DAEM: Spectral-Deformable Entropy Modeling for Scalable Learned Image Compression

Yiming Ding
College of Intelligence and Computing
Tianjin University
Tianjin, China
yiming_ding@tju.edu.cn

Jianguo Wei*
College of Intelligence and Computing
Tianjin University
Tianjin, China
jianguo@tju.edu.cn

***Abstract*—Context modeling is a fundamental component distinguishing modern learned image compression (LIC) from classical transform codecs. However, state-of-the-art entropy models increasingly rely on global self-attention, the quadratic complexity of which becomes prohibitive for high-resolution images. Simultaneously, local context modules are often constrained by fixed grid sampling, which is hindered by misaligned structures and window boundary artifacts. To overcome these limitations, we propose Dual-Aware Entropy Modeling (DAEM), a plug-and-play framework that integrates spectral global context with offset-adapted local context. For global modeling, we employ content-adaptive spectral filtering in the frequency domain via FFT on causal side information (hyperprior features and previously decoded slices), achieving an efficient global receptive field with $O(n \log n)$ complexity. For local modeling, we introduce an offset-adapted context module that predicts content-conditioned sampling offsets atop overlapped window attention. This mechanism improves structural alignment for non-anchor symbols while preserving parallel decoding capabilities. Extensive evaluations on Kodak, Tecnick, and CLIC Professional Validation show that DAEM achieves PSNR-based BD-rate reductions of 12.67% (Tecnick) and 10.28% (CLIC) relative to VTM-17.0, and attains the best MS-SSIM BD-rate on Tecnick (-47.55%).**

***Index Terms*—Learned image compression, entropy modeling, frequency domain, deformable context**

I. Introduction

The exponential growth of high-resolution visual content imposes significant demands on storage and transmission infrastructure. While classical image and video codecs (e.g., JPEG2000, BPG, VVC) [1]–[3] rely on hand-crafted transforms and explicit probability models, their capacity to capture complex, content-adaptive dependencies remains limited. Learned image compression (LIC) addresses these constraints by substituting fixed components with end-to-end optimized neural transforms and learned entropy models [4], [5]. Among these components, entropy modeling is critical, as a more accurate estimation of the conditional distribution of latents directly translates to reduced bitrates for a given reconstruction quality.

A prevailing trend in LIC entropy modeling is the incorporation of increasingly rich contexts. Hyperpriors utilize global side information [5], while autoregressive models exploit previously decoded symbols [6]. Recent architectures integrate multi-slice, checkerboard, and attention-based mechanisms to capture channel-wise, local, and global dependencies with enhanced parallelism [7]–[10]. Despite these advancements, two significant challenges impede the deployment of these models in scalable, high-resolution compression scenarios.

First, global context modeling becomes computationally intractable at high resolutions. While attention-based mechanisms offer robust long-range dependency capture, their $O(n^2)$ complexity relative to the number of latent positions $n{=}HW$ scales poorly. For instance, a 4K resolution image generates tens of thousands of latent tokens, rendering global attention unfeasible within practical memory and latency constraints. Although approximate attention variants exist, they typically compromise modeling capacity, a detrimental trade-off for rate-critical entropy coding.

Second, local context modeling is often compromised by rigid sampling and boundary discontinuities. Architectures utilizing windowed attention or masked convolution achieve parallel decoding but operate on fixed grid locations. In the context of compression, where latents represent aggregated receptive fields, semantic features such as edges, thin structures, and repeating textures frequently exhibit misalignment with these fixed grids. Consequently, rigid sampling yields suboptimal local correspondence, while non-overlapping windows introduce boundary artifacts that degrade both bit allocation efficiency and visual quality.

To address these issues, we introduce Dual-Aware Entropy Modeling (DAEM), a unified framework that is frequency-aware for efficient global context and offset-aware for precise local context. Our approach is grounded in the insight that global dependencies can be modeled efficiently via spectral filtering: the convolution theorem establishes that element-wise multiplication in the frequency domain is equivalent to global convolution in the spatial domain. Unlike attention, spectral filtering provides a global receptive field with near-linear complexity and stable memory scaling. Furthermore, we observe that difficult local coding scenarios often stem from geometric misalignment. We therefore predict content-conditioned offsets to adaptively re-sample local context for non-anchor symbols, utilizing overlapped window attention to eliminate boundary discontinuities.

