



ARTICLE

Learned Image Compression via Text-Semantic Guidance and Content-Aware Bitrate Control

Kaisen Li¹, Yunwei Zhang^{1,*}, Guoying Sun¹ and Bin Li^{2,*}

¹Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

²Information Center, Yunnan Tobacco and Leaf Company, Kunming, China

*Corresponding Authors: Yunwei Zhang. Email: zhangyunwei72@gmail.com; Bin Li. Email: libin_home@126.com

Received: 28 April 2026; Accepted: 24 June 2026; Published: 13 August 2026

ABSTRACT: With the development of vision-language pre-trained models, effectively exploiting high-level semantics and precisely controlling bitrate in learned image compression remains a challenging problem. Existing methods mainly rely on image feature modeling alone, making it difficult to jointly preserve fine-grained details and semantic consistency under a given bitrate budget. To address this issue, this paper proposes a learned image compression framework that integrates text-semantic guidance with content-aware bitrate control. The framework combines Bootstrapping Language-Image Pre-training (BLIP) and Contrastive Language-Image Pre-training (CLIP) to extract image semantic information, and performs conditional modulation on multi-scale visual features through feature-wise linear modulation to construct semantically enhanced latent representations, thereby improving structural preservation and visual reconstruction quality under low-bitrate conditions. Meanwhile, we design a bitrate control module that integrates content awareness with historical feedback. Regional importance is estimated by channel attention, and the quantization strength is adaptively adjusted according to the target and actual bitrates, enabling the model to approach the preset bitrate without manual tuning. Experimental results on different datasets show that, compared with representative learned image compression methods, including Cheng2020, Efficient Learned Image Compression (ELIC), and Frequency-aware Transformer for Learned Image Compression (FTIC), the proposed method achieves a more competitive rate-distortion trade-off. In particular, it consistently exhibits superior rate-distortion (RD) performance, delivers higher objective reconstruction quality at the same bitrate, and maintains favorable subjective perceptual consistency.

KEYWORDS: Learned image compression; feature-wise linear modulation; semantic guidance; semantically enhanced latent representation; content-aware bitrate control

1 Introduction

Traditional lossy image compression standards, such as Joint Photographic Experts Group (JPEG) [1], JPEG2000 [2], and Better Portable Graphics (BPG) [3], have achieved remarkable success in the field of image coding. However, such methods rely on fixed linear transforms, quantization, and entropy coding modules, making them difficult to adapt to the diverse redundancy characteristics present in complex images, and their overall performance has gradually approached saturation. In recent years, end-to-end image compression methods based on deep learning have achieved remarkable progress owing to their ability to learn expressive feature representations. Ballé et al. [4] pioneered a convolutional neural network (CNN)-based end-to-end compression framework, which established the technical foundation of this field by optimizing a joint rate-distortion objective. Subsequent studies have improved the model's efficiency in extracting complex

image features and further enhanced compression performance by incorporating network architectures such as self-attention mechanisms [5,6], residual blocks [7,8], Transformer-based modules [9,10], reversible structures [11], and wavelet-based modules [12,13].

Nevertheless, learned image compression methods still exhibit clear limitations in entropy modeling and bitrate allocation. Early methods predominantly modeled latent variables using unimodal parametric distributions, such as Gaussian scale mixtures. However, such formulations are insufficient to capture the multimodal and highly complex distribution characteristics of image latents. As a result, the mismatch in probability estimation introduces additional rate redundancy. Agustsson et al. [14] pointed out that, under highly textured image scenarios, the probability estimation error caused by such unimodal distribution assumptions may lead to an increase in bitrate redundancy of approximately 10%–20%. In addition, most existing entropy models focus solely on the distribution of low-level features of latent variables and do not incorporate high-level semantic information from images, resulting in the loss of structure and detail in semantically critical regions at high compression ratios. Li et al. [15] reported that, when the compression ratio exceeded 40:1, traditional entropy models exhibited over 30% greater structural distortion in semantically important regions than semantic-aware models. Bit allocation is a core component of image compression frameworks for balancing coding cost against reconstruction distortion. Its fundamental objective is to allocate limited bit resources preferentially to regions with high information density, thereby achieving an optimal global trade-off between rate and distortion. Existing bitrate allocation strategies are generally implemented as block-level uniform regulation under global constraints, but they lack fine-grained local adaptivity. Most methods first determine the average bit allocation for image blocks according to the global target bitrate, and then distribute bits within each block using a fixed model, without achieving content-aware fine-grained regulation. Minnen et al. [16] demonstrated that such a block-level uniform allocation strategy may result in bit over-allocation or under-allocation in regions with drastic local fluctuations in information density, such as edge-to-flat transition areas, thereby significantly degrading local reconstruction quality.

Therefore, this paper proposes a text-semantics-guided adaptive image compression framework. This framework uses a pre-trained model to convert images into textual descriptions, combines these with a text encoder to map them into semantic vectors, and employs feature-wise linear modulation to perform affine modulation on multi-scale visual latent features, thereby enabling high-level semantic information to guide the compression process. Meanwhile, we design a dual-channel rate control module that integrates content awareness with historical feedback. By jointly leveraging the semantic complexity of the image and historical bit statistics, the module adaptively adjusts the quantization scaling factor and the Lagrange multiplier, thereby enabling more stable and precise bitrate control.

The main contributions of this paper are as follows:

- A semantic-guided compression framework is proposed, which leverages a pre-trained vision-language model to extract high-level semantic cues from images and embeds them into the multi-scale latent space through feature-wise linear modulation, thereby enabling the encoding process to consistently preserve critical content, structural information, and contextual details.
- A three-branch convolutional fusion mechanism based on channel concatenation is designed, in which shallow features, textual features, and deep features are aligned along the channel dimension and jointly integrated through convolution. Compared with element-wise addition or simple linear projection, this strategy provides a more flexible interaction paradigm between semantic and visual signals.
- A dynamic bitrate control strategy that integrates content awareness with historical feedback is proposed. By constructing a dual-channel controller that jointly adjusts the quantization scaling factor and

the Lagrangian multiplier, the strategy adaptively allocates bits to achieve more precise and stable global bitrate control under non-uniform content and texture distributions.

2 Related Work

2.1 Semantic Text-Guided Image Compression

To improve content understanding, many approaches incorporate high-level semantic information into the compression process. Lee et al. [17] proposed the Text-Adaptive Compression (TACO) framework, which introduces a text-adaptive module at the encoder side and employs a joint image-text loss, thereby taking into account both pixel-level Peak Signal-to-Noise Ratio (PSNR) and Learned Perceptual Image Patch Similarity (LPIPS)-based perceptual quality. Qin et al. [18] proposed a perceptual image compression method assisted by cross-modal side information, which enhances the subjective perceptual quality of reconstructed images by fusing textual and visual features and leveraging semantic masks to guide the reconstruction process at the pixel level. Segmentation-Prior-Guided Image Compression (SegPIC) [19] employs region-adaptive transformations on different regions based on semantic segmentation priors, thereby better preserving structural information. In addition, semantic-assisted compression methods, such as Semantic-Assisted Image Compression (SAIC) [20], embed task semantics or semantic features into the bitstream to improve the performance of downstream vision tasks. Unlike segmentation priors or semantic-assisted schemes, recent methods such as Prompt Inversion Compressor with Sketch (PICS) and Multimodal Image Semantic Compression (MISC) directly exploit textual descriptions and generative models to reconstruct images that are semantically consistent with the original ones at extremely low bitrates [21,22]. However, these methods typically sacrifice pixel-level fidelity in exchange for improved subjective quality through semantic alignment. In contrast, our method utilizes pre-trained vision-language models such as Contrastive Language-Image Pre-training (CLIP) and Bootstrapping Language-Image Pre-training (BLIP) to extract descriptive text, and performs channel-level adjustment of encoded features at the latent level through feature-wise linear modulation. This allows semantic information to be implicitly embedded in the bitstream as parameters, eliminating the need for additional text input at the decoding stage while improving metrics such as PSNR.

2.2 Variable Bit Rate and Content-Aware Bit Rate Control

Variable-rate compression and precise rate control are critical issues for the practical deployment of learned compression methods. The content-weighted compression proposed by Li et al. [15] explicitly models variable bit allocation in the latent space using an importance map, thereby achieving strong content awareness. Building upon this idea, Song et al. [23] proposed a variable-rate compression framework based on spatially adaptive feature transformation, in which feature transformation is guided by a quality map, allowing a single model to cover a continuous range of bitrates. From an optimization perspective, Guerin et al. [24] proposed a learning-based image compression method with bitrate constraints, which imposes constraints on a given target bitrate. Gupta et al. [25] further controlled local bit allocation by having users specify priority regions, thereby achieving user-guided variable-rate compression.

In addition to the aforementioned methods, numerous studies have further improved the rate adaptability of a single model from the perspectives of model architecture and training strategy. Choi et al. [26] proposed a conditional autoencoder framework that uses the Lagrange multiplier and quantization step size as conditional inputs, simultaneously controlling coarse-grained and fine-grained bitrate adjustment to achieve continuous target bitrate coverage. Yang et al. [27] superimposed a lightweight modulation network on a shared encoder by modulating the autoencoder, applying channel gain to the latent variables, and effectively approximating the rate-distortion upper bound for multi-model training. Kamisli et al. [28]

employed multi-objective optimization and quantization-reconstruction offset on a single baseline model to improve the variable-bit-rate performance of schemes that rely solely on a uniform quantization step size. In the context of regions of interest (ROI), Ma et al. [29,30] proposed an ROI compression method and a subsequent Transformer-based ROI encoder. By using ROI masks or prompt-generation networks to identify areas of interest, these approaches guide the model to allocate more bits to critical regions, thereby improving the visual quality of regions of interest at the same bitrate.

Inspired by the work described above, we adopt a similar single-model variable-bit-rate design, but focus further on the expressive power of the bitrate control signal itself. Existing methods use scalar metrics such as importance maps, quality maps, or target bitrates as constraints; however, the control signals employed still lack a strong correlation with the high-level semantics of the image and the actual encoding history. Therefore, this paper introduces a dual-rate controller within a framework based on a hyperprior and contextual entropy model. By using text-modulated semantic features and historical bit statistics as joint conditions, it jointly adjusts the quantization scaling factor and the Lagrange multiplier within a unified network, thereby achieving fine-grained bit allocation that balances content awareness and historical stability.

3 Methods

3.1 Overall Architecture

Adaptive compression, as illustrated in Fig. 1, differs from traditional learning-based compression methods that rely solely on visual features from a single pathway. This framework systematically incorporates textual semantic information into both feature extraction and probabilistic modeling to achieve adaptive compression of image content.

The specific workflow is as follows. Given an input image x , BLIP first generates a textual description of the image, which is then encoded by the text-to-tensor encoder g_{a2} into a compact semantic representation y_2 . The shallow image encoder g_{a1} performs local structural analysis on x to obtain shallow visual features y_1 , while the deep image encoder g_{a3} extracts high-level semantic and global layout information over a larger receptive field, yielding deep features y_3 . Taking the semantic tensor y_2 as input, the feature-wise linear modulation parameter generator predicts the affine scaling and bias parameters $\{\gamma_1, \beta_1, \gamma_3, \beta_3\}$ for the two visual branches. It then applies channel-wise affine modulation to y_1 and y_3 , producing the semantically enhanced features $\tilde{y}_1$ and $\tilde{y}_3$. Finally, $\tilde{y}_1$, y_2 , and $\tilde{y}_3$ are concatenated along the channel dimension to form a unified latent representation y . In the rate control module, the Lagrange multiplier and scaling factor $\{\lambda, \alpha\}$ are used for global bitrate scheduling. During quantization and entropy coding, the hyperprior analysis network h_a first maps the main latent variable to the hyper-latent variable z , and then y and z are quantized to obtain $\hat{y}$ and $\hat{z}$. $\hat{z}$ is processed by arithmetic encoding/arithmetic decoding (AE/AD) to form an auxiliary bitstream, which is fed into the hyper-synthesis network h_s at the decoder to generate the parameters of a Gaussian distribution for the main latent variable $\hat{y}$. Finally, $\hat{y}$ is fed into the main image decoder g_s , and the reconstructed image $\hat{x} = g_s(\hat{y})$ is progressively obtained through multiple layers of TConv and ResGroup.

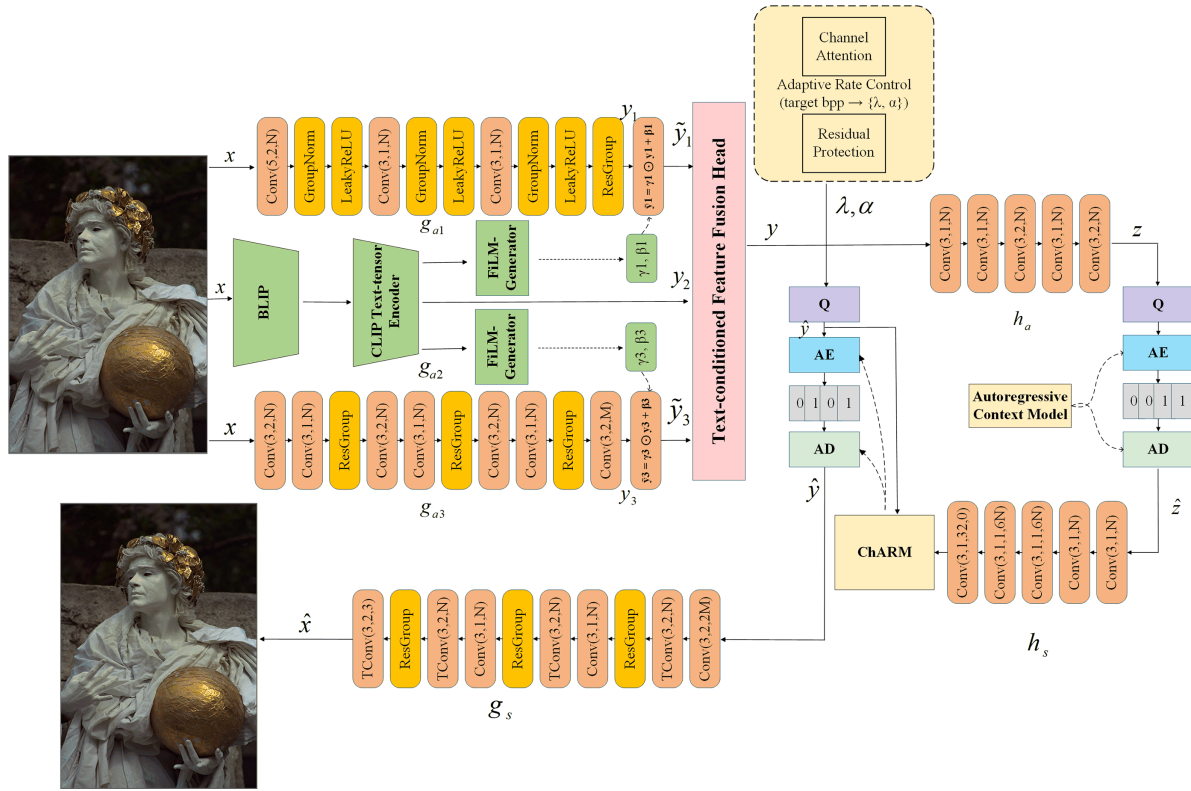


Figure 1: Overall architecture of the proposed text-conditioned image compression framework. $\text{Conv}(k, s, n)$ denotes a convolutional layer, where the kernel size is $k \times k$, the stride is s , and the number of output channels is n . $\text{TConv}(k, s, n)$ denotes a transposed convolutional layer, and ResGroup represents a residual group. The text branch exploits BLIP and CLIP to extract semantic guidance information, which is then injected into the visual encoding network through a feature-wise linear modulation generator and a text-conditioned feature fusion head. Q, AE, and AD denote quantization, arithmetic encoding, and arithmetic decoding, respectively. ChARM, together with an autoregressive context model, is employed for entropy modeling.

3.2 Text Semantic Fusion Module

The upper and lower visual branches are denoted as g_{a1} and g_{a3} , respectively. g_{a1} is a shallow image encoder composed of multiple convolutional blocks and residual groups, which captures fine-grained structural information such as local textures and edges in the image. Feature-wise linear modulation uses the text conditions (γ_1, β_1) to perform channel-wise affine modulation on these low-level features, enabling the model to achieve fine-grained control according to the text in terms of color and detail. g_{a3} is a deep image encoder that, through repeated downsampling and stacked residual blocks, produces a deep feature tensor y_3 containing high-level information such as the overall image structure. Feature-wise linear modulation applies semantic modulation at this scale through (γ_3, β_3) , allowing the network to perform global control according to the text in terms of object shape and regional style. The shallow feature y_1 and deep feature y_3 obtained from the two visual branches are

$$y_1 \in \mathbb{R}^{B \times C_1 \times H \times W}, \quad y_3 \in \mathbb{R}^{B \times C_3 \times H \times W} \quad (1)$$

where B denotes the batch size, C_1 and C_3 denote the numbers of feature channels in the two branches, and H and W denote the spatial resolution of the latent feature maps. By inserting upsampling/downsampling layers within the branches, the features from both branches are aligned to the same spatial size $H \times W$ at the

output. g_{a2} first employs BLIP to generate a description, and then uses the CLIP text encoder to obtain the semantic text vector

$$y_2 = g_{a2}(x) \in \mathbb{R}^{B \times D} \quad (2)$$

where D denotes the text embedding dimension. The feature-wise linear modulation parameter generator predicts the modulation parameters for the two visual branches based on y_2 :

$$(\gamma_1, \beta_1), (\gamma_3, \beta_3) = \text{FiLMG}_{en}(y_2) \quad (3)$$

$$\gamma_1, \beta_1 \in \mathbb{R}^{B \times C_1 \times 1 \times 1}, \quad \gamma_3, \beta_3 \in \mathbb{R}^{B \times C_3 \times 1 \times 1} \quad (4)$$

The visual features from the two channels are subjected to channel-wise affine transformation to obtain the visual features under text conditions:

$$\tilde{y}_1 = \gamma_1 \odot y_1 + \beta_1, \quad \tilde{y}_3 = \gamma_3 \odot y_3 + \beta_3 \quad (5)$$

where $\odot$ denotes element-wise multiplication, $\tilde{y}_1 \in \mathbb{R}^{B \times C_1 \times H \times W}$, and $\tilde{y}_3 \in \mathbb{R}^{B \times C_3 \times H \times W}$.

To ensure stable training, the feature-wise linear modulation (FiLM) is initialized close to an identity mapping. FiLM performs a text-conditioned channel-wise affine transformation on the visual features, where the scaling parameters γ regulate the response strength of each visual channel, and the shifting parameters β introduce semantic-dependent feature offsets. At the initial stage of training, γ is initialized around 1 and β is initialized around 0, so that the modulated features remain close to the original visual features. This identity-like initialization prevents the semantic branch from excessively disturbing the visual representation at the early training stage, thereby improving the stability of model optimization and convergence.

As training progresses, the FiLM generator gradually learns to perform adaptive channel-wise modulation according to the text-semantic embedding. In this process, channels associated with semantically important structures, such as object contours, salient regions, and meaningful textures, are selectively enhanced, whereas channels mainly corresponding to semantically irrelevant or unstable high-frequency responses are relatively suppressed. Therefore, the text branch acts as a conditional modulation signal rather than a direct image generation prior. Since the reconstructed image content is still primarily determined by the visual latent representation, the proposed mechanism reduces the risk of generating semantically plausible but structurally inconsistent textures. In this way, text-semantic guidance helps alleviate texture artifacts and texture mismatch in low-bitrate reconstruction.

After obtaining the feature-wise linearly modulated visual features $\tilde{y}_1$ and $\tilde{y}_3$, together with the text-semantic representation y_2 , feature-level fusion is performed. Since the text-semantic representation and the multi-scale visual feature maps may have inconsistent spatial resolutions, all visual and semantic features are explicitly projected and aligned to a unified channel dimension and spatial resolution before fusion. The three-branch features are then concatenated along the channel dimension after being mapped to a unified channel number C . For the 1×1 convolution projection of the visual features:

$$\hat{y}_1 = \text{Conv}_{1 \times 1}^C(\tilde{y}_1) \in \mathbb{R}^{B \times C \times H \times W}, \quad \hat{y}_3 = \text{Conv}_{1 \times 1}^C(\tilde{y}_3) \in \mathbb{R}^{B \times C \times H \times W} \quad (6)$$

Since the global text vector $y_2 \in \mathbb{R}^{B \times D}$, it is first mapped to C dimensions through a fully connected layer:

$$\tilde{y}_2 = W_t y_2 + b_t \in \mathbb{R}^{B \times C} \quad (7)$$

and then broadcast over the spatial dimensions $H \times W$:

$$\hat{y}_{2,b,c,h,w} = \bar{y}_{2,b,c}, \quad \hat{y}_2 \in \mathbb{R}^{B \times C \times H \times W} \quad (8)$$

We extract the token-level feature sequence from the Transformer text encoder, denoted as $H \in \mathbb{R}^{B \times L \times D}$, where L is the token sequence length. First, a linear projection is applied to the feature of each token to map the dimension from D to C , yielding $\hat{H} \in \mathbb{R}^{B \times L \times C}$. Then, $\hat{H}$ is reshaped into a two-dimensional feature map, and bilinear interpolation is adopted for upsampling when necessary so that its spatial resolution is consistent with that of the visual features, thereby obtaining $\hat{y}_2 \in \mathbb{R}^{B \times C \times H \times W}$.

For global semantic guidance, the sentence-level CLIP embedding is projected and broadcast to the spatial resolution of the visual latent features. For token-level semantic guidance, the token sequence is first projected to the latent channel dimension and then reshaped or interpolated to match the spatial size of the visual features. Therefore, all semantic and visual features are explicitly aligned to the same spatial resolution before channel-wise concatenation.

The aligned semantic feature $\hat{y}_2$ is then concatenated with $\hat{y}_1$ and $\hat{y}_3$ along the channel dimension:

$$F_{cat} = [\hat{y}_1; \hat{y}_2; \hat{y}_3] \in \mathbb{R}^{B \times 3C \times H \times W} \quad (9)$$

Since concatenation merely completes information aggregation without modeling cross-modal and cross-scale channel interactions, and the increase of channel number from C to $3C$ introduces redundancy and additional computational overhead, we further introduce a fusion module to jointly model F_{cat} and compress the channels, yielding the fused feature:

$$y = F_{fusion} = \phi \left(\text{Conv}_{3 \times 3}^{(3C \rightarrow N)}(F_{cat}) \right) \in \mathbb{R}^{B \times N \times H \times W} \quad (10)$$

where N denotes the number of channels in the final latent features, and $\phi(\cdot)$ represents the normalization and nonlinear activation following the convolution. After the above fusion and compression under feature-wise linear modulation, a compact and semantically enhanced latent variable y is obtained for subsequent hyperprior analysis and entropy coding.

From the perspective of latent feature distribution, the proposed semantic modulation does not activate all latent channels uniformly. Instead, semantic guidance redistributes the response energy across channels and spatial regions, shifting it from background areas, smooth regions, and semantically weakly related channels toward object contours, salient regions, and structurally meaningful details. As a result, the latent representation does not simply become denser; rather, it shows higher energy concentration and stronger structured sparsity. This property enables the entropy model to allocate bits more effectively to visually and semantically important components.

3.3 Hyperprior Entropy Model and Interaction with Text-Semantic Features

After text-semantic modulation and multi-scale feature fusion, the compact latent representation y contains both visual structural information and high-level semantic guidance. To accurately estimate the coding probability of this semantic-enhanced latent representation, we employ a hyperprior entropy model together with a channel-wise autoregressive entropy model. The hyper-analysis network h_a maps the main latent representation y into the hyper-latent variable z , and the quantized hyper-latent $\hat{z}$ is transmitted as an auxiliary bitstream. At the decoder side, the hyper-synthesis network h_s transforms $\hat{z}$ into the hyperprior

feature ψ , which provides side information for estimating the probability distribution of the quantized main latent variable $\hat{y}$:

$$z = h_a(y), \quad \hat{z} = Q(z), \quad \psi = h_s(\hat{z}). \quad (11)$$

Here, ψ represents the decoded hyperprior information, which reflects the spatial dependency, channel-wise correlation, and local scale variation of the semantic-enhanced latent representation. Instead of directly entropy coding $\hat{y}$ using a fixed prior, the proposed entropy model estimates the probability of $\hat{y}$ conditioned on both the hyperprior information and channel-wise context.

To further exploit the dependency among latent channels, we adopt a channel-wise autoregressive entropy model, namely ChARM. The quantized main latent variable $\hat{y}$ is divided into G channel groups along the channel dimension:

$$\hat{y} = [\hat{y}^{(1)}, \hat{y}^{(2)}, \dots, \hat{y}^{(G)}]. \quad (12)$$

In our implementation, $G = 4$, and the $N = 128$ channels are evenly divided into four groups, with each group containing 32 channels. The first channel group is encoded only conditioned on the hyperprior feature ψ , while each subsequent group is encoded conditioned on both ψ and the previously decoded channel groups. Therefore, the conditional probability of $\hat{y}$ can be formulated as

$$p(\hat{y}|\hat{z}) = \prod_{g=1}^G p(\hat{y}^{(g)}|\psi, \hat{y}^{(<g)}), \quad (13)$$

where $\hat{y}^{(<g)} = \hat{y}^{(1)}, \dots, \hat{y}^{(g-1)}$ denotes the previously decoded channel groups. For the first group, $\hat{y}^{(<1)}$ is empty, and the probability estimation relies only on the hyperprior feature ψ . For the following groups, the already decoded channel groups provide additional context information, which helps reduce the uncertainty of probability estimation. Compared with fully serial spatial autoregressive models, this channel-wise grouping strategy achieves a better balance between entropy modeling accuracy and decoding efficiency.

For the g -th channel group, the context feature is generated from the previously decoded groups:

$$c^{(g)} = \begin{cases} 0, & g = 1, \\ f_{\text{ctx}}^{(g)}(\hat{y}^{(<g)}), & g > 1. \end{cases} \quad (14)$$

where $f_{\text{ctx}}^{(g)}(\cdot)$ denotes the channel-wise context transform. The hyperprior feature ψ and the channel context feature $c^{(g)}$ are then fused and fed into the entropy-parameter network:

$$\Theta^{(g)} = f_{\text{ep}}^{(g)}([\psi, c^{(g)}]). \quad (15)$$

The entropy-parameter network predicts the parameters of a Gaussian mixture model for each channel group. In this paper, a three-component Gaussian mixture model is adopted, i.e., $K = 3$. For the g -th channel group, the predicted parameter set is written as

$$\Theta^{(g)} = \left\{ \omega_{i,k}^{(g)}, \mu_{i,k}^{(g)}, s_{i,k}^{(g)} \right\}_{k=1}^K, \quad K = 3. \quad (16)$$

where $\omega_{i,k}^{(g)}$, $\mu_{i,k}^{(g)}$, and $s_{i,k}^{(g)}$ denote the mixture logit, mean, and pre-activation scale of the k -th Gaussian component for the i -th latent element, respectively. The mixture weight and positive scale parameter are obtained by

$$\pi_{i,k}^{(g)} = \frac{\exp(\omega_{i,k}^{(g)})}{\sum_{j=1}^K \exp(\omega_{i,j}^{(g)})}, \quad \sigma_{i,k}^{(g)} = \text{softplus}(s_{i,k}^{(g)}) + \epsilon. \quad (17)$$

where ϵ is a small constant used for numerical stability. Accordingly, the conditional probability of the g -th quantized channel group is modeled as

$$p(\hat{y}^{(g)} | \psi, \hat{y}^{(<g)}) = \prod_i \left[\sum_{k=1}^K \pi_{i,k}^{(g)} \mathcal{N}(\hat{y}_i^{(g)}; \mu_{i,k}^{(g)}, (\sigma_{i,k}^{(g)})^2) \right] * \mathcal{U}(-0.5, 0.5), \quad (18)$$

where $\mathcal{U}(-0.5, 0.5)$ is used to model the uncertainty introduced by quantization. The estimated probability distribution is then used for arithmetic coding of the main latent variable. The total bitrate term consists of the coding cost of the main latent variable and the hyper-latent variable:

$$R = \mathbb{E} \left[- \sum_{g=1}^G \log_2 p(\hat{y}^{(g)} | \psi, \hat{y}^{(<g)}) \right] + \mathbb{E} [-\log_2 p_z(\hat{z})]. \quad (19)$$

The interaction between the text-semantic hidden representation and the hyperprior entropy model is achieved through an implicit statistical pathway. The text-semantic representation is not transmitted as an independent textual bitstream. Instead, the BLIP-generated description is encoded by the CLIP text encoder into a semantic embedding, which is further used by the FiLM generator to produce channel-wise scaling and shifting parameters. These parameters modulate the shallow and deep visual latent features before entropy coding. As a result, the fused latent representation y already contains text-semantic guidance before being fed into the hyper-analysis network h_a .

3.4 Bitrate Control Combining Content Awareness and Historical Feedback

Through the text-semantics-guided feature-wise linear modulation fusion introduced in the previous section, we obtain the semantically enhanced feature $y \in \mathbb{R}^{B \times N \times H \times W}$, which simultaneously encodes the local structural information of the image and the textual semantic information along the channel dimension. To achieve more stable and fine-grained rate control under the target bitrate constraint, we further design a rate control module that integrates content awareness and historical feedback based on this representation.

First, a channel attention module is employed to extract the content-aware prior from the semantically enhanced feature y . Global average pooling is performed on y along the spatial dimensions, followed by a lightweight mapping to obtain the channel response vector:

$$c = \text{CA}(y) \in \mathbb{R}^{B \times N} \quad (20)$$

where c_n represents the response intensity of the n -th channel to the current input. Since y has already been modulated by text semantics, c_n not only reflects the sensitivity of the corresponding channel to local texture complexity, but also characterizes its importance to semantically critical content. Based on this content-aware representation, an initial scaling factor $\alpha_0 = \psi(c)$ is further generated through a mapping function. As the initial control variable driven by content awareness, α_0 provides an adaptive initial rate operating point

for different images, enabling coarse-grained cross-content compensation and reducing the search range of subsequent feedback adjustment.

Next, to further approach the target bitrate, a historical feedback mechanism is introduced to dynamically refine the initial control variable. Let the target bitrate be r^* , and let the estimated batch bitrate obtained in the t -th forward pass be $\hat{r}_t$. We compute it using the conditional probability $p_0(\hat{y}_i | \mu_i, \sigma_i)$ predicted by the entropy model:

$$\hat{r}_t = -\frac{1}{BHW} \sum_i \log_2 p_0(\hat{y}_i | \mu_i, \sigma_i) \quad (21)$$

The historical bitrate is then obtained through exponential smoothing:

$$\bar{r}_t = \beta \bar{r}_{t-1} + (1 - \beta) \hat{r}_t, \quad 0 < \beta < 1 \quad (22)$$

The bitrate deviation is defined as:

$$\Delta_t = \bar{r}_t - r^* \quad (23)$$

when $\Delta_t < 0$, it indicates under-coding, whereas $\Delta_t > 0$ indicates over-coding. To enable the system to adaptively adjust to both under-coding and over-coding conditions, we employ multiplicative exponential feedback to update the scaling factor α and the rate-distortion trade-off multiplier λ , respectively:

$$\alpha_{t+1} = \Pi_{[\alpha_{\min}, \alpha_{\max}]} (\alpha_t \odot \exp(-k_\alpha \Delta_t)), \quad \lambda_{t+1} = \Pi_{[\lambda_{\min}, \lambda_{\max}]} (\lambda_t \cdot \exp(-k_\lambda \Delta_t)) \quad (24)$$

Among them, $k_\alpha > 0$ and $k_\lambda > 0$ are feedback intensity hyperparameters, and the operator $\Pi_{[a,b]}(\cdot)$ denotes the projection constraint onto the interval $[a, b]$, which is used to limit the parameter range and improve numerical stability. Its element-wise form is given by:

$$\Pi_{[a,b]}(x) = \min(\max(x, a), b) \quad (25)$$

when the bitrate exceeds the target, i.e., $\Delta_t > 0$, $\exp(-k_\alpha \Delta_t) < 1$ causes α to shrink, thereby reducing the amplitude of the latent representation before quantization. Meanwhile, $\exp(-k_\lambda \Delta_t) < 1$ decreases λ , which relatively strengthens the influence of the bitrate term R in the objective $\mathcal{L} = R + \lambda D$ and shifts the optimization toward reducing bitrate consumption. Conversely, when the bitrate falls below the target, i.e., $\Delta_t < 0$, α is appropriately relaxed and λ is increased, allowing the model to place more emphasis on reconstruction fidelity and avoid performance degradation caused by excessive compression. To further dampen short-term fluctuations, we apply EMA smoothing to the updated scaling factor:

$$\bar{\alpha}_{t+1} = \eta \alpha_{t+1} + (1 - \eta) \bar{\alpha}_t, \quad 0 < \eta < 1 \quad (26)$$

The obtained $\bar{\alpha}_{t+1}$ is used as the pre-quantization scaling factor and applied to the latent features:

$$\tilde{y} = \bar{\alpha}_{t+1} \odot y \quad (27)$$

To further clarify the quantization process, the proposed framework adopts a differentiable quantization relaxation during training. After the adaptive scaling factor is applied to the semantically enhanced latent representation, the scaled latent feature $\tilde{y}$ is quantized before entropy coding. Since the standard rounding operation is non-differentiable, we follow the common practice in learned image compression and replace hard rounding with additive uniform noise during training. Therefore, the relaxed latent representation

can be formulated as $\tilde{y} = \hat{y} + u$, where $u \sim \mathcal{U}(-0.5, 0.5)$. This relaxation allows gradients to be propagated through the analysis transform, the text-semantic modulation module, the entropy model, and the bitrate controller. During inference, the relaxed quantization is replaced by deterministic rounding, i.e., $\hat{y} = \text{round}(\tilde{y})$, followed by arithmetic encoding and decoding.

It is worth noting that the proposed method does not rely on an additional independent network to compensate for post-quantization reconstruction offsets. Rather, the quantization-reconstruction behavior is implicitly controlled through the adaptive latent scaling factor, the entropy-model distribution parameters, and the dynamically updated Lagrange multiplier. Consequently, the proposed controller affects both the quantization strength and the probability estimation of latent variables, while ensuring end-to-end differentiability during training and consistency with deterministic hard quantization during inference.

This process adjusts the amplitude of the latent features to alter the statistical distribution of the encoded latent variables, and further affects the entropy model's estimation of conditional probability distributions; meanwhile, λ_{t+1} , as the global rate-distortion weighting parameter, participates in the optimization objective:

$$\mathcal{L}_t = \mathbb{E}[\hat{r}] + \lambda_t \mathbb{E}[D(x, \hat{x})], \quad (28)$$

$D(x, \hat{x})$ serves as a metric for mean squared error (MSE) distortion. To avoid over-suppression of semantically important channels, we apply a more conservative lower bound $\alpha_{\min}^{\text{sem}} > \alpha_{\min}$ to high-attention channels to protect their dynamic range, while low-importance channels can be scaled more aggressively.

The threshold for distinguishing high-attention and low-attention channels is not manually fixed for specific image categories. Instead, it is adaptively determined according to the attention response distribution of each input image. Given the channel attention vector a , we define an adaptive threshold τ as the mean response plus a scaled standard deviation, i.e., $\tau = \text{mean}(a) + \kappa \text{std}(a)$, where κ is a control coefficient that determines the selectivity of important channels. Channels with attention responses higher than τ are regarded as semantically important channels and are assigned a more conservative lower bound $\alpha_{\min}^{\text{sem}}$ to avoid excessive suppression, while the remaining channels use the normal lower bound $\alpha_{\min}$ and can be more aggressively compressed. In practice, τ can also be clipped into a stable range to avoid extremely large or small thresholds.

This adaptive threshold design enables the bitrate controller to adjust to different image contents without requiring category-specific manual settings. For portrait images, channels related to facial contours, eyes, and body structures usually obtain higher attention responses and are therefore protected more conservatively. For landscape images, important structural regions such as object boundaries, sky-ground transitions, or salient natural textures can be adaptively preserved. In contrast, high-frequency noise or random texture responses with weak semantic relevance tend to produce lower attention scores and are more strongly compressed. Therefore, the proposed thresholding strategy can adapt to different image categories while suppressing semantically unimportant high-frequency bitrate consumption.

For high-frequency regions, the controller does not allocate bits solely according to local texture magnitude. Since the content-aware prior is extracted from the semantically modulated latent representation, channels with high semantic relevance are preserved more conservatively, whereas visually noisy but semantically less relevant high-frequency responses can be suppressed more strongly. Therefore, high-frequency patterns with weak semantic relevance are less likely to consume excessive bits. The historical feedback branch further suppresses systematic over-allocation by correcting the deviation between the actual output bitrate and the target bitrate.

This design enables the text-conditioned latent feature y to determine which regions are important, the channel attention module to provide content-aware amplitude priors, and the historical bitrate feedback to

indicate whether the model meets the target bitrate. The two control variables, (λ, α) , are jointly incorporated into the quantization and hyperprior entropy model, thereby realizing a bitrate control mechanism that integrates content awareness and historical feedback and can stably approach the target bitrate across different images and target bitrate settings.

3.5 Loss Function

In this paper, the overall network is trained using a joint rate-distortion optimization objective. The loss function consists of two components: the bitrate term R and the distortion term D . Specifically, R is used to constrain the coding cost of the latent variables fused with semantic information and their hyperprior information, while D is used to measure the reconstruction error between the original image x and the reconstructed image $\hat{x}$. The mean squared error is adopted as the distortion measure, and the Lagrange multiplier λ is used to balance compression rate and reconstruction quality. The overall objective function is defined as

$$\mathcal{L} = R + \lambda D \quad (29)$$

$$R = \mathbb{E}[-\log_2 p_{y|\hat{z}}(\hat{y} | \hat{z})] + \mathbb{E}[-\log_2 p_{\hat{z}}(\hat{z})] \quad (30)$$

$$D = \mathbb{E}_{x \sim P_x} [\|x - \hat{x}\|_2^2] \quad (31)$$

where λ represents the Lagrangian weight, which is used to balance the trade-off between bitrate and distortion. A larger value of λ prioritizes reconstruction quality, while a smaller value of λ favors reducing the bitrate.

In our implementation, the Lagrange multiplier λ in the rate-distortion objective is not fixed to a single value. To cover multiple bitrate targets during training, the base multiplier λ_0 is uniformly sampled from $\{0.0016, 0.0032, 0.0075, 0.015\}$ for each mini-batch. These values provide different initial rate-distortion trade-off points from low to high bitrate ranges. The sampled base multiplier is further adjusted by the proposed bitrate controller according to the historical bitrate deviation between the actual output bitrate and the target bitrate. When the actual bitrate is higher than the target bitrate, the effective multiplier is decreased to relatively strengthen the bitrate constraint and encourage stronger compression. Conversely, when the actual bitrate is lower than the target bitrate, the effective multiplier is increased to place more emphasis on reconstruction fidelity and avoid over-compression. Therefore, the effective multiplier used in optimization becomes an adaptive value rather than a manually fixed constant.

With this design, the proposed framework can learn compression behaviors corresponding to different bitrate constraints within a single unified model. During inference, different target bitrates can be achieved by specifying the target bitrate and allowing the controller to update the quantization scaling factor and the effective Lagrange multiplier accordingly. Therefore, the proposed method supports variable-rate compression within the trained bitrate range, rather than requiring a separately trained model for each target bitrate.

4 Experiments

4.1 Experimental Design

The training phase was conducted on the Microsoft Common Objects in Context (MS-COCO) dataset. During training, we first randomly cropped all images to a resolution of 256×256 and applied strategies such

as random horizontal flipping and mild color perturbations to enhance the model's generalization ability. For text semantics, we use a pre-trained BLIP model to generate several descriptive sentences for each image. In each iteration, we randomly select one sentence and feed it into a pre-trained CLIP text encoder to extract a 512-dimensional text semantic embedding. We then use this semantic embedding as conditional information input to a feature-dimensional linear modulation module to perform channel-level modulation on multi-scale visual latent features.

During the testing stage, we evaluate the proposed method on Kodak24, Tecnick, the Challenge on Learned Image Compression (CLIC) validation set, Urban100, and the DIVERse 2K-resolution image dataset (DIV2K). Kodak24 is used as a standard benchmark for learned image compression, Tecnick provides high-resolution natural images, the CLIC validation set is included to evaluate cross-dataset generalization under more diverse image contents and resolutions, Urban100 is used to further assess the robustness of the proposed method on dense edges, repeated structures, and high-frequency textures, and DIV2K is used to evaluate the generalization ability on high-resolution natural images with diverse contents. All test images are evaluated at their original resolutions without resizing or center cropping, ensuring the practical relevance of the evaluation results. The textual descriptions are also automatically generated by BLIP and encoded into semantic features by the CLIP text encoder, so as to maintain consistency between the training and testing procedures.

The encoder and decoder adopt symmetric convolutional neural networks, with the base channel number N set to 128, consistent with the learned image compression framework of Ballé2018 et al. [31] to ensure fair comparison. The channel number of the hyperprior network is set to 192, and the entropy model adopts a three-component Gaussian mixture model ($K = 3$). The model is trained for 150 epochs using the Adam optimizer with a batch size of 8. The learning rate is maintained at 1×10^{-4} for the first 100 epochs and then reduced to 1×10^{-5} for the remaining 50 epochs for fine-tuning. To cover multiple target bitrates from low to high, we predefine four base Lagrange multipliers, $\{0.0016, 0.0032, 0.0075, 0.015\}$. During training, one base multiplier is uniformly sampled at random from this set for each mini-batch and denoted as λ_0 . Under the historical bitrate feedback mechanism, λ_0 and the initial scaling factor α_0 of the content-aware module in Section 3.4 are dynamically updated to (λ_t, α_t) , enabling the model to learn compression strategies that adapt to different target bitrate constraints. All experiments are conducted on a server equipped with an RTX 4090 GPU with 48 GB of memory.

4.2 Main Experimental Results

We use three objective quality metrics to comprehensively evaluate the quality of the reconstruction: (1) Peak Signal-to-Noise Ratio (PSNR), which measures pixel-level fidelity and is expressed in dB; (2) Multi-scale Structural Similarity (MS-SSIM), which assesses the degree to which structural information is preserved; and (3) Perceptual Loss (LPIPS), calculated based on AlexNet features, which reflects human visual perception of quality, where a lower LPIPS value indicates better perceptual quality. We compare our method with several representative baselines, including the traditional coding standards JPEