MRT encoder, RCM, and MRT decoder. The encoder employs multiple 1-D transformation layers designed to capture both inter-window and intra-window dependencies, followed by RCM for latent space compression. The decoder adopts a symmetric architecture as the encoder with a pixel generator to reconstruct images from latent representations. The subsequent sections provide a detailed exposition of each module.

Encoding Given an input image $X \in \mathbb{R}^{3\times H\times W}$, we first use a CNN layer to embed it into patch tokens $Emb(X) \in \mathbb{R}^{w\times h\times c}$, where $w = H/16$ and $h = W/16$. The embedded patch tokens are then partitioned into N non-overlapping 16×16 windows, each of which is flattened to a token sequence $L_i \in \mathbb{R}^{256\times c}$. To compress images into highly compact 1-D latent representations, we follow (Yu et al. 2024) and append a learnable latent embedding $L_{latent} \in \mathbb{R}^{32\times c}$ to each window, forming an extended token sequence $L_w^i \in \mathbb{R}^{288\times c}$, as illustrated in Figure 3(a), where blue squares represent the learnable latent embeddings.

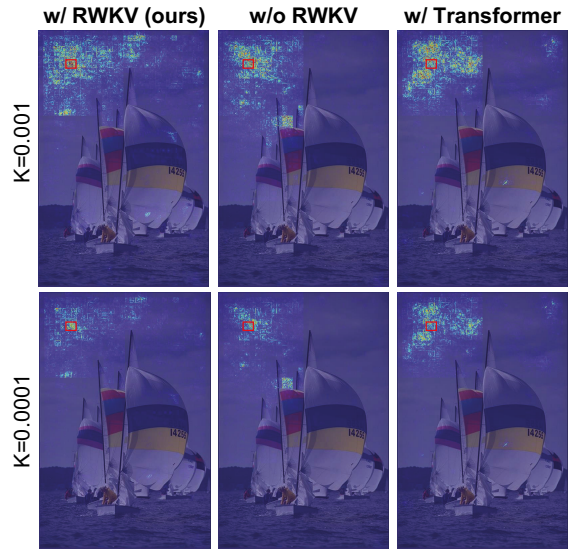

Figure 4: Effective receptive field visualization for different architectural configurations on *kodim09*. Columns: our proposed model with global Bi-RWKV blocks, model without cross-window dependency modeling, and model with global ViT blocks. Rows: gradient thresholds (0.0001 and 0.001).

After that, MRT adopts a 1-D transform module to jointly processe all window tokens, as illustrated in Figure 3 (b). The process consists of three steps: (1) All windows tokens are concatenated together to form a global representation $f_g \in \mathbb{R}^{(288\times N)\times c}$, enabling global context modeling; (2) A Bi-RWKV is applied to capture long-range dependencies across windows within f_g; (3) The transformed global representation is reshaped back to individual windows and passed through ViT blocks for local spatial modeling, yielding $f_w^i \in \mathbb{R}^{288\times c}$.

Finally, we extract the learnable embeddings L_{latent} from each processed window and concatenate them to form a highly compact 1-D representation $L \in \mathbb{R}^{(N\times 32)\times c}$, which serves as the final latent representation of the entire image and is subsequently compressed by the RCM, as illustrated in Figure 3 (a).

Decoding On the decoder side, given the quantized representation $\hat{L} \in \mathbb{R}^{(N\times 32)\times c}$, we partition it into N segments. Each segment is concatenated with mask features, which share the same shape as flattened features $L_w^i \in \mathbb{R}^{288\times c}$, to reconstruct window representation $\hat{L}_w^i \in \mathbb{R}^{288\times c}$. The windows are processed through 1-D transformation layers: Bi-RWKV for global dependencies, then ViT for local relationships. After removing embeddings, features are rearranged to 2-D map $f \in \mathbb{R}^{w\times h\times c}$, and pixel generator G produces output $\hat{X} \in \mathbb{R}^{3\times H\times W}$.

