Adaptive Learned Image Compression with Graph Neural Networks

Supplementary Material

1. Advanced CNNs and Transformers

We provide additional quantitative comparisons against advanced CNN- and Transformer-based operators. Specifically, we replace our GNN module with either deformable/dynamic convolutions [3, 4] or Swin/Neighborhood Attention (NAT) [7, 13]. We keep the entropy model and FFN unchanged, and only swap the graph operator with these alternative operators. All models are trained with $\lambda = 0.05$, yielding the loss curves in Fig. 1 and the R–D points in Fig. 2.

At $\lambda = 0.05$, GLIC reduces bitrate by more than 9% compared with deformable/dynamic convolutional variants, and by more than 5% compared with Swin/NAT variants, confirming the advantage of our GNN-based design.

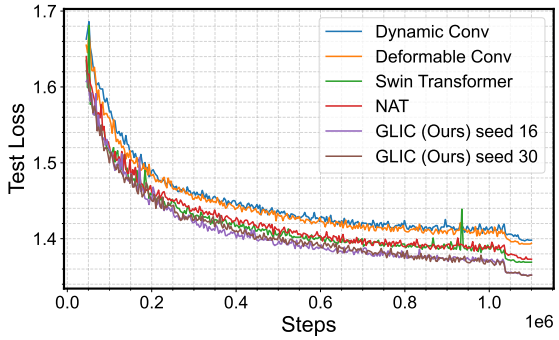


Figure 1. Test Loss on Kodak vs. Training Step

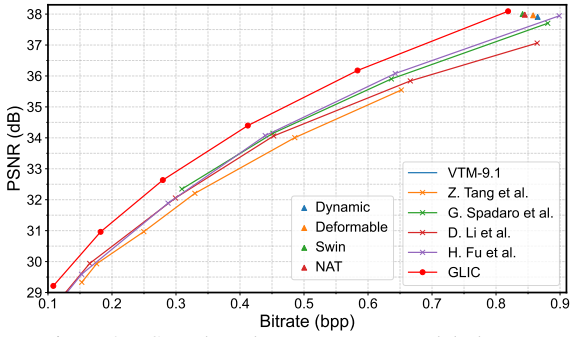


Figure 2. PSNR-based RD curves on Kodak dataset.

2. Training Stability

Our GLIC model can be trained stably under different random seeds. The test loss curves of GLIC with multiple seeds (Fig. 1) consistently indicate stable optimization. Compared with other operators and baselines, GLIC exhibits smooth convergence and consistently outperforms the alternatives when using identical entropy models.

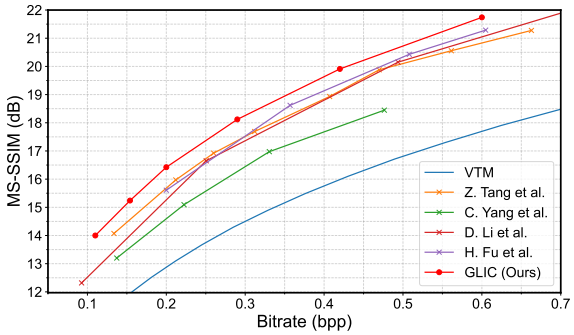


Figure 3. MS-SSIM-based RD curves on Kodak dataset.

3. Comparisons with GNN-based LIC models

We provide detailed PSNR and MS-SSIM R–D curves in Fig. 2 and Fig. 3. We compare against three graph-based LIC models, Z. Tang *et al.* [16], G. Spadaro *et al.* [15], and C. Yang *et al.* [17], as well as two deformable CNN-based LIC models, D. Li [9] and H. Fu *et al.* [6]. Our GLIC model significantly outperforms all of them, demonstrating the effectiveness of our GNN design.

Our GLIC differs fundamentally from existing methods (Tab. 1). These differences constitute the key innovations of our approach:

1. Flexible Receptive Fields via Dual-scale Sampling. Fully connected or global KNN graphs in [16, 17] provide global context but incur $\mathcal{O}(N^2)$ complexity, while the linear-complexity window graph in [15] restricts the receptive field to a local patch. Our dual-scale graph combines both types of graphs, achieving a near-global receptive field with linear complexity.

2. Complexity-aware Neighbor Quota Assignment. Instead of fixing the neighbor count to a constant K, we adaptively allocate node degrees according to the RMS gradient magnitude, assigning denser connections to visually complex regions.

4. Details of ERF Computation

The effective receptive field (ERF) [14] characterizes how strongly each input pixel *actually* affects a designated output unit, complementing the theoretical receptive field implied by the architecture. Given an input image $\mathbf{x}_0 \in \mathbb{R}^{3\times H\times W}$ and a network output (or intermediate feature) $\mathbf{y}$, we denote by $x_{0,i,j}$ the input pixel at spatial location (i, j), and by $y_{c,h,w}$ the feature at channel c and spatial location (h, w).

Following our implementation, we compute the ERF on the analysis transform $g_a(\cdot)$. For an input image, we first

Table 1. Comparison with other GNN-based models. N: total pixels, M: pixels per window [2], K: the GNN neighborhood size, n: sample size for our graphs. Complexity: *Construction* + *Aggregation* Complexity.

	[16]	[17]	[15]	**GLIC (ours)**
Construction	Fully-connected graph	Global K-NN graph	Local window K-NN	Dual graph
Node Degree	Fixed $N-1$ (full)	Fixed K	Fixed K	Adaptive, average n
Aggregation	Graph Attention	ECC	Graph Attention	Edge-conditioned Aggregation
Receptive field	Full	Full	Bounded by Window	Approximately Full
Complexity	$O(1)+O(N^2)$	$O(N^2)+O(NK)$	$O(NM)+O(NK)$	$O(Nn)+O(Nn)$

obtain the latent feature

$$\mathbf{y} = g_a(\mathbf{x}_0), \tag{1}$$

and select its spatial center (h_c, w_c). To measure the influence on the center latent unit, we form a scalar response by summing all channels at that position:

$$s = \sum_c y_{c,h_c,w_c}. \tag{2}$$

The ERF map is then defined as the magnitude of the gradient of s with respect to the input:

$$\mathrm{ERF}(i,j) = \left| \frac{\partial s}{\partial x_{0,i,j}} \right|. \tag{3}$$

Intuitively, a larger $\mathrm{ERF}(i,j)$ indicates that the input pixel $x_{0,i,j}$ contributes more strongly to the center latent response.

In practice, we back-propagate from s to the input and post-process the resulting gradients: (i) clamp large values by an upper bound τ (we use τ=0.2); (ii) apply ReLU to retain only positive contributions; and (iii) aggregate over channels to obtain a 2D spatial map:

$$\tilde{\mathbf{E}}(i,j) = \sum_c \mathrm{ReLU}\left(\min\left(\frac{\partial s}{\partial x_{0,c,i,j}}, \tau \right) \right). \tag{4}$$

5. VTM Setting and Commands

We follow [5] to ensure a widely used baseline. Specifically, we directly adopt the R–D results reported in the CompressAI repository [1], which are obtained using VTM-9.1. For our VTM-9.1 evaluations, we use the QP set $\{22, 27, 32, 37, 42, 47\}$, with the following command:

```
VTM/bin/EncoderAppStatic -i [input.yuv] \
  -c VTM/cfg/encoder_intra_vtm.cfg \
  -o [output.yuv] -b [output.bin] \
  -wdt [width] -hgt [height] -q [QP] \
  --InputBitDepth=8 -fr 1 -f 1 \
  --InputChromaFormat=444
```

Table 2. BD-rate (%) comparison between the GNN-based SCCTX entropy model (GNN-SCCTX) and the convolution-based SCCTX entropy model in GLIC. Lower is better.

Method	Kodak(%)	Tecnick(%)	CLIC(%)	Dec.Lat(s)
GNN-SCCTX	-19.08	-22.86	-19.78	0.506
Conv-SCCTX (Ours)	-19.29	-21.69	-18.71	0.395

6. GNNs in the Entropy Model

Our experiments so far demonstrate that the proposed GNN is highly effective for nonlinear transform networks. In the main GLIC model, we follow [8] and adopt convolutional networks and MLPs for spatial-channel entropy modeling. Here, we further explore the potential of GNNs for improving the entropy model itself.

As shown in Tab. 2, we construct a variant (GNN-SCCTX) that uses our GNNs to model the spatial and channel context in the SCCTX entropy model, and compare it with the convolution-based SCCTX (Conv-SCCTX). We observe that the GNN-SCCTX variant achieves better R–D performance on the high-resolution datasets CLIC and Tecnick by more than 1.2%. However, this comes at the cost of a 28% increase in decoding latency. We consider this level of BD-rate gain insufficient to justify such a substantial slowdown in decoding.

Nevertheless, these results indicate that the entropy model can be further enhanced by GNNs or other advanced architectures, suggesting that the overall performance of GLIC could be further boosted when combined with a stronger entropy model.

7. Ablations on GNN Settings

We further ablate the GNN design by varying the *average connection number* and the *candidate sampling size*, as summarized in Tab. 3. Increasing these two hyperparameters to "64 / 16×16" substantially improves the R–D performance, while only mildly increasing the decoding latency.

We summarize the observation as follows: **(1) Sparsity.** Increasing the candidate set size does not yield unbounded RD gains. We find that 16×16 sampling, though sparse, is already near saturation. A denser setting (e.g., 32×32)

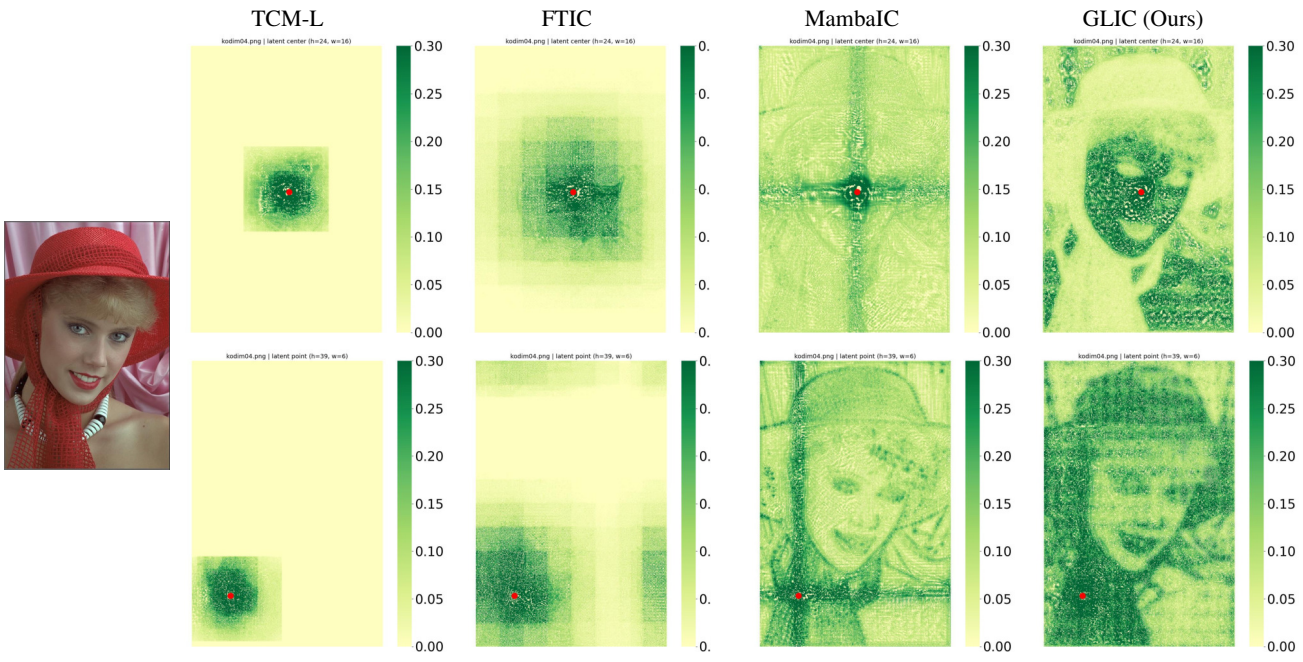


Figure 4. Additional single-image ERF visualizations (set 1).

