Enhanced Compression Model: A Hyperprior-Based Neural Image Compression Framework with Context Modeling, Entropy Parameter Network, Semantic Gating, and Scale Refinement

1st Abhinand Meethele Valappil
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Modinagar, India
mvabhinand2005@gmail.com

2th Sanidhya Singh
Department of IoT and Intelligent Systems
Manipal University Jaipur
Jaipur, India
sanidhya11719@gmail.com

3nd Rishabh Gupta
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Modinagar, India
rishabh886068@gmail.com

4rd Garv Randhar
Department of Computer Science and Engineering
SRM Institute of Science and Technology
Modinagar, India
garvrandhar18@gmail.com

5th Bhawna Upadhayay
Assistant Professor
SRM Institute of Science and Technology
Modinagar, India
bhawnaupadhayay123@gmail.com

***Abstract*—Over the years, various techniques have been developed for image compression. For a long time, the old standards were the two major players in the industry. The newly proposed learning-based image compression methods have been revealed to be very effective Yet to date, there remains a performance gap on compression efficiency, reconstruction quality, and visual integrity. In this paper, we develop the Enhanced Compression Model (ECM), an architecture for neural image compression which is based on a variational auto-encoder with a hyperprior. ECM improves the rate-distortion trade-off via semantic gating, channel attention, scale refinement, and autoregressive context modelling. These elements not only support the feature extraction but also the entropy modelling, resulting in a better estimation of the latent distribution and a higher level of quality of the picture reconstruction. Trained on DIV2K with a rate-distortion objective, ECM joins hyperprior entropy coding with semantic feature handling to compress in a highly efficient manner while still preserving the fundamental picture structures. The experimental results on the Kodak dataset demonstrate that ECM attained 30.882 dB PSNR and 0.985 MS-SSIM at 0.176 BPP, which means the proposed method is quite effective and opens the way for advanced entropy modelling and semantic feature gating in neural image compression.**

***Index Terms*—Neural Image Compression, Hyperprior Model, Entropy Modeling, Context Modeling, Semantic Gating, Channel Attention, Rate-Distortion Optimization, Deep Learning**

I. Introduction

Compressing images is key in making digital communication efficient in today's visual-oriented world. Without it, things like HD video streaming and huge cloud storage would be impossible. Several years ago, JPEG and other conventional compression algorithms were the main solutions. But, they have a hard time handling very low bit rates [1]–[3]. Because of this shortcoming, the new learned image compression that relies on autoencoder-based neural networks that are able to learn compression from the data was introduced [4], [5]. We take our inspiration from the scale hyperprior architecture, which makes use of side information to model spatial dependencies [9]. Standard hyperprior models have difficulty capturing fine-grained spatial dependencies and assigning greater weight to semantically significant features. The **Enhanced Compression Model (ECM)** is proposed to address this issue. ECM integrates several complementary components: semantic gating to dynamically modulate feature channels, channel attention to amplify important representations [13], [14], autoregressive context modeling for accurate probability estimation, and refined entropy parameter networks. These components contribute to enhanced compression efficiency and improved reconstruction quality. The model is trained via patch-based optimization on the high-resolution DIV2K dataset [18] and evaluated on the Kodak dataset [19] using BPP, PSNR, and MS-SSIM [16], [17].

Results show that ECM achieves better rate-distortion performance, particularly at very low bitrates, reaching **0.176 BPP** with **30.882 dB PSNR** and **0.985 MS-SSIM** —

5.5 dB PSNR higher than the Scale Hyperprior Q1 baseline (25.380 dB) at a lower bitrate (0.204 BPP), and **2.18 dB** higher than Scale Hyperprior Q3 (28.704 dB) at less than one-third the bitrate.

The paper's primary contributions are:

- **Novel integration of semantic gating with hyperprior-based entropy modelling for adaptive suppression of non-informative latent channels before entropy coding.**
- A **scale refinement** method that refines hyperprior Gaussian scale estimates using direct latent content to achieve an improved entropy accuracy for difficult textures.
- **Improved entropy parameter network** that conditions on hyperprior statistics and autoregressive context features concurrently for improved probability estimation.
- **Competitive performance at low bitrates**, attaining 0.176 BPP with 30.882 dB PSNR and 0.985 MS-SSIM on the Kodak dataset, outperforming classical and learned baselines at similar bitrates.

II. Background and Related Work

A. Classical image compression

For many years, classical image compression methods have been used to facilitate efficient digital image transmission. PNG is a standard lossless format that preserves original image quality and works best with synthetic content and text [2]. JPEG is a lossy transform codec that uses quantization, block-wise discrete cosine transform (DCT), and entropy coding to achieve significant compression of natural images [1]. WebP is a modern container format that supports both lossy and lossless modes, providing improved compression efficiency [3].

B. End-to-end learned compression and hyperpriors

Learned image compression discards manually designed coding pipelines by using trainable neural analysis and synthesis transforms [8]. The latent representation is quantized and entropy coded after an encoder maps an image to it, whereas a decoder is used for the image reconstruction. The entire system is trained end-to-end through a rate-distortion loss. Hierarchical hyperpriors in latent probability modeling and rate-distortion have led to substantial breakthroughs in the field, citeballe2018. To capture global latent statistics, Ball'e et al. [9] came up with the scale hyperprior. Minnen et al.

[10] further improved the results by autoregressively modeling the local dependencies. ECM is quite different from these approaches in that it uses semantic gating to deprioritize less informative latent channels in the first step of compression, it re-estimates hyperprior scales by using latent content, and it predicts entropy parameters jointly by hyperprior and context features resulting in a better compression method.

C. Autoregressive and attention-based entropy models

Side information can be used to boost compression efficiency by using hyperpriors to model spatial dependencies. However, they may still fail to completely capture complex interactions between latent coefficients in neighbouring places. Autoregressive context models utilise the previously decoded latents to drive the entropy estimation of the current symbols, and so achieve superior probability estimation and redundancy reduction, despite the sequential decoding dependency that it introduces [10]. However, attention mechanisms have been employed in neural compression designs for better latent feature representations. Attention modules, which dynamically adjust the weights of feature channels or spatial regions according to their relevance to the semantic level, can preserve critical visual details and generate more compact representations [13], [14], thereby allocating more representational capacity to complex textures and edges even at extremely low bit rates. The entropy coder is more capable of adapting to varied image statistics by attention-based feature selection and autoregressive modelling, so that learning codecs can be comparable to established standards .

D. Perceptual quality and Rate-distortion tradeoffs

To verify the claim that learned codecs rival established standards, it is necessary to examine how their performance is quantified. Mean squared error and PSNR are common distortion metrics used to evaluate image compression quality. However, low distortion does not always correspond to high perceptual quality. MS-SSIM and other structural metrics provide important information about structural preservation [16], [17]. The rate-distortion framework formalises the tradeoff between bitrate and reconstruction fidelity, which modern neural methods optimise directly during training, ensuring that bits are allocated to preserve the most semantically important visual information.

Table I shows a conceptual comparison of existing image compression approaches, generative latent models, and the proposed Enhanced Compression Model (ECM). It focuses on the main differences in architecture and modeling. The methods in Table I are chosen to reflect the main architectural types, and Section VII provides the full list of baselines utilized for the quantitative experimental evaluation.

TABLE I: Conceptual comparison of representative image compression methods and the proposed ECM.

Feature	JPEG / PNG	Ballé et al. [9]	Minnen et al. [10]	LDM [21]	ECM (Proposed)
Neural Network Based	×	✓	✓	✓	✓
Hyperprior	×	✓	✓	×	✓
Autoregressive Context Modeling	×	×	✓	×	✓
Channel Attention	×	×	×	✓	✓
Semantic Gating	×	×	×	×	✓
Entropy Parameter Network	×	✓	✓	×	✓
Scale Refinement	×	×	×	×	✓
Adaptive Feature Prioritization	×	×	Partial	✓	✓

III. Proposed Hybrid Compression Framework

A. System overview

Enhanced Compression Model is based on a learned image compression pipeline that uses an improved hyperprior model. The encoder transforms the input image into a compact latent representation that is semantically gated and processed through a hyperprior branch to estimate entropy. A context model extracts autoregressive spatial information from the latents, and an entropy parameter network uses hyperprior statistics and context features to predict means and scales for Gaussian

conditional coding. The quantised latent representation is then decoded to reconstruct the image.

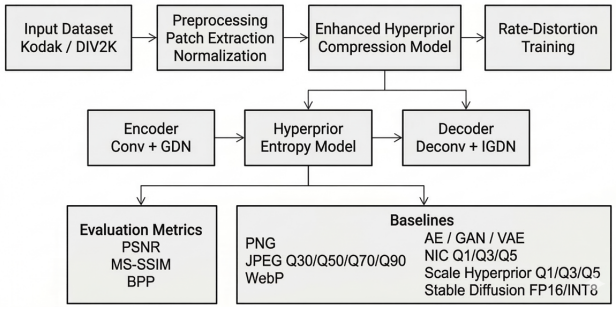


Fig. 1: System overview of *Enhanced Compression Model*

B. Design rationale

The proposed model's design is grounded in the principle that compression efficiency depends on the compactness of the latent representation and the accuracy of its entropy model. The encoder must produce feature maps that are information-rich yet compressible. The entropy model must estimate latent probabilities as accurately as possible to minimise the number of coded bits.

By focusing on channel attention we are able to extract features better by highlighting important channels and making the less relevant ones quiet. Next, semantic gating is done to the latent representation before processing the hyperprior to retain more semantic content, This way making the latent codes not only compact but also structurally relevant.

The hyperprior branch uses hyperlatents to model global statistical structure, while the context model captures local spatial dependencies between latent coefficients. The entropy parameter network combines these two information sources to predict means and scales for Gaussian conditional coding. A scale refinement step is applied before the likelihood calculation to further improve the predicted scales. This combination facilitates more accurate entropy estimation and enables encoding at lower bitrates.

C. What makes Enhanced Compression Model unique

The *Enhanced Compression Model* is unique because it combines several complementary modules into a single learned compression architecture.

- **Channel attention** is used in both analysis and synthesis transforms to improve feature learning [13], [14].
- **Semantic gating** adaptively filters the latent representation so that channels carrying more useful information have a greater impact on entropy modeling and reconstruction.
- By taking into account nearby latent dependencies, **autoregressive context modeling** improves probability estimation [10].
- **Scale refinement** corrects the predicted Gaussian scales before entropy coding.
- The entropy parameter prediction uses both hyperprior and context features to obtain a more accurate characterization of the coding distribution.

The standard scale hyperprior backbone is actually kept in house; these modules do the job of upgrading it. Each component of the baseline is progressively improved. Ultimately, their combination yields the highest benefit. The importance of their cooperative interaction is highlighted in Section IX, where it is shown that a change in any one module results in a drop in both PSNR and compression efficiency.

Semantic gate is a soft channel-selection instrument that is trained end-to-end with the rate-distortion objective and doesn't require explicit supervision. It is tough in the sense that it picks out useful features like edges and textures, while at the same time it suppresses redundant information. This not only promotes compression efficiency, but also results in lesser latent entropy.

IV. Enhanced Hyperprior Model

A. Architecture summary

The proposed Enhanced Hyperprior Model extends the standard Scale Hyperprior design with additional modules that improve latent feature processing and entropy parameter estimation [9].

The analysis transform g_a uses reflection-padded convolutions and generalised divisive normalisation (GDN) layers to transform the input image into a latent representation [7]. Reflection padding facilitates consistent border handling during convolutional encoding. Channel attention is applied to the encoder output to emphasize informative latent channels [13], [14].

The synthesis transform g_s uses transposed convolutions and inverse GDN layers to mirror the encoder. Channel attention is also applied to the decoder input to facilitate feature reconstruction from quantised latents [13], [14].

The hyper-analysis transform h_a maps the latent representation to hyperlatents, and the hyper-synthesis transform h_s uses quantised hyperlatents to predict hyperprior scale features. A dedicated scale refinement module then improves these predicted scales using direct latent information.

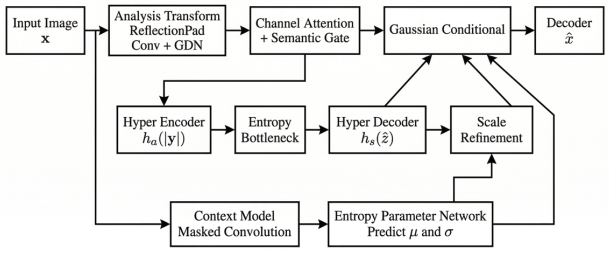


Fig. 2: Model Architecture of *Enhanced Compression Model*

B. Semantic gating and context modeling

After the analysis process, the latent tensor passes through a semantic gate that consists of: The gate suppresses the less informative channels and retains the semantically significant features. The context model uses a masked-convolution approach to gather neighbouring information in a causal manner and provide the autoregressive context for the prediction

of entropy parameters. By taking advantage of the spatial correlations between nearby latent coefficients, it improves the probability estimation and reduces the bitrate [10].

Although decoding becomes more difficult because of the sequential dependencies in autoregressive context modelling, the resulting" bitrate savings justifies the additional expenditure. Also, ways of parallelising such as chequerboard decoding [10] can be considered for future implementation in tackling the issue of decoding latency.