Bi-RWKV Block To efficiently capture cross-window dependencies and support variable-resolution images, we adopt a Bi-RWKV block that consists of two primary components: spatial mix and channel mix. Firstly, the spatial mix module applies layer normalization and three linear projections to generate receptance R_s, key K_s, and value V_s. After that, given K_s and V_s, global attention is computed using the Bi-directional Weighted Key-Value (BiWKV) mechanism, modulated by sigmoid-gated receptance $\theta(R_s)$ to produce spatial output O_s. The BiWKV attention for the t-th token is defined as:

$$\mathrm{wkv}_t = \frac{\sum_{i=1,\, i\neq t}^{T} e^{-\frac{|t-i|-1}{T}\cdot w + k_i} v_i + e^{u+k_t} v_t}{\sum_{i=1,\, i\neq t}^{T} e^{-\frac{|t-i|-1}{T}\cdot w + k_i} + e^{u+k_t}} \tag{1}$$

where k_i and v_i denote the key and value of the i-th token, and T represents the sequence length of 1-D representations. This enables linear computational complexity for long-range dependencies. Finally, the channel mix module takes O_s as input and adopts two linear projections to produce receptance R_c and key K_c. A squared ReLU activation is applied to K_c, and the final channel output O_c is obtained via sigmoid-gated modulation, analogous to the spatial pathway.

To evaluate Bi-RWKV's effectiveness in modeling long-range dependencies, we conduct an effective receptive field (ERF) analysis following (Liu, Sun, and Katto 2023). We compare Bi-RWKV against two baselines: (1) replacing global Bi-RWKV blocks with global RoPE-ViT blocks and (2) removing global Bi-RWKV blocks entirely. The ERF is computed as the absolute gradient magnitude with respect to a target region (marked by a red square). We clip the top K gradients to better visualize global gradients. As shown in Figure 4, Bi-RWKV achieves superior global dependency capture with broader and more uniform gradient distribution across the entire image. In contrast, models without Bi-RWKV blocks show localized attention patterns concentrated around the target area, while models with global RoPE-ViT blocks exhibit incomplete cross-window interaction. The visualization confirms that Bi-RWKV effectively models long-range dependencies for comprehensive spatial coverage in image compression.

RWKV Compression Model

Compared to conventional 2-D latent representations produced by 2-D architectures (Muckley et al. 2023; Mentzer et al. 2020; Jia et al. 2024), our proposed 1-D latent representations are highly compact. However, due to the absence of spatial priors typically leveraged in 2-D representations, 1-D latent do not support straightforward downsampling for hyperprior extraction. To address this challenge, we propose a dedicated RWKV compression model (RCM) that directly operates on the 1-D latent features. As depicted in Figure 5, the 1-D latent representations are first projected into a higher-dimensional space and then transformed into a latent variable $y \in \mathbb{R}^{(N\times 32)\times c_y}$, where $c_y = 320$. The variable y is subsequently quantized to obtain $\hat{y}$. To model the distributional properties of y, a hyper network is employed to generate hyper representations $z \in \mathbb{R}^{(N\times 32)\times c_z}$ by processing y through two stacked Bi-RWKV blocks, followed by a multilayer perceptron (MLP) that projects the output into a lower-dimensional hyper space. To maximize codebook utilization, we employ Look-up Free Quantization