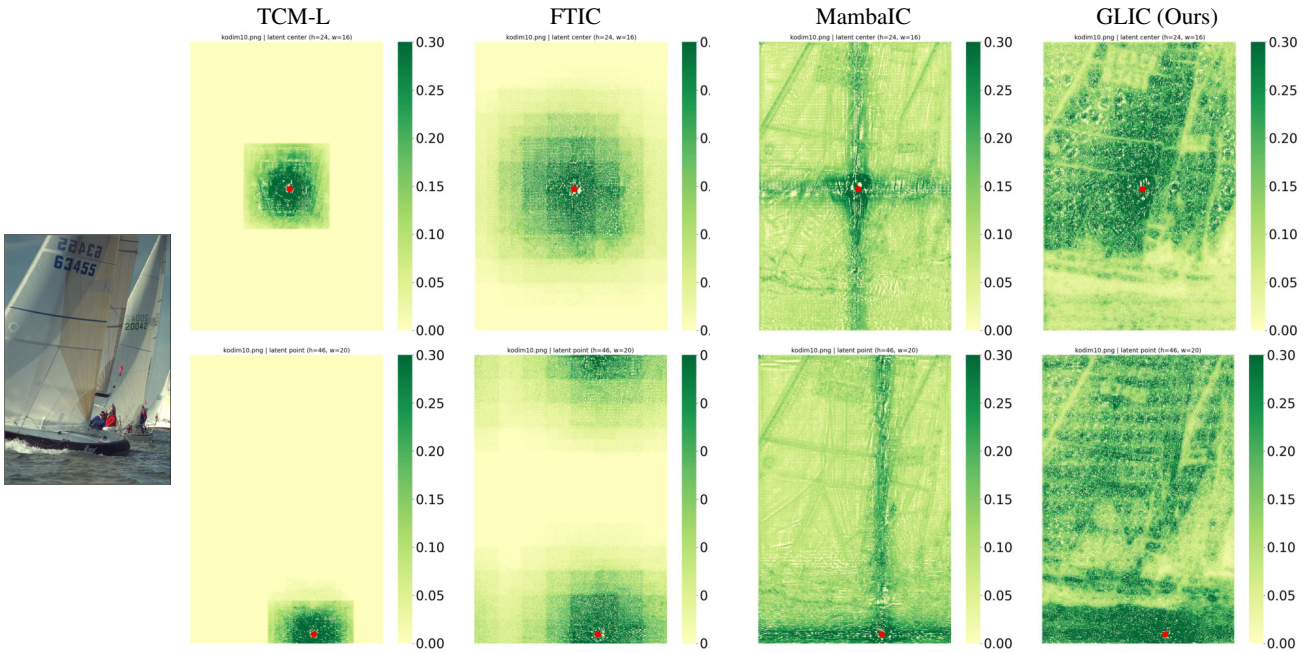


Figure 5. Additional single-image ERF visualizations (set 2).

yields no significant gain while incurring much higher latency. **(2) Budget.** Varying the connection budget (average degree) shows that 64 neighbors are sufficient. Expanding to a fully connected graph (256 connections) does not further improve RD performance as it may result in unnecessary interactions for redundancy elimination.

8. Clarification on Table 3 in the Main Paper

In Table 3 ("Ablations on dual-scale graph flexibility") of the main paper, the entries "w/o Global (Local only)" and "w/o Local (Global only)" indicate that one of the two graphs is replaced by the other one. In other words, each variant replaces the dual-scale design with a single-scale design, but does not change the network depth.

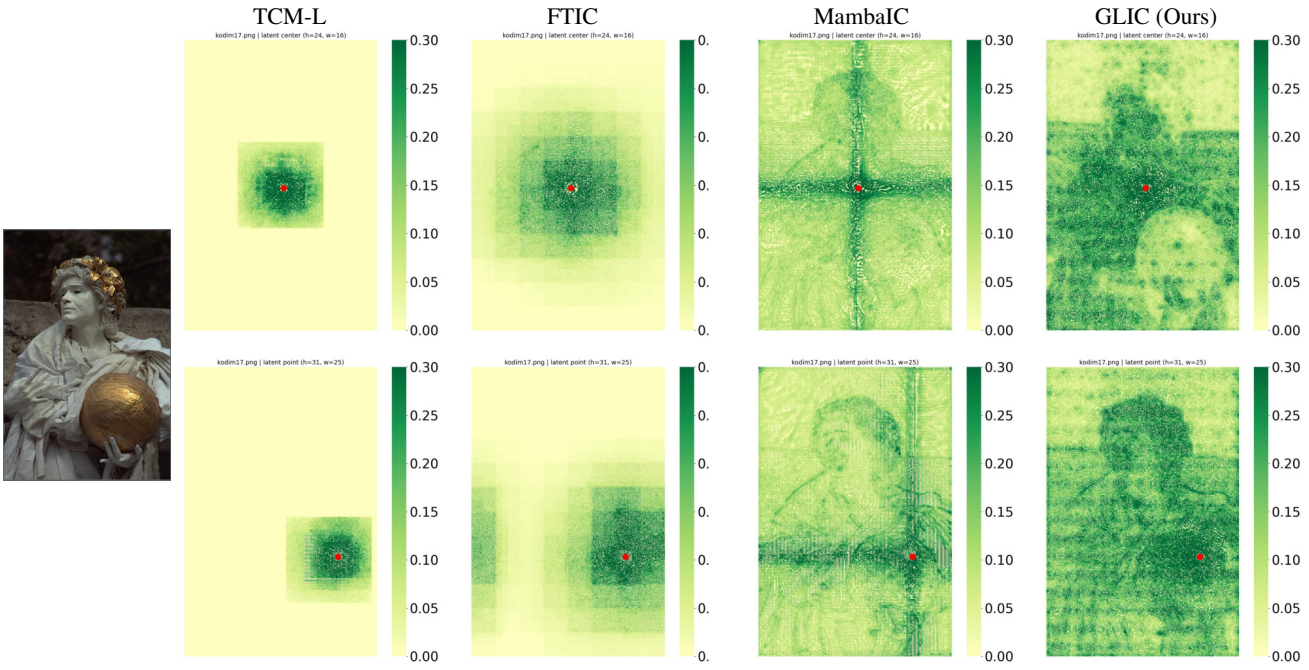


Figure 6. Additional single-image ERF visualizations (set 3).

Table 3. Ablation on average connection number and candidate sampling size. BD-rate (%) is computed on Kodak; lower is better.

Average Connection / Candidate Size	Kodak(%)	Dec.Lat(s)
32 / 8×8	-15.33	0.298
64 / 16×16 (Ours)	-19.29	0.395
256 / 16×16	-19.17	0.402
512 / 32×32	-19.97	0.891

Table 4. Comparison between standard and accelerated training under different λ values on Kodak (lower test loss is better).

	$\lambda = 0.05$	$\lambda = 0.0017$
Standard Training	1.344	0.258
Accelerated Training	1.343	0.258

9. Ablations on Training Strategy

We also ablate the accelerated training strategy proposed in [11] at two operating points, $\lambda = 0.05$ and $\lambda = 0.0017$. We compare the best test losses of GLIC obtained with this accelerated strategy against those obtained with the standard 2M-step training schedule used in [8, 10, 12]. The results on Kodak are reported in Tab. 4.

The test losses under the two training regimes are almost identical, and the remaining gaps are very small. Therefore, we conclude that adopting the accelerated training strategy does not noticeably affect the performance of our GLIC model.

10. More Adaptivity Visualizations

We provide additional single-image ERF visualizations to further illustrate the adaptivity of our method. As shown in Fig. 4–6, GLIC produces ERFs that are both more global and more content-adaptive than those of prior approaches. The activated ERF patterns vary noticeably across different images and also across different target positions within the same image.

In contrast, previous Transformer- and Mamba-based methods, including TCM [12], FTIC [10], and MambaIC [18], tend to yield highly similar ERF structures across images and target locations. Their high-response regions are largely isotropic and appear content-agnostic rather than truly content-dependent.

Our method, however, exhibits clear content-driven behavior: for example, in Fig. 4, it strongly attends to the face, the red hat and the scarf, respectively in the two examples; in Fig. 6, it focuses on the white sculpture and the golden sphere, respectively and separately for the two examples. Importantly, regions with large ERF values align with areas that exhibit substantial visual redundancy consistent with human perception and semantic structure, further validating the effectiveness of our approach.

11. More Visualization Comparisons

We also provide more subjective visual comparisons to demonstrate the effectiveness of our GLIC model. We compare GLIC with the traditional codec VTM-9.1 [2] and the LIC model TCM [12]. A detailed inspection of image

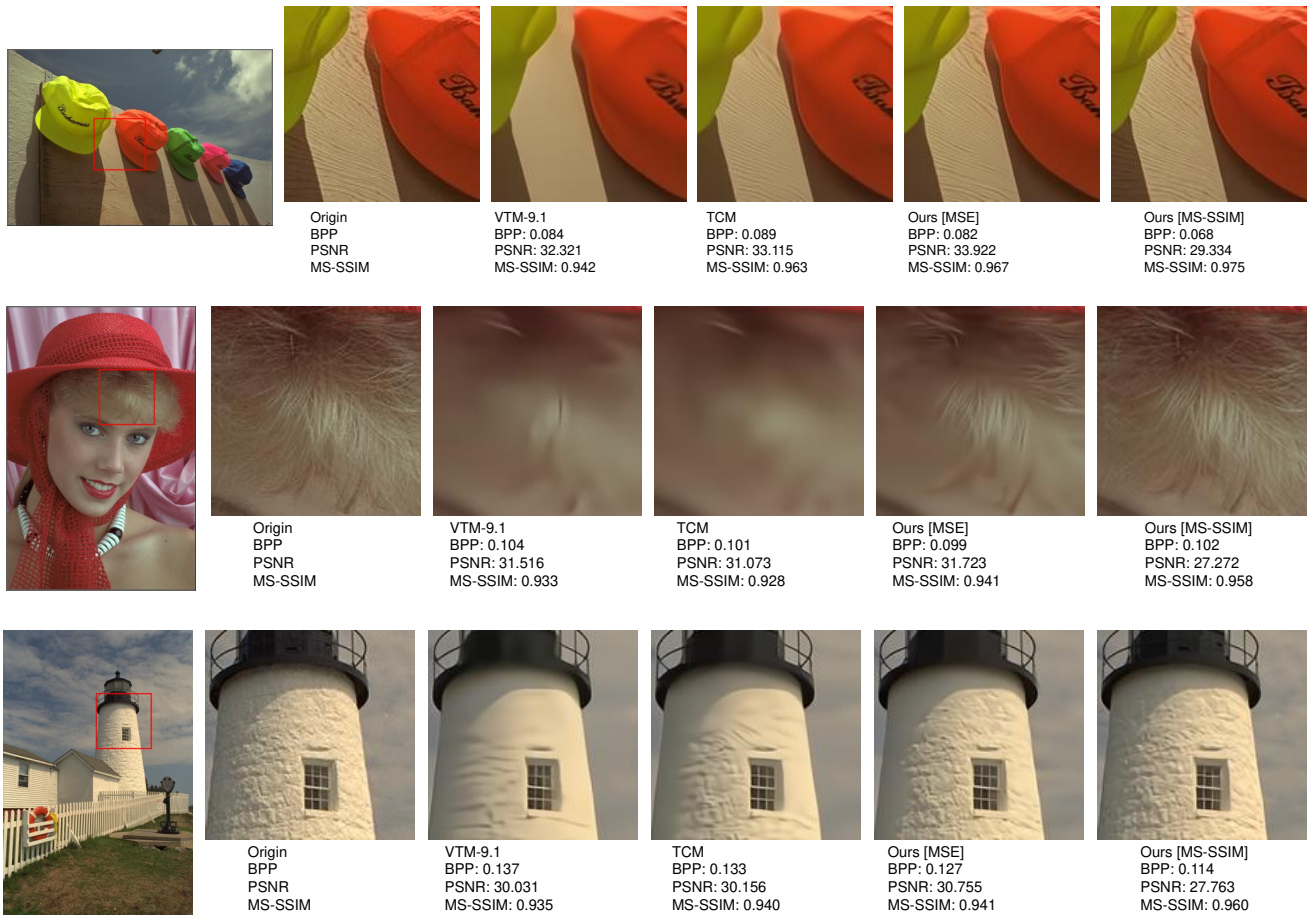
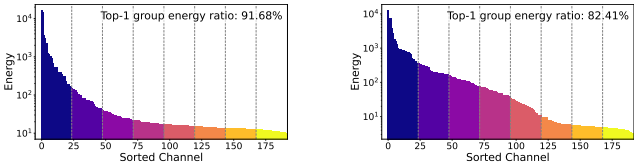


Figure 7. More Visual Comparisons.

patches reveals that GLIC better preserves high-frequency details and textures, effectively mitigating the severe detail loss and over-smoothing observed in other methods. This improvement stems from our complexity-aware neighbor quota assignment, which adaptively allocates more connections in complex regions, leading to better compression and reconstruction of intricate structures.

12. Why RMS-Gradient

As shown in Fig. 8, for LIC models, both the latent features and the Sobel-gradient responses exhibit strong energy compaction: a small subset of channels concentrates most of the energy (carrying the most critical information), while the majority of channels remain low-energy. Since RMS emphasizes larger magnitudes, it is better suited than the plain averaging. RMS highlights the high-energy, informative channels instead of being diluted by numerous low-energy ones. This is also empirically validated by our experimental results in main manuscript Tab. 4.



(a) Feature energy compaction (b) Gradient energy compaction

Figure 8. Channel-wise energy distribution of Stage-2 features, averaged over Kodak images. The channels are sorted into eight groups.

13. Limitations and future work

A key limitation is that we have not evaluated our dynamic GNNs' effectiveness in exploiting cross-frame redundancy for video compression. In future work, we will extend our GNN to video codecs and investigate how adaptive graphs can better model temporal correlations.

References

[1] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv e-prints*, pages arXiv–2011, 2020. 2

[2] Jianle Chen, Yan Ye, and S Kim. Algorithm description for versatile video coding and test model 1 (vtm 1). *Joint Video Experts Team (JVET) of ITu-T SG*, 16:3–12, 2020. 4

[3] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. 1

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1

[5] Donghui Feng, Zhengxue Cheng, Shen Wang, Ronghua Wu, Hongwei Hu, Guo Lu, and Li Song. Linear attention modeling for learned image compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7623–7632, 2025. 2

[6] Haisheng Fu, Feng Liang, Jie Liang, Yongqiang Wang, Zhenman Fang, Guohe Zhang, and Jingning Han. Fast and high-performance learned image compression with improved checkerboard context model, deformable residual module, and knowledge distillation. *IEEE Transactions on Image Processing*, 2024. 1

[7] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6185–6194, 2023. 1

[8] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. 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*, pages 5718–5727, 2022. 2, 4

[9] Daowen Li, Yingming Li, Heming Sun, and Lu Yu. Deep image compression based on multi-scale deformable convolution. *Journal of Visual Communication and Image Representation*, 87:103573, 2022. 1

[10] Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware transformer for learned image compression. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[11] Han Li, Shaohui Li, Wenrui Dai, Maida Cao, Nuowen Kan, Chenglin Li, Junni Zou, and Hongkai Xiong. On disentangled training for nonlinear transform in learned image compression. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[12] Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14388–14397, 2023. 4

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[14] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 1

[15] Gabriele Spadaro, Alberto Presta, Enzo Tartaglione, Jhony H Giraldo, Marco Grangetto, and Attilio Fiandrotti. Gabic: Graph-based attention block for image compression. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1802–1808. IEEE, 2024. 1, 2

[16] Zhisen Tang, Hanli Wang, Xiaokai Yi, Yun Zhang, Sam Kwong, and C-C Jay Kuo. Joint graph attention and asymmetric convolutional neural network for deep image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):421–433, 2022. 1, 2

[17] Chunhui Yang, Yi Ma, Jiayu Yang, Shiyi Liu, and Ronggang Wang. Graph-convolution network for image compression. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2094–2098. IEEE, 2021. 1, 2

[18] Fanhu Zeng, Hao Tang, Yihua Shao, Siyu Chen, Ling Shao, and Yan Wang. Mambaic: State space models for high-performance learned image compression. *arXiv preprint arXiv:2503.12461*, 2025. 4