Our contributions are summarized as follows:

*Corresponding author: jianguo@tju.edu.cn

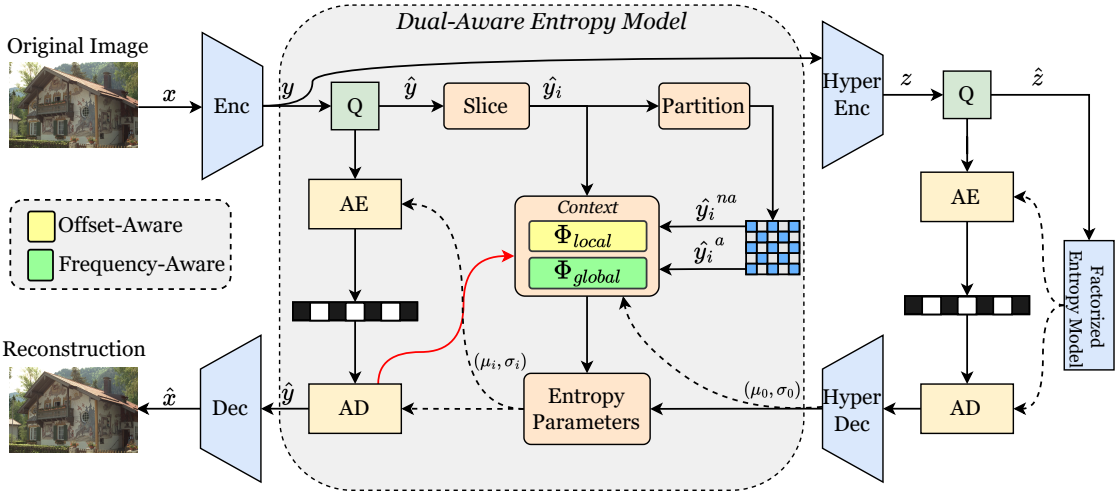

Fig. 1. Overall framework of DAEM. The entropy model predicts Gaussian parameters (μ, σ) for arithmetic coding (AE/AD). DAEM comprises Spectral Global Context (SGC) operating on causal side information, and Offset-Adapted Local Context (OLC) operating on decoded anchors for non-anchor prediction.

- We propose spectral global context for LIC entropy modeling: an FFT-based, content-adaptive frequency filtering module applied to causal side information, yielding a global receptive field with $O(n \log n)$ complexity.
- We introduce an offset-adapted local context module that learns content-conditioned sampling offsets over overlapped window attention, enhancing structural alignment for parallel non-anchor decoding.
- DAEM improves the rate–distortion–speed trade-off on the Kodak, Tecnick, and CLIC Professional Validation datasets, outperforming strong LIC baselines and reducing encoding latency compared to attention-heavy entropy models.

II. Related Work

A. Entropy Models for LIC

Hyperprior-based LIC [5] augments a factorized density model with side information, while joint autoregressive-hyperprior models further refine conditional entropy by exploiting decoded context [6]. To accelerate serial autoregression, parallelizable context schedules have been developed, including checkerboard context modeling [7] and channel-wise autoregressive entropy models [8]. Multi-context and multi-reference entropy models combine cross-slice, local, and global contexts to maximize coding gain [9]. Transformer-based entropy models (e.g., Entroformer) enhance long-range modeling but often incur substantial computational and memory overheads [10].

B. Frequency-Domain Modeling in Learned Compression

Recent research has revisited insights from classical transform coding. Frequency-aware transformer blocks have been proposed to better capture directional and multi-scale frequency components within LIC transforms [11]. Similarly, WeConvene embeds wavelet transforms into both convolution layers and entropy coding [12]. These approaches typically modify the analysis/synthesis transforms or perform coding directly in wavelet subbands. In contrast, DAEM treats frequency modeling as a plug-in component for the entropy model, preserving the backbone codec while utilizing spectral filtering to construct an efficient global context prior from causal side information.

III. Method

A. Notation and Codec Overview

Given an input image $\mathbf{x}$, an analysis transform g_a produces the latent representation $\mathbf{y}{=}g_a(\mathbf{x})$. Quantization yields $\hat{\mathbf{y}}{=}Q(\mathbf{y})$, which is losslessly encoded via arithmetic coding using a learned probability model $p(\hat{\mathbf{y}})$. A hyperprior transform [5] encodes side information: $\mathbf{z}{=}h_a(\mathbf{y})$, $\hat{\mathbf{z}}{=}Q(\mathbf{z})$, and the hyper-decoder h_s generates hyper features $\boldsymbol{\phi}{=}h_s(\hat{\mathbf{z}})$.