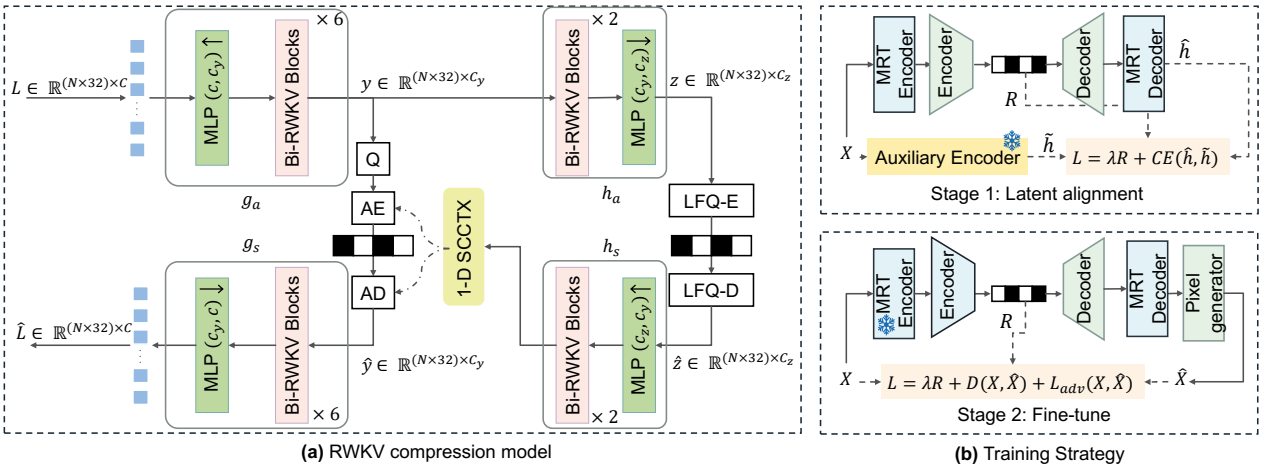

Figure 5: (a) RWKV compression model architecture. MLP$(c_1, c_2)\uparrow$ and MLP$(c_2, c_1)\downarrow$ denote dimensionality expansion and reduction, respectively. AE, AD, and Q represent arithmetic encoding, arithmetic decoding, and quantization. LFQ-E, LFQ-D stand for look-up free (LFQ) encoding and decoding. 1-D SCCTX denotes the 1-D Spatial-Channel Context Module. (b) Two-stage training strategy. Stage 1: auxiliary encoder aligns quantized codes with discrete targets via cross-entropy loss. Stage 2: RCM and MRT decoder are optimized end-to-end using pixel-level and perceptual losses.

(LFQ) to quantize the hyper representations, rather than the prior VQ-based hyper modules (Jia et al. 2024). We define LFQ as:

$$\hat{z} = q(z) = 2 \cdot \mathbb{I}[z \geq 0] - 1. \tag{2}$$

In our implementation, we set $c_z = 14$, resulting in a codebook of size 16384 for the binary hyper representations. To further enhance the compression efficiency, we introduce a 1-D Spatial-Channel Context Module (SCCTX) to estimate the entropy of the quantized hyper representations. SCCTX facilitates more accurate probability modeling for entropy coding, thereby reducing the overall bitrate. Finally, the quantized latent representations $\hat{y}$ are passed to the MRT decoder to reconstruct the compact 1-D latent features $\hat{L} \in \mathbb{R}^{(N\times32)\times c}$.

Two-Stage Training Strategy

Inspired by recent works (Yu et al. 2024; Sargent et al. 2025), we train MRT in a two-stage manner, including latent alignment and decoder fine-tuning, as shown in Figure 5(b).

Latent Alignment. Instead of training the entire model with pixel-level loss functions, we align the codes $\hat{h}$ with the discrete codes $\tilde{h}$ generated by a pre-trained VQGAN model (Chang et al. 2022) using cross-entropy loss denoted $CE(\hat{h}, \tilde{h})$ to measure the reconstruction loss. In order to make the RCM module more effective in compressing the intermediate 1-D latent, multiple dedicated loss function are also introduced. Firstly, for stability, we also introduce a space alignment loss to minimize the distortion from the RCM (Li et al. 2024b):

$$\mathcal{L}_{\text{latent}} = \mathcal{L}_{MSE}(\hat{L}, L) \tag{3}$$

Additionally, following (Yu et al. 2023), an auxiliary loss is introduced to maximize the usage of the LFQ codebook, which is defined as:

$$\mathcal{L}_{\text{ent}} = \mathbb{E}\left[H(q(\hat{z})) - H\left(\mathbb{E}[q(\hat{z})]\right)\right], \tag{4}$$

$$\mathcal{L}_{\text{commit}} = \mathbb{E}\left[\|\hat{z} - q(\hat{z})\|_2^2\right]. \tag{5}$$

$$\mathcal{L}_{\text{LFQ}} = \lambda_{ent} \cdot \mathcal{L}_{\text{ent}} + \lambda_{commit} \cdot \mathcal{L}_{\text{commit}}. \tag{6}$$

where λ_{ent} and λ_{commit} are the weights of the entropy loss and commitment loss, respectively. We set $\lambda_{ent} = 0.25$ and $\lambda_{commit} = 0.00625$. Finally, we optimize the estimated entropy of the representation y, resulting in R. Overall, we train MRT encoder and MRT decoder and RCM with the following loss function:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{LFQ}} + \mathcal{R}(\hat{y}) + \mathcal{L}_{\text{latent}} \tag{7}$$

In stage 1, we do not modify the lambda of rates to achieve different bitrates.

Decoder Fine-tuning. In the decoder fine-tuning stage, we jointly optimize the RCM, MRT decoder and pixel generator using pixel-level supervision. Specifically, we employ a combination of pixel-wise ℓ_1 loss, perceptual loss (Simonyan and Zisserman 2014), and adversarial loss (Yu et al. 2024) to enhance both the fidelity and perceptual quality of the reconstructed images. The overall objective for this stage is formulated as:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_1 + \mathcal{L}_{\text{perceptual}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda\mathcal{R}(\hat{y}) \tag{8}$$

where $\mathcal{L}_1$ denotes the pixel-wise ℓ_1 loss, $\mathcal{L}_{\text{perceptual}}$ is the widely used VGG-based perceptual loss (Simonyan and Zisserman 2014), $\mathcal{L}_{\text{adv}}$ is the adversarial loss, and $\mathcal{R}$ represents exstimated bitrates. The hyperparameters λ_{adv} and λ are used to balance the contributions of the adversarial loss and the rate, respectively. In our experiments, we set $\lambda_{\text{adv}} = 0.05$ by default.

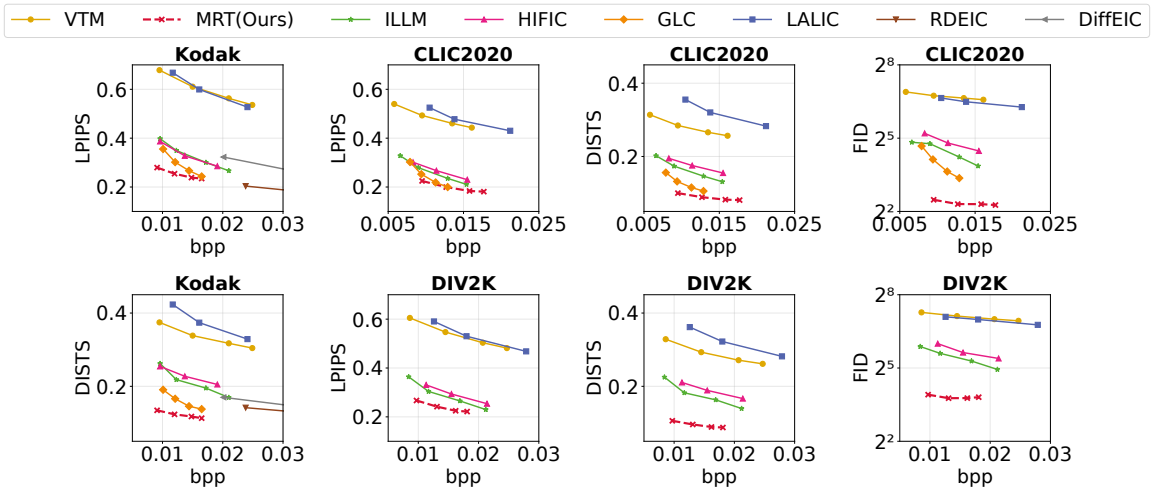

Figure 6: Rate-distortion curves on the Kodak, the CLIC2020 and the DIV2K datasets.