C. Entropy parameter net and scale refinement

The entropy parameter network combines refined hyperprior scale features and autoregressive context features by concatenating them channel-wise and projecting them pointwise through convolutional layers. The network predicts the means and scales for Gaussian conditional entropy coding. Predicted scales are clamped before likelihood calculation to ensure numerical stability.

Scale refinement is important because hyperprior scale estimates may be inaccurate for highly textured or structurally complex image content. Refinement improves entropy estimation accuracy and overall coding efficiency.

V. Mathematical Formulation

A. Transforms and quantization

Let $x \in \mathbb{R}^{H\times W\times 3}$ be the input. The analysis transform produces latents:

$$y = g_a(x;\theta). \tag{1}$$

Quantization (training-time relaxation may replace rounding with uniform noise):

$$\hat{y} = Q(y). \tag{2}$$

Hyper-analysis and hyper-synthesis:

$$z = h_a(y;\phi), \quad \hat{z} = Q(z), \quad \tilde{z} = h_s(\hat{z};\psi). \tag{3}$$

B. Entropy model

An entropy model parameterizes $p(\hat{y} \mid \tilde{z}, \text{context})$ using distribution parameters (e.g., mean/scale; or mixture parameters):

$$p(\hat{y} \mid \tilde{z}) = \prod_i p(\hat{y}_i \mid \mu_i(\tilde{z}), \sigma_i(\tilde{z})). \tag{4}$$

C. Decoder and reconstruction

Reconstruction:

$$\hat{x} = G(\hat{y}, c; \gamma), \tag{5}$$

where c denotes conditioning signals from the entropy modeling pipeline or auxiliary refinement paths.

D. Training objective

A Rate-distortion objective template:

$$\mathcal{L} = \lambda D(x,\hat{x}) + R(\hat{y},\hat{z}) + \alpha\mathcal{L}_{\text{perc}} + \beta\mathcal{L}_{\text{GAN}}. \tag{6}$$

In the present implementation, training primarily optimizes a Rate-distortion objective using BPP together with either MSE-based distortion or MS-SSIM-based distortion, depending on the selected criterion. The reported evaluation metrics in this work are Train Loss, PSNR, MS-SSIM, and BPP.

VI. Training Setup

A. Datasets

The DIV2K high-resolution dataset, comprising 800 high-quality images widely used for image restoration and compression research, is used to train the model [18]. Patch-based training is employed to increase data diversity and reduce memory consumption. Random crops of size 256×256 are extracted from each image during training, with random horizontal and vertical flips applied as augmentation.

The Kodak dataset is used for validation and performance evaluation. It contains 24 uncompressed natural images [19] spanning a wide range of textures, colours, and spatial structures, making it a standard benchmark for compression model evaluation.

Images are processed through a consistent validation pipeline, and reconstructed outputs are compared to originals using objective metrics including PSNR and MS-SSIM [16], [17].

B. Training hardware

An NVIDIA RTX 5080 GPU with CUDA acceleration is used for all experiments. Mixed precision training was enabled to improve computational efficiency and reduce GPU memory usage. Gradient scaling and automatic mixed precision were utilized during training to maintain numerical stability.

The model was implemented in PyTorch and trained with the AdamW optimiser. Training checkpoints were saved periodically to enable recovery and evaluation of intermediate models. The training pipeline records validation metrics including BPP, PSNR, and MS-SSIM after each epoch to monitor convergence.

C. Hyperparameters

The training configuration follows the implementation used for the proposed model:

- Patch size: 256×256
- Optimiser: AdamW for the main network parameters and Adam for the auxiliary entropy bottleneck parameters
- Learning rate: 1×10^{-4}
- Auxiliary learning rate: 1×10^{-3}
- Batch size: 4
- Number of epochs: 50
- Random seed: 42
- Mixed precision training: enabled on CUDA
- Validation batch size: 1

D. Training protocol

During each training epoch, the model processes randomly cropped training patches from DIV2K. The main optimiser applies gradient clipping to stabilise network parameter updates, while the auxiliary optimiser independently updates the entropy bottleneck quantile parameters. After each epoch, model parameters are updated for entropy coding and the Kodak dataset is used for validation. BPP, PSNR, and MS-SSIM values are logged to a CSV file alongside the training loss.

VII. Experimental Setup

A. Baselines

The proposed model is compared against classical, neural, and generative compression baselines:

- PNG (lossless) [2]
- JPEG at multiple quality factors [1]
- WebP [3]
- Autoencoder-based baseline [4]
- GAN-based baseline [6]
- VAE-based baseline [5]
- NIC baselines at multiple quality settings [8]
- Scale Hyperprior baselines [9]
- Stable Diffusion VAE-based baselines [21]

These baselines provide a broad comparison across traditional coding methods, simple neural compression methods, entropy-model-based learned compression models, and generative latent compression approaches.

B. Metrics

The metrics used to evaluate the methods are:

- BPP (bits per pixel): measures bitrate efficiency; lower is better.
- PSNR (dB): measures pixel-level reconstruction quality; higher is better.
- MS-SSIM (unitless): measures structural similarity between reconstructed and original images; higher is better [16], [17].
- Train Loss: reports the combined training objective value; lower is better.

C. Evaluation protocol

The evaluation protocol compares reconstructed outputs against original Kodak images. Bitrate is measured in bits per pixel, and reconstruction quality is assessed using PSNR and MS-SSIM. Train Loss is reported to characterise model convergence and the final operating point. Together, these metrics provide a comprehensive picture of compression efficiency, distortion, and structural preservation.

VIII. Results and Evaluation

A. Compression Efficiency Analysis

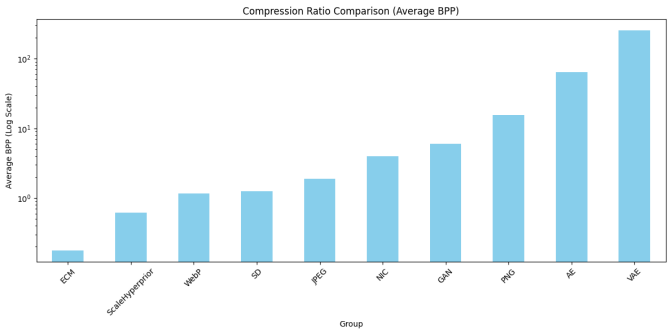


Fig. 3: Comparison of average BPP across the evaluated compression methods. Lower BPP indicates better compression efficiency.

Figure 3 summarises the average bits-per-pixel compression efficiency across all evaluated methods. The proposed Enhanced Compression Model achieves **0.176 BPP**, which is **13.7%** lower than Scale Hyperprior Q1 (0.204 BPP) and **65.9%** lower than Scale Hyperprior Q3 (0.517 BPP), while preserving significantly greater reconstruction quality than both baselines. The model runs at a considerably lower bitrate than most conventional codecs, while still delivering high-quality visual reconstructions.

B. Rate-distortion Performance

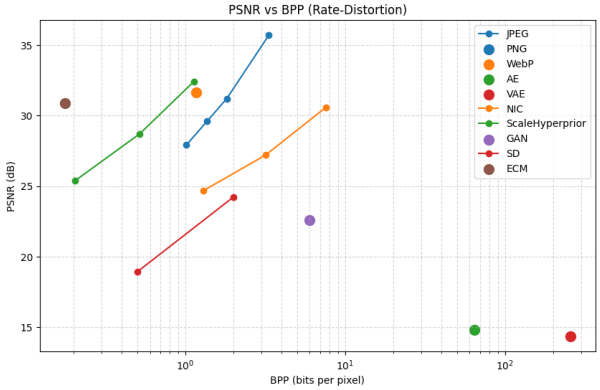


Fig. 4: PSNR versus BPP Rate-distortion comparison of the evaluated methods. Higher PSNR at lower BPP indicates superior compression performance.

Figure 4 shows the PSNR–BPP rate-distortion performance. ECM achieves **30.882,dB PSNR** at **0.176 BPP**, outperforming Scale Hyperprior Q1 by **5.5,dB** and approaching the quality of Scale Hyperprior Q5 while using only **15.6%** of its bitrate. These results demonstrate ECM's ability to maintain high reconstruction quality at low bitrates.

C. Structural Quality Evaluation

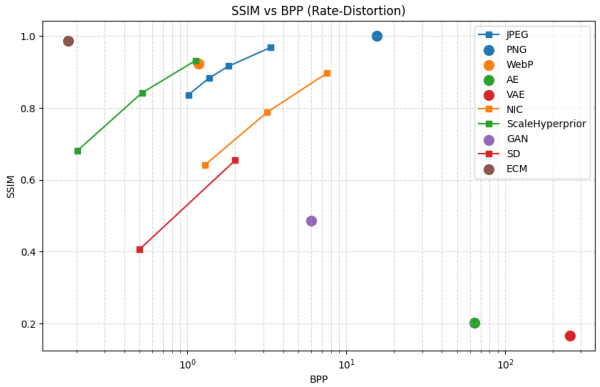


Fig. 5: MS-SSIM versus BPP comparison for the evaluated methods. Higher MS-SSIM at lower BPP indicates better structural preservation under compression.

Figure 5 compares structural preservation across methods. ECM achieves the **highest MS-SSIM of 0.985** across all evaluated methods at only 0.176 BPP, compared to 0.932 for

Scale Hyperprior Q5 at 1.127 BPP and 0.968 for JPEG Q=90 at 3.328 BPP. This confirms that semantic gating and improved entropy modeling contribute not only to rate efficiency but also to stronger structural preservation.

D. Average Reconstruction Quality



Fig. 6: Average PSNR comparison across the evaluated compression methods. Higher values indicate better reconstruction quality.

Figure 6 compares average PSNR across all evaluated methods. While some high-bitrate classical baselines achieve higher PSNR values due to their operating points, the proposed method achieves strong fidelity at a substantially lower bitrate, making it more suitable for bandwidth-constrained applications.

E. Perceptual Distribution Comparison

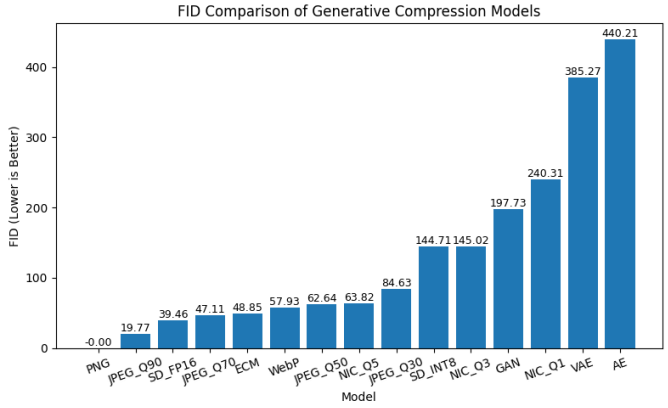


Fig. 7: FID comparison of the evaluated compression models. Lower FID values indicate better perceptual realism of reconstructed images.

Figure 7 presents a perceptual comparison using FID among the evaluated methods [15]. The primary objective of this study is rate-distortion performance as measured by BPP, PSNR, MS-SSIM, and Train Loss. The perceptual comparison illustrates how different models behave with respect to image distribution fidelity upon reconstruction.

TABLE II: Codec comparison on the primary test set (Kodak). Metrics include BPP, PSNR, MS-SSIM, and Train Loss (LOSS ↓).

Method	BPP	PSNR (dB)	MS-SSIM	LOSS ↓
PNG (lossless)	15.528	∞	1.000	0.000
JPEG Q=30	1.014	27.916	0.836	0.063
JPEG Q=50	1.365	29.589	0.883	0.027
JPEG Q=70	1.815	31.202	0.916	0.015
JPEG Q=90	3.328	35.696	0.968	0.004
WebP	1.169	31.621	0.923	0.018
AE	64.000	14.825	0.203	0.845
GAN	6.000	22.602	0.486	0.534
VAE	256.000	14.356	0.166	0.843
NIC Q1	1.293	24.675	0.642	0.369
NIC Q3	3.180	27.207	0.788	0.200
NIC Q5	7.555	30.572	0.896	0.048
Scale Hyperprior Q1	0.204	25.380	0.682	0.349
Scale Hyperprior Q3	0.517	28.704	0.842	0.149
Scale Hyperprior Q5	1.127	32.398	0.932	0.022
SD FP16	2.000	24.232	0.654	0.058
SD INT8	0.500	18.934	0.407	0.204
Enhanced Compression Model	**0.176**	**30.882**	**0.985**	**0.0428**

Table II shows the quantitative comparison across all evaluated methods. The Enhanced Compression Model achieves the lowest bitrate among all lossy methods while maintaining a high PSNR and the highest MS-SSIM score, confirming effective structural preservation under heavy compression.

TABLE III: Rate-distortion operating points. Choose matched BPP points and report PSNR and MS-SSIM.

Method	Setting	BPP	PSNR (dB)	MS-SSIM
JPEG	Q=30	1.013	27.916	0.835
JPEG	Q=50	1.365	29.588	0.882
JPEG	Q=70	1.815	31.201	0.915
JPEG	Q=90	3.328	35.696	0.967
WebP	Q=75 , method=6	1.464	33.359	0.945
Scale Hyperprior	λ=0.1	1.390	28.822	0.970
ECM	λ=0.1	**0.176**	**30.882**	**0.985**

Table III presents rate-distortion operating points at comparable settings. At λ=0.1, ECM achieves **30.882 dB PSNR** at **0.176 BPP**, versus Scale Hyperprior at 28.822 dB at 1.390 BPP delivering **2.06 dB higher PSNR** at **less than one-eighth the bitrate**. This demonstrates that ECM is substantially more rate-distortion efficient than both traditional and learned baselines.

TABLE IV: FID results on the Kodak dataset. Lower FID values indicate better image quality and distribution matching.

Method	Dataset	FID ↓
JPEG Q=30	Kodak	84.633
JPEG Q=50	Kodak	62.635
JPEG Q=70	Kodak	47.110
JPEG Q=90	Kodak	19.766
PNG (Lossless)	Kodak	0.000
WebP	Kodak	57.926
AE	Kodak	440.209
VAE	Kodak	385.269
GAN	Kodak	197.731
NIC Q1	Kodak	240.310
NIC Q3	Kodak	145.021
NIC Q5	Kodak	63.822
SD FP16	Kodak	39.456
SD INT8	Kodak	144.712
Enhanced Compression Model	Kodak	**48.848**

Table IV shows FID scores for the Kodak dataset. The proposed model achieves a competitive FID of 48.848, comparable to JPEG Q=70 (47.110) and substantially better than WebP (57.926), demonstrating that rate-distortion-optimised ECM maintains reasonable perceptual quality relative to established codecs.

IX. Ablation Study

The contribution of the key components of the proposed Enhanced Compression Model (ECM) was analysed through an ablation study. Each variant alters/removes a particular module, while the training configuration, dataset, and evaluation procedure remain unchanged. This paper studies the role of entropy modelling and semantic gating, and presents the performance results on the Kodak dataset for BPP, PSNR, and MS-SSIM. The results are presented in Table V. The basic configuration (A0) incorporates the entire ECM with hyperprior, autoregressive context model, entropy parameter network, and semantic gating. This entire model achieves the highest performance of 30.882, dB PSNR and 0.985 MS-SSIM, indicating that the addition of these components improves both compression efficiency and reconstruction quality.

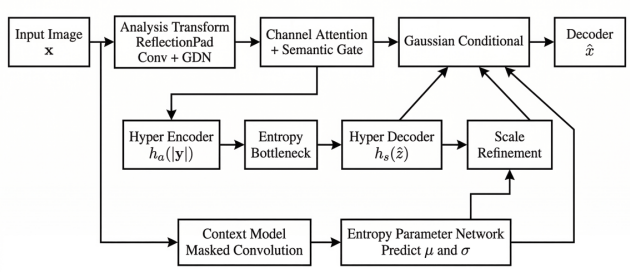


Fig. 8: The complete Enhanced Compression Model (A0) architecture.

Experiment A1 removes the context model and the entropy parameter network altogether. The model without them is unable to estimate latent probability distributions for entropy coding correctly. As a result, the bitrate goes up drastically to 1.865 BPP and the PSNR drops to 28.753 dB. So, this really supports that for good compression, one needs to accurately model latent probabilities.

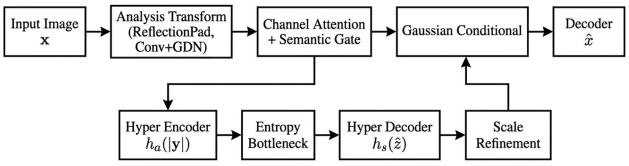


Fig. 9: Architecture variant A1: Removal of autoregressive context model and entropy parameter network

In Experiment A2, we only took out the context model but kept the hyperprior module and the entropy parameter network. Getting rid of the context model means the system only gets to know the overall statistical information from the hyperprior but still doesn't find the local spatial dependencies between neighboring latent elements. This explains the increase in bitrate to 0.602 BPP and the decrease in PSNR to 28.562 dB which further corroborates the significance of autoregressive context modeling in effective entropy estimation.

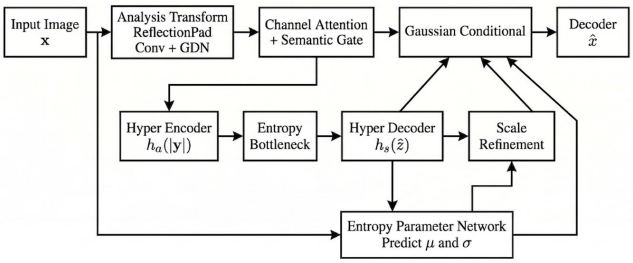


Fig. 10: Architecture variant A2: Removal of the autoregressive context model.

Experiment A3 investigates the contribution of the semantic gating mechanism. The semantic gate modulates the latent representation before hyperprior analysis, enabling the network to focus on informative features. Removing this component causes a moderate degradation, with bitrate rising to 0.283 BPP and PSNR falling to 28.465 dB, confirming that semantic gating produces more compact and informative latent representations.

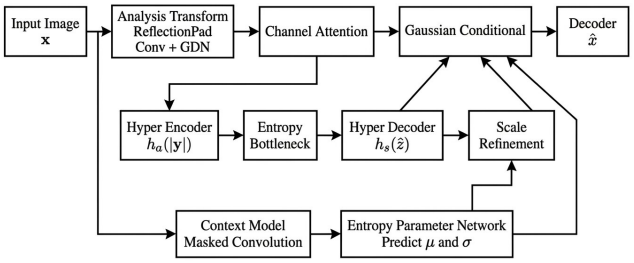


Fig. 11: Architecture variant A3: Removal of the semantic gating mechanism.

The ablation results confirm that entropy modeling components contribute most significantly to bitrate reduction, while semantic gating further improves latent representation quality and compactness. The full Enhanced Compression Model consistently achieves the best rate-distortion performance across all tested configurations.

TABLE V: Ablation study results on the Kodak dataset. Each variant disables one component; all other training conditions are held constant.

Variant	BPP	PSNR (dB)	MS-SSIM	TRAIN LOSS ↓
A0: Full Model	0.176	30.882	0.985	8.245
A1: w/o context model and entropy parameter	1.865	28.753	0.968	12.836
A2: w/o context model	0.602	28.562	0.970	12.838
A3: w/o semantic gate	0.283	28.465	0.977	11.228

In summary, the single largest improvement is context-aware entropy modelling (A1 vs. A0), which reduces BPP from 1.865 to 0.176 a **90.6% bitrate reduction** while regaining **2.13 dB PSNR**. Semantic gating (A3 vs. A0) provides an extra **0.107 BPP reduction** and **0.417 dB PSNR improvement** compared to the gate-removed variation. These results demonstrate that the performance advantages of ECM are not due to a single architectural modification but rather to the cooperation of numerous modules.

X. Conclusion

The paper introduced the Enhanced Compression Model (ECM), a neural image compression system working on the hyperprior principle and improving compression efficiency and reconstruction quality through semantic gating, channel attention, autoregressive context modelling, entropy parameter prediction, and scale refinement.

The experimental results on the Kodak dataset show that ECM achieves 30.882 dB PSNR and 0.985 MS-SSIM at a very low bit rate of only 0.176 BPP. Additional ablation tests confirm the importance of context modelling, entropy parameter prediction and semantic gating in achieving better rate-distortion performance.

These results highlight the power of sophisticated entropy modelling and latent feature modulation in learnt image compression. The proposed system offers a solid base for further research, potentially addressing applications where limited bandwidth is a problem, like in satellite imaging telemedicine intelligent vehicles, and mobile media communication. The next steps will be focused on reducing the complexity of autoregressive decoding and further extend the scope of ECM by incorporating video compression.

References

[1] ITU-T, "Information technology—Digital compression and coding of continuous-tone still images—Requirements and guidelines," ITU-T Rec. T.81 / ISO/IEC 10918-1, Sep. 1992.

[2] W3C PNG Working Group, "Portable Network Graphics (PNG) Specification (Third Edition)," W3C Recommendation, Jun. 2025.

[3] Google Developers, "Compression Techniques," *WebP Documentation*, Aug. 2025.

[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014.

[6] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.

[7] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density Modeling of Images Using a Generalized Normalization Transformation," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2016.

[8] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-End Optimized Image Compression," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017.

[9] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational Image Compression with a Scale Hyperprior," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2018.

[10] D. Minnen, J. Ballé, and G. D. Toderici, "Joint Autoregressive and Hierarchical Priors for Learned Image Compression," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[11] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7939–7948.

[12] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-Fidelity Generative Image Compression," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 11913–11924.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132–7141.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[17] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale Structural Similarity for Image Quality Assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.

[18] E. Agustsson and R. Timofte, "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 126–135.

[19] R. Franzen, "Kodak Lossless True Color Image Suite," 2013. [Online]. Available: `http://r0k.us/graphics/kodak/`

[20] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "CompressAI: A PyTorch Library and Evaluation Platform for End-to-End Compression Research," *arXiv preprint arXiv:2011.03029*, 2020.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.