Following prior work [7], [9], we partition $\hat{\mathbf{y}}$ into S channel slices $\{\hat{\mathbf{y}}^{(i)}\}_{i=1}^{S}$ and apply a checkerboard split within each slice: $\hat{\mathbf{y}}^{(i)} = \{\hat{\mathbf{y}}_A^{(i)}, \hat{\mathbf{y}}_{\bar{A}}^{(i)}\}$, where A and $\bar{A}$ denote anchor and non-anchor positions, respectively. Anchors are encoded/decoded first in parallel and subsequently serve as context for non-anchors. The conditional probability factorization is given by:

$$p(\hat{\mathbf{y}}|\hat{\mathbf{z}}) = \prod_{i=1}^{S} p(\hat{\mathbf{y}}_A^{(i)} \mid \hat{\mathbf{z}}, \hat{\mathbf{y}}^{(<i)})\; p(\hat{\mathbf{y}}_{\bar{A}}^{(i)} \mid \hat{\mathbf{z}}, \hat{\mathbf{y}}^{(<i)}, \hat{\mathbf{y}}_A^{(i)}), \quad (1)$$

where $\hat{\mathbf{y}}^{(<i)}$ denotes all previously decoded slices.

Each symbol is modeled by a Gaussian convolved with a unit uniform distribution [5]:

$$p(\hat{y} \mid \mu, \sigma) = \big(\mathcal{N}(\mu, \sigma^2) * \mathcal{U}(-\tfrac{1}{2}, \tfrac{1}{2})\big)(\hat{y}), \quad (2)$$

with parameters (μ, σ) predicted by our entropy model.

B. Dual-Aware Entropy Model (DAEM)

DAEM predicts distribution parameters for anchors and non-anchors utilizing two complementary context sources:

(i) Spectral Global Context (SGC) $\Phi_g^{(i)}$, derived from hyper features and previously decoded slices; and (ii) Offset-Adapted Local Context (OLC) $\Phi_l^{(i)}$, computed from decoded anchors within the current slice.

We first construct a causal side-information tensor for slice i:

$$\mathbf{t}^{(i)} = \mathrm{Concat}\big(\boldsymbol{\phi}^{(i)},\ \psi(\hat{\mathbf{y}}^{(<i)})\big), \tag{3}$$

where $\boldsymbol{\phi}^{(i)}$ denotes the slice-wise hyper features (a channel projection of $\boldsymbol{\phi}$), and $\psi(\cdot)$ is a lightweight projection aggregating previously decoded slices (implemented as a 1×1 convolution over concatenated slices; for $i{=}1$, this is initialized as zeros).

The anchor parameters are predicted by:

$$(\mu_A^{(i)}, \sigma_A^{(i)}) = f_A\big(\mathbf{t}^{(i)}, \Phi_g^{(i)}\big), \tag{4}$$

and the non-anchor parameters are predicted by:

$$(\mu_{\bar{A}}^{(i)}, \sigma_{\bar{A}}^{(i)}) = f_{\bar{A}}\big(\mathbf{t}^{(i)}, \Phi_g^{(i)}, \Phi_l^{(i)}\big), \tag{5}$$

where f_A and $f_{\bar{A}}$ are convolutional predictors.

C. Spectral Global Context (SGC)

Given $\mathbf{t}^{(i)} \in \mathbb{R}^{B\times C\times H\times W}$, we apply a 2D real FFT (rFFT) along the spatial dimensions:

$$\mathcal{F}(\mathbf{t}^{(i)}) \in \mathbb{C}^{B\times C\times H\times (\lfloor \frac{W}{2} \rfloor + 1)}. \tag{6}$$

We introduce N learnable complex-valued spectral filters $\{\mathbf{G}_n\}_{n=1}^N$, where each filter is defined on a reference rFFT grid and resized to match the current spectral resolution. In practice, we bilinearly interpolate the real and imaginary parts of $\mathbf{G}_n$ to the current $\big(H, \lfloor \frac{W}{2} \rfloor + 1\big)$ grid. For each input, we predict mixture weights $\boldsymbol{\alpha} \in \mathbb{R}^{B\times N}$ using global average pooling (GAP) followed by an MLP:

$$\boldsymbol{\alpha} = \mathrm{softmax}\big(\mathrm{MLP}(\mathrm{GAP}(\mathbf{t}^{(i)}))\big). \tag{7}$$