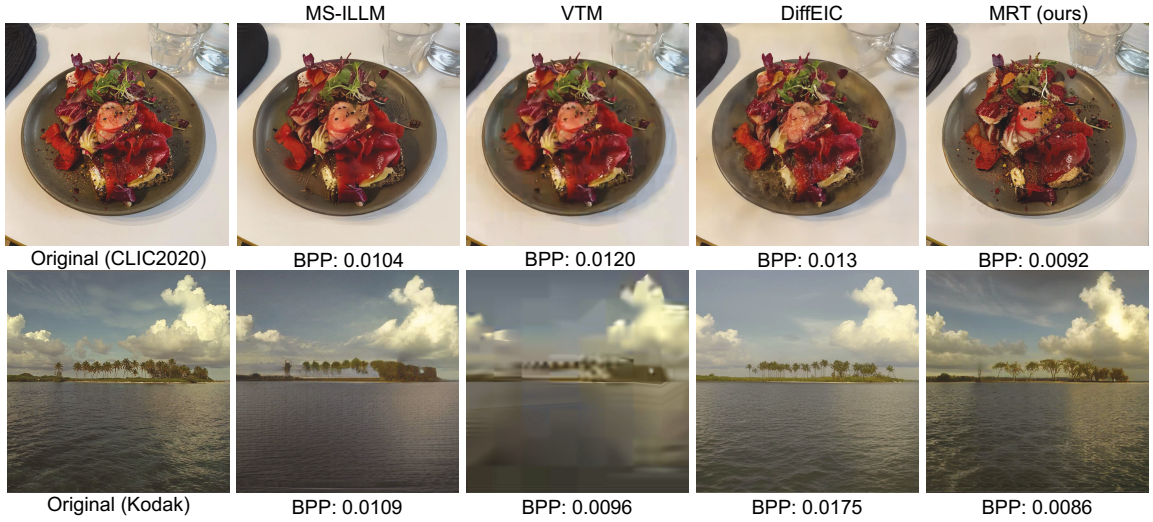

Figure 7: Qualitative examples of different methods on Kodak and CLIC2020. Zoom in for better visualization.

Experiments

Implementation Details

Training Details. Our MRT model is trained on the OpenImages training set (Krasin et al. 2017) following a two-stage training strategy. In the first stage, we train the model using randomly cropped 512×512 image patches with a batch size of 4, initialized with a pre-trained 1-D tokenizer (Yu et al. 2024) to accelerate convergence. In the second stage, we fine-tune the MRT decoder, RCM, and pixel generator using the same patch size and batch size. To achieve different rate-distortion trade-offs, we set the rate-distortion parameter λ to $\{20, 10, 5, 2.5\}$. All models are trained on a single NVIDIA RTX 5090 GPU.

Evaluation Datasets. We evaluate our MRT model on three widely-used image compression benchmarks: the Kodak dataset (Franzen 1999), CLIC2020 test set (Toderici et al. 2020), and DIV2K validation set (Agustsson and Timofte 2017). We utilize full-resolution images to assess performance across diverse image characteristics. More results are provided in the extended version(Liu et al. 2025).

Evaluation Metrics. We employ multiple evaluation metrics to assess performance. For reconstruction fidelity, we use perceptual metrics LPIPS (Zhang et al. 2018) and DISTS (Ding et al. 2020). Generation realism is quantified using FID (Heusel et al. 2017). Bitrate is calculated in bits per pixel (bpp). Following (Mentzer et al. 2020), we do not report FID for Kodak due to its limited size. Additional PSNR and MS-SSIM (Wang, Simoncelli, and Bovik 2003) results are provided in the extended version(Liu et al. 2025).

Comparison Methods. We compare our method with state-of-the-art image compression approaches. Traditional baselines include VTM-23.11 (Bross et al. 2021) and the RWKV-based neural codec LALIC (Feng et al. 2025). For generative compression, we evaluate against

Model variants	Kodak	CLIC2020	DIV2K
MRT w/o RWKV	9.62%	32.41%	25.76%
RCM w/ ViT	6.99%	8.61%	9.08%
RCM w/ VQ	34.52%	40.66%	40.40%
Ours	0.0%	0.0%	0.0%

Table 1: BD-Rate on DISTS metric of different model variants. We use the proposed methods as the anchor, denoted as **Ours**.

HiFiC (Mentzer et al. 2020), MS-ILLM (Muckley et al. 2023), GLC (Jia et al. 2024), and Diffusion-based codecs DiffEIC (Li et al. 2024b), RDEIC (Li et al. 2025).

Main Results

Quantitative Evaluation Figure 6 shows rate-distortion curves across datasets and metrics (Bjontegaard 2001). MRT achieves the best rate-distortion trade-offs on three datasets (Kodak, CLIC2020, DIV2K) and three metrics (LPIPS, DISTS, FID). On Kodak, MRT achieves 19.82% and 43.75% bitrate savings for LPIPS and DISTS compared to GLC, outperforming diffusion-based methods. On CLIC2020, MRT maintains 30.59% bitrate saving for DISTS. MRT achieves competitive FID performance on both DIV2K and CLIC2020 datasets.

Qualitative Evaluation Figure 7 shows reconstruction quality comparisons. On CLIC2020 image, MRT (0.0092 bpp) maintains sharp textures and fine details, while MS-ILLM shows texture softening, VTM exhibits blurring, and DiffEIC loses details. On Kodak image, MRT (0.0086 bpp) produces clear reconstructions and preserves textures. VTM suffers from banding artifacts and blurring, while MS-ILLM shows reduced sharpness. DiffEIC requires higher bitrate with mismatch. MRT generates more visually pleasing results at extreme bitrates.

Ablation Study

In this section, we conduct ablation studies to analyze the contribution of MRT, Bi-RWKV compression model and look-up free quantization. All model variations are evaluated on the Kodak, CLIC2020 and DIV2K datasets using DISTS.

Mixed RWKV-Transformer Architecture. We evaluate the contribution of global Bi-RWKV blocks by removing them and keeping only local ViT blocks (denoted as *MRT w/o RWKV*). Table 1 shows bitrate increases of 9.62%, 32.41%, and 25.76% on Kodak, CLIC2020, and DIV2K datasets, respectively. The more pronounced performance degradation on CLIC2020 and DIV2K (high-resolution images) highlights MRT's superior capability in compressing high-resolution images without requiring fine-tuning.

RWKV Compression Model. We replace Bi-RWKV blocks with RoPE-ViT blocks in RCM (denoted as *RCM w/ ViT*). Table 1 shows bitrate increases of 6.99%, 8.61%, and 9.08% on the three datasets, indicating that Bi-RWKV blocks are more effective for modeling long-range dependencies in the compression domain.

Quantization method	Kodak	CLIC2020	DIV2K
MRT w/ VQ	5.42	5.65	5.68
MRT w/ LFQ	7.59	10.38	10.39

Table 2: Entropy of the codebook of different quantization methods. Larger entropy indicates more effective usage of the codebook.

Model variants	Latency (s)		BD-DISTS
	Enc.	Dec.	
MS-ILLM	**0.0350**	**0.0304**	0.00%
DiffEIC	0.1456	3.9079	-34.52%
Ours	0.1641	0.1909	**-77.64%**

Table 3: Complexity analysis and BD-DISTS of different methods. All tests are conducted on the Kodak dataset with a single NVIDIA A100 GPU.

Look-up Free Quantization. We compare LFQ against conventional vector quantization (denoted as *RCM w/ VQ*). As observed in Table 1, replacing VQ with LFQ results in significant bitrate reductions of 34.52%, 40.68%, and 42.52% across three benchmarks. Additionally, as illustrated in Table 2, LFQ achieves higher entropy than VQ, indicating superior codebook utilization. This advantage facilitates more complex distribution modeling and enable more efficient compression in the 1-D latent space.

Complexity Analysis

We analyze the computational complexity of MRT. As shown in Table 3, MRT has longer encoding and decoding time than traditional 2-D methods (e.g., MS-ILLM), but offers superior performance that justifies this trade-off. Compared to diffusion-based methods (e.g., DiffEIC), MRT achieves better rate-distortion performance while maintaining similar encoding time and significantly reducing decoding time. Additional results are provided in the extended version(Liu et al. 2025).

Conclusion

We propose a novel Mixed RWKV-Transformer (MRT) architecture for extreme image compression. By combining ViT blocks for local modeling and Bi-RWKV blocks for global modeling, MRT efficiently encodes images into compact 1-D representations. Additionally, we design a dedicated RWKV compression model (RCM) to compress these representations effectively. Experimental results demonstrate that MRT achieves superior rate-distortion performance across multiple datasets at extreme bitrates below 0.02 bpp. Future work will focus on optimizing the architectural design and training strategies to achieve comparable performance while reducing computational complexity.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2023YFA1008500), the National Natural Science Foundation of China (NSFC) under grants U22B2035 and 62502116, and China Post-Doctoral Science Foundation under Grant 2025M774315.

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Bachmann, R.; Allardice, J.; Mizrahi, D.; Fini, E.; Kar, O. F.; Amirloo, E.; El-Nouby, A.; Zamir, A.; and Dehghan, A. 2025. FlexTok: Resampling Images into 1D Token Sequences of Flexible Length. In *Forty-second International Conference on Machine Learning*.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*.

Bjontegaard, G. 2001. Calculation of average PSNR differences between RD-curves. *ITU SG16 Doc. VCEG-M33*.

Blau, Y.; and Michaeli, T. 2019. Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 675–685. PMLR.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 3736–3764.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11315–11325.

Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020. Learned Image Compression With Discretized Gaussian Mixture Likelihoods and Attention Modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2567–2581.

Duan, Y.; Wang, W.; Chen, Z.; Zhu, X.; Lu, L.; Lu, T.; Qiao, Y.; Li, H.; Dai, J.; and Wang, W. 2024. Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv preprint arXiv:2403.02308*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Feng, D.; Cheng, Z.; Wang, S.; Wu, R.; Hu, H.; Lu, G.; and Song, L. 2025. Linear Attention Modeling for Learned Image Compression. *arXiv preprint arXiv:2502.05741*.

Franzen, R. 1999. Kodak PhotoCD dataset.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jia, Z.; Li, J.; Li, B.; Li, H.; and Lu, Y. 2024. Generative latent coding for ultra-low bitrate image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26088–26098.

Jiang, W.; Yang, J.; Zhai, Y.; Ning, P.; Gao, F.; and Wang, R. 2023a. Mlic: Multi-reference entropy model for learned image compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7618–7627.

Jiang, X.; Tan, W.; Tan, T.; Yan, B.; and Shen, L. 2023b. Multi-modality deep network for extreme learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1033–1041.

Krasin, I.; Duerig, T.; Alldrin, N.; Ferrari, V.; Abu-El-Haija, S.; Kuznetsova, A.; Rom, H.; Uijlings, J.; Popov, S.; Veit, A.; Belongie, S.; Gomes, V.; Gupta, A.; Sun, C.; Chechik, G.; Cai, D.; Feng, Z.; Narayanan, D.; and Murphy, K. 2017. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*.

Körber, N.; Kromer, E.; Siebert, A.; Hauke, S.; Mueller-Gritschneder, D.; and Schuller, B. 2024. PerCo (SD): Open Perceptual Compression. arXiv:2409.20255.

Li, D.; Bai, Y.; Wang, K.; Jiang, J.; Liu, X.; and Gao, W. 2024a. GroupedMixer: An Entropy Model with Group-wise Token-Mixers for Learned Image Compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Li, Z.; Zhou, Y.; Wei, H.; Ge, C.; and Jiang, J. 2024b. Towards Extreme Image Compression with Latent Feature Guidance and Diffusion Prior. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, Z.; Zhou, Y.; Wei, H.; Ge, C.; and Mian, A. 2025. RDEIC: Accelerating Diffusion-Based Extreme Image Compression with Relay Residual Diffusion. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, H.; Man, H.; Wang, X.; Li, W.; and Zhao, D. 2025. MRT: Learning Compact Representations with Mixed RWKV-Transformer for Extreme Image Compression. *arXiv preprint arXiv:2511.06717*.

Liu, J.; Sun, H.; and Katto, J. 2023. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14388–14397.

Man, H.; Fan, X.; Lu, R.; Yu, C.; and Zhao, D. 2024. MetaIP: Meta-Network-Based Intra Prediction With Customized Parameters for Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10): 9591–9605.

Man, H.; Fan, X.; Xiong, R.; and Zhao, D. 2023. Tree-Structured Data Clustering-Driven Neural Network for Intra Prediction in Video Coding. *IEEE Transactions on Image Processing*, 32: 3493–3506.

Mao, Q.; Yang, T.; Zhang, Y.; Wang, Z.; Wang, M.; Wang, S.; Jin, L.; and Ma, S. 2024. Extreme image compression using fine-tuned vqgans. In *2024 Data Compression Conference (DCC)*, 203–212. IEEE.

Mentzer, F.; Toderici, G. D.; Tschannen, M.; and Agustsson, E. 2020. High-fidelity generative image compression. *Advances in neural information processing systems*, 33: 11913–11924.

Muckley, M. J.; El-Nouby, A.; Ullrich, K.; Jégou, H.; and Verbeek, J. 2023. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, 25426–25443. PMLR.

Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Biderman, S.; Cao, H.; Cheng, X.; Chung, M.; Grella, M.; et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Sargent, K.; Hsu, K.; Johnson, J.; Fei-Fei, L.; and Wu, J. 2025. Flow to the mode: Mode-seeking diffusion autoencoders for state-of-the-art image tokenization. *arXiv preprint arXiv:2503.11056*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Toderici, G.; Shi, W.; Timofte, R.; Theis, L.; Balle, J.; Agustsson, E.; Johnston, N.; and Mentzer, F. 2020. Workshop and Challenge on Learned Image Compression (CLIC2020).

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, 1398–1402. Ieee.

Xue, N.; Mao, Q.; Wang, Z.; Zhang, Y.; and Ma, S. 2024. Unifying generation and compression: Ultra-low bitrate image coding via multi-stage transformer. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Yang, R.; and Mandt, S. 2024. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36.

Yang, Z.; Li, J.; Zhang, H.; Zhao, D.; Wei, B.; and Xu, Y. 2024. Restore-rwkv: Efficient and effective medical image restoration with rwkv. *arXiv preprint arXiv:2407.11087*.

Yu, L.; Lezama, J.; Gundavarapu, N. B.; Versari, L.; Sohn, K.; Minnen, D.; Cheng, Y.; Birodkar, V.; Gupta, A.; Gu, X.; et al. 2023. Language Model Beats Diffusion–Tokenizer is Key to Visual Generation. *arXiv preprint arXiv:2310.05737*.

Yu, Q.; Weber, M.; Deng, X.; Shen, X.; Cremers, D.; and Chen, L.-C. 2024. An image is worth 32 tokens for reconstruction and generation. *Advances in Neural Information Processing Systems*, 37: 128940–128966.

Yuan, H.; Li, X.; Qi, L.; Zhang, T.; Yang, M.-H.; Yan, S.; and Loy, C. C. 2024. Mamba or RWKV: Exploring High-Quality and High-Efficiency Segment Anything Model. *arXiv preprint*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.