The filtered spectrum is computed as:

$$\tilde{\mathbf{T}}_b^{(i)} = \sum_{n=1}^{N} \alpha_{b,n}\ \big(\mathbf{G}_n \odot \mathcal{F}(\mathbf{t}_b^{(i)})\big), \quad b = 1, \ldots, B, \tag{8}$$

where $\odot$ denotes element-wise complex multiplication (broadcast over channels). Finally, the global context is obtained via inverse rFFT and a channel mixing layer:

$$\Phi_g^{(i)} = \mathrm{Conv}_{1\times1}\big(\mathcal{F}^{-1}(\tilde{\mathbf{T}}^{(i)})\big). \tag{9}$$

The global nature of SGC is established by the convolution theorem: $\mathcal{F}^{-1}(\mathbf{G} \odot \mathcal{F}(\mathbf{t}))$ is mathematically equivalent to the circular convolution of $\mathbf{t}$ with a kernel whose Fourier transform is $\mathbf{G}$. Consequently, SGC implements an image-wide linear operator with a global receptive field, while the dynamic mixture in Eq. (8) ensures content adaptivity. Crucially, SGC relies exclusively on causal side information $\mathbf{t}^{(i)}$ (Eq. (3)), guaranteeing validity during decoding. Implementation parameterizes the real and imaginary parts of $\mathbf{G}_n$ on the rFFT grid; the imaginary components at self-conjugate bins (e.g., the DC bin, and the Nyquist bin when applicable) are fixed to zero to satisfy the rFFT Hermitian constraint.

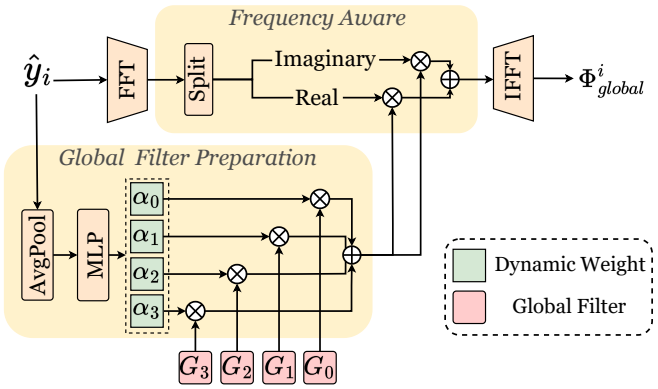

Fig. 2. Spectral Global Context (SGC). A small set of learnable spectral filters are dynamically mixed to modulate the frequency response of causal side information.

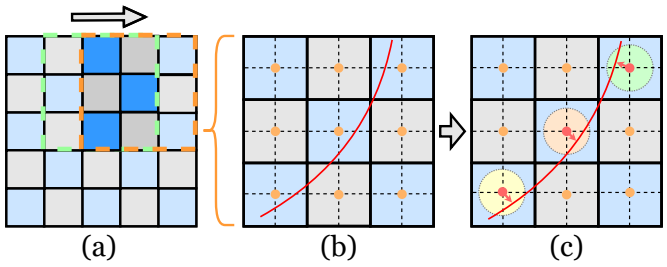

Fig. 3. Offset-adapted local context. Overlapped windows enable cross-window information flow, while predicted offsets enhance alignment for structures ill-suited to fixed grids.

D. Offset-Adapted Local Context (OLC)

While SGC provides global priors, non-anchor symbols necessitate precise local correspondence. OLC predicts local context for $\hat{\mathbf{y}}_{\bar{A}}^{(i)}$ using decoded anchors $\hat{\mathbf{y}}_A^{(i)}$.

1) Overlapped Window Attention: We partition the feature map into $K\times K$ windows with stride $K/2$ (50% overlap). Queries are derived from side information $\mathbf{t}^{(i)}$ at non-anchor positions, while keys/values are derived from embeddings of decoded anchors:

$$\mathbf{Q} = W_q\, \mathbf{t}_{\bar{A}}^{(i)}, \quad \mathbf{K} = W_k\, \eta(\hat{\mathbf{y}}_A^{(i)}), \quad \mathbf{V} = W_v\, \eta(\hat{\mathbf{y}}_A^{(i)}), \tag{10}$$

where $\eta(\cdot)$ is an embedding layer. Within each window, masked multi-head attention [13] generates intermediate local features:

$$\mathbf{A}^{(i)} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{B} + \mathbf{M}\right)\mathbf{V}, \tag{11}$$

where $\mathbf{B}$ is a learnable relative position bias and $\mathbf{M}$ masks invalid (non-causal) connections, ensuring non-anchor queries attend solely to anchor keys. All operations are conditional on decoded anchors, preventing access to future symbols.

2) Offset-Adapted Sampling: Inspired by deformable convolution [14], we predict a 2D offset field $\Delta\mathbf{p}^{(i)}$ from the attention features $\mathbf{A}^{(i)}$ using a lightweight CNN $\xi(\cdot)$:

$$\hat{\mathcal{G}}^{(i)} = \mathcal{G} + \Delta\mathbf{p}^{(i)}, \quad \Delta\mathbf{p}^{(i)} = \mathrm{clip}\big(\xi(\mathbf{A}^{(i)}), -\delta, \delta\big), \tag{12}$$

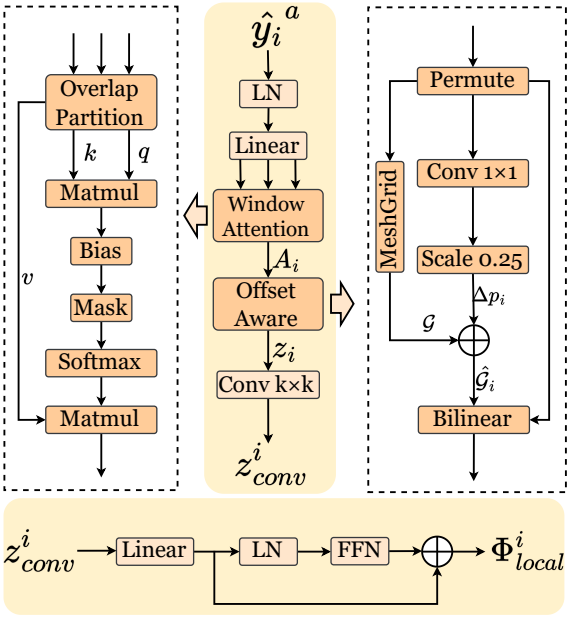

Fig. 4. OLC block structure employed for non-anchor context prediction.

where $\mathcal{G}$ represents the regular normalized sampling grid and δ=0.25 bounds the offsets for stability. We then perform bilinear sampling:

$$\mathbf{Z}^{(i)} = \mathcal{T}(\mathbf{A}^{(i)}, \hat{\mathcal{G}}^{(i)}), \tag{13}$$

followed by local aggregation utilizing a $K\times K$ convolution and a feed-forward network (FFN):

$$\Phi_l^{(i)} = \mathbf{Z}_{conv}^{(i)} + \mathrm{FFN}(\mathbf{Z}_{conv}^{(i)}), \quad \mathbf{Z}_{conv}^{(i)} = \mathrm{Conv}_{K\times K}(\mathbf{Z}^{(i)}). \tag{14}$$

E. Complexity Analysis

Let n=HW denote the number of latent positions per slice. Standard global self-attention scales as $O(n^2)$, which dominates computational costs at high resolutions. In contrast, SGC requires one FFT and one inverse FFT per slice, scaling as $O(n \log n)$, plus $O(Nn)$ for spectral modulation (where N is small, e.g., N=4). OLC utilizes window attention with a constant window size, resulting in $O(nK^2)$ complexity. Thus, DAEM scales near-linearly with resolution, making it highly effective for high-resolution entropy modeling.

F. Rate–Distortion Objective

We optimize the standard rate–distortion loss:

$$\mathcal{L} = \lambda\, \mathcal{D}(\mathbf{x}, \hat{\mathbf{x}}) + \mathbb{E}\big[-\log_2 p(\hat{\mathbf{y}}|\hat{\mathbf{z}})\big] + \mathbb{E}\big[-\log_2 p(\hat{\mathbf{z}})\big], \tag{15}$$

where $\mathcal{D}$ is either MSE or MS-SSIM [15], and the bitrate is computed utilizing the learned likelihoods.

IV. Experiments

A. Experimental Setup

Our implementation is built upon PyTorch [16] and CompressAI [17]. We train all models on the Vimeo-90K dataset [18]. For evaluation, we primarily report results on the Kodak dataset [19], and further assess generalization on two high-resolution benchmarks: the Tecnick dataset [20] and the CLIC Professional Validation set [21]. DAEM employs S=10 slices, a window size of K=8 with 50% overlap, and N=4 spectral filters.

Optimization is performed using Adam with an initial learning rate of 10^{-4}, β_1=0.9, and β_2=0.999, over 2M training steps. Following standard LIC protocols, we train 10 models in total: five MSE-optimized models with $\lambda \in \{0.0035, 0.0067, 0.013, 0.025, 0.0483\}$, and five MS-SSIM-optimized models with $\lambda \in \{4.58, 8.73, 16.64, 31.73, 60.50\}$.

Metrics. We report PSNR and MS-SSIM as distortion metrics. For rate–distortion (RD) curves and Bjøntegaard-Delta (BD) computations, MS-SSIM is converted to decibels as $-10\log_{10}(1 - \text{MS-SSIM})$. BD-rate and BD-PSNR are computed following the Bjøntegaard-Delta method [22].

TABLE I
BD-rate and BD-PSNR (BD-MSSSIM) performance on Kodak compared to BPG.

Method	PSNR		MS-SSIM	
	BD-rate	BD-PSNR	BD-rate	BD-MSSSIM
BPG [2]	0.00%	0.00dB	0.00%	0.00dB
VTM-17.0 [3]	-22.46%	+1.26dB	-25.91%	+1.38dB
Cheng20 [23]	-18.34%	+0.99dB	-57.98%	+4.33dB
Entroformer [10]	-19.89%	+1.11dB	-59.80%	+4.41dB
STF [24]	-23.72%	+1.31dB	-62.64%	+4.61dB
MLIC [9]	*-29.08%*	*+1.59dB*	*-63.19%*	*+4.69dB*
DAEM (Ours)	**-32.06%**	**+1.88dB**	**-65.45%**	**+4.74dB**

TABLE II
Encoding and decoding latency on Kodak (ms per image). All neural methods are measured on the same GPU. VTM is reported for reference.

Method	Encoding (ms)	Decoding (ms)
VTM-17.0	104921.8	235.4
Cheng20	3708.2	8658.6
STF	259.4	*162.9*
MLIC	*220.2*	169.9
DAEM (Ours)	**181.5**	**136.4**

B. Comparison with State-of-the-Art

We compare DAEM against classical codecs, including BPG [2] and VTM-17.0 [3], as well as representative learned image compression (LIC) methods: Cheng20 [23], Entroformer [10], STF [24], MLIC [9], and WeConvene [12].

As illustrated in Fig. 5 and Table I, DAEM consistently outperforms all compared learned codecs across the 0.1–0.8 bpp range. Using BPG as the reference in Table I, DAEM achieves -32.06% BD-rate and +1.88 dB BD-PSNR on Kodak. Under the same reference, this corresponds to a further 9.60 percentage-point BD-rate reduction and a 0.62 dB BD-PSNR gain over VTM-17.0. Notably, the performance gains are more

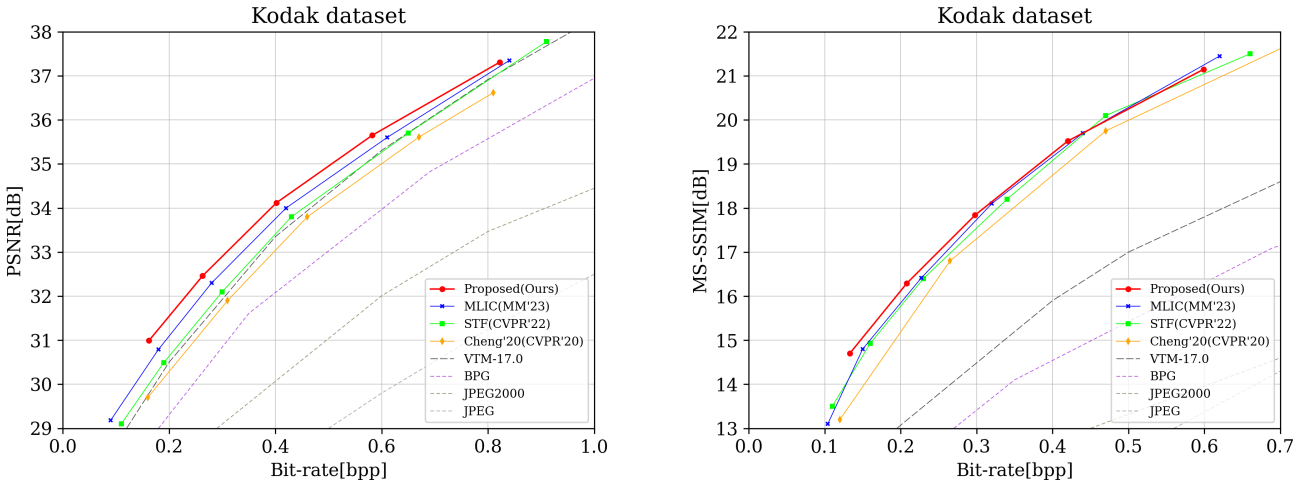

Fig. 5. Rate–distortion curves on Kodak. Left: PSNR vs. bitrate. Right: MS-SSIM (dB) vs. bitrate.

TABLE III
BD-RATE (%) RELATIVE TO VTM-17.0 ON HIGH-RESOLUTION BENCHMARKS. LOWER IS BETTER.

Method	Tecnick		CLIC Pro. Val.	
	PSNR	MS-SSIM	PSNR	MS-SSIM
VTM-17.0	0.00	0.00	0.00	0.00
STF [24]	-2.75	-45.81	+0.42	-44.82
MLIC [9]	**-12.73**	*-47.26*	-8.79	**-45.79**
WeConvene [12]	-9.46	-45.57	*-10.15*	*-45.41*
DAEM (Ours)	*-12.67*	**-47.55**	**-10.28**	-45.21

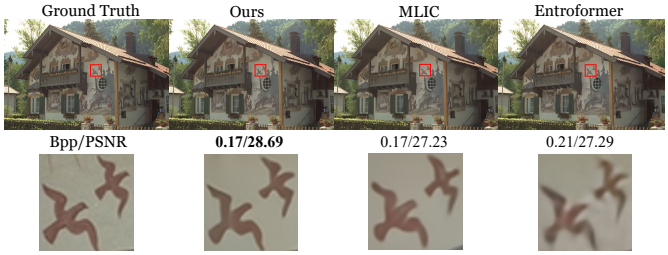

Fig. 6. Visual comparison on Kodak (Kodim24). DAEM reduces ringing artifacts and preserves fine structures more effectively than Entroformer and MLIC.

pronounced on the MS-SSIM metric, where DAEM attains a -65.45% BD-rate relative to BPG, exceeding MLIC by 2.26 percentage points. This improvement can be attributed to the proposed spectral global context, which effectively captures perceptually relevant long-range dependencies.

Regarding computational efficiency, Table II demonstrates that DAEM is $1.21\times$ faster in encoding and $1.25\times$ faster in decoding than MLIC, while simultaneously achieving superior RD performance. These results validate that spectral-domain global modeling successfully circumvents the latency bottlenecks inherent in attention-based global context mechanisms. Compared to Cheng20, which relies on autoregressive decoding, DAEM reduces encoding and decoding latency by over an order of magnitude, underscoring the practical advantages of parallel entropy coding.

C. Generalization to High-Resolution Content

To assess the scalability of DAEM on larger spatial resolutions, we evaluate its performance on the Tecnick and CLIC Professional Validation datasets. As reported in Table III, DAEM maintains robust coding efficiency on high-resolution content, achieving BD-rate reductions of 12.67% and 10.28% (PSNR) relative to VTM-17.0 on Tecnick and CLIC, respectively.

Notably, DAEM consistently surpasses STF by a significant margin and remains highly competitive with state-of-the-art methods such as MLIC and WeConvene. On the Tecnick dataset, DAEM achieves superior MS-SSIM performance (-47.55%) compared to all competing methods. On CLIC Professional Validation, DAEM attains the best PSNR-based BD-rate (-10.28%), outperforming both MLIC and WeConvene. These results confirm that the proposed spectral filtering mechanism effectively captures long-range spatial correlations in high-resolution imagery without suffering from the receptive field limitations of purely local models or the computational overhead of dense attention mechanisms. However, we observe that DAEM yields a slightly inferior MS-SSIM BD-rate on CLIC Professional Validation (-45.21%) compared to MLIC (-45.79%), suggesting that content with diverse scene categories may benefit from more adaptive global modeling, which we leave for future exploration.

D. Ablation Study

We conduct ablation studies to isolate the individual contributions of DAEM's key components. The baseline model employs a hyperprior [5] combined with a checkerboard

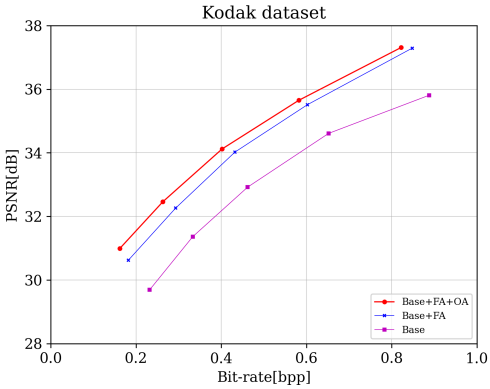

Fig. 7. Ablation study on Kodak (PSNR). Base configuration: hyperprior with checkerboard local context.

local context model, following prior parallel entropy coding designs [7].

As shown in Fig. 7, incorporating the Spectral Global Context (SGC) module (Base+SGC) yields significant RD gains across all bitrates, demonstrating that efficient global priors remain essential even within parallel coding frameworks. The addition of the Offset-adapted Local Context (OLC) module (Base+SGC+OLC) provides further consistent improvements, particularly at lower bitrates where accurate entropy estimation is critical. Qualitatively, OLC enhances reconstruction quality around edges and thin structures, confirming the efficacy of offset-adapted alignment for capturing local dependencies in regions with geometric discontinuities.

V. Conclusion

In this paper, we presented DAEM, a dual-aware entropy model designed for scalable learned image compression. DAEM substitutes quadratic global attention with spectral global context derived from causal side information, achieving a global receptive field with $O(n \log n)$ complexity. Additionally, DAEM enhances parallel local context via offset-adapted alignment over overlapped windows, improving structural correspondence for non-anchor decoding. Experimental results on Kodak, Tecnick, and CLIC Professional Validation demonstrate that DAEM achieves superior rate–distortion performance with significantly reduced latency, confirming its scalability to high-resolution content. We anticipate that spectral and offset-aware context modeling will constitute a pivotal direction for the development of high-resolution and latency-sensitive learned codecs.

References

[1] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Proceedings DCC 2000. Data compression conference*. IEEE, 2000, pp. 523–541.

[2] F. Bellard, "BPG Image format," 2015. [Online]. Available: https://bellard.org/bpg/

[3] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the Versatile Video Coding (VVC) Standard and its Applications," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 3736–3764, Oct. 2021.

[4] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end Optimized Image Compression," in *International Conference on Learning Representations*, 2016.

[5] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

[6] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, 2018.

[7] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard Context Model for Efficient Learned Image Compression," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021.

[8] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.

[9] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "MLIC: Multi-Reference Entropy Model for Learned Image Compression," in *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa ON Canada: ACM, Oct. 2023, pp. 7618–7627.

[10] Y. Qian, X. Sun, M. Lin, Z. Tan, and R. Jin, "Entroformer: A Transformer-based Entropy Model for Learned Image Compression," in *International Conference on Learning Representations*, 2021.

[11] H. Li, S. Li, W. Dai, C. Li, J. Zou, and H. Xiong, "Frequency-Aware Transformer for Learned Image Compression," Dec. 2024, arXiv:2310.16387 [eess].

[12] H. Fu, J. Liang, Z. Fang, J. Han, F. Liang, and G. Zhang, "WeConvene: Learned Image Compression with Wavelet-Domain Convolution and Entropy Model," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, vol. 15108, pp. 37–53, series Title: Lecture Notes in Computer Science.

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

[14] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[15] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[17] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "CompressAI: a PyTorch library and evaluation platform for end-to-end compression research," Nov. 2020, arXiv:2011.03029 [cs, eess].

[18] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video Enhancement with Task-Oriented Flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[19] E. Kodak, "Kodak lossless true color image suite," 1993. [Online]. Available: https://r0k.us/graphics/kodak/

[20] N. Asuni and A. Giachetti, "TESTIMAGES: a Large-scale Archive for Testing Visual Devices and Basic Image Processing Algorithms." in *STAG*, 2014, pp. 63–70.

[21] G. Toderici, L. Theis, N. Johnston, E. Agustsson, F. Mentzer, J. Ballé, W. Shi, and R. Timofte, "Clic 2020: Challenge on learned image compression," *Retrieved March*, vol. 29, p. 2021, 2020.

[22] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," ITU-T Video Coding Experts Group (VCEG), ITU-T SG16/Q6, Austin, TX, USA, Tech. Rep. VCEG-M33, Apr. 2001. [Online]. Available: https://www.itu.int/wftp3/av-arch/video-site/0104_Aus/VCEG-M33.doc

[23] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.

[24] R. Zou, C. Song, and Z. Zhang, "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*, 2022, pp. 17 492–17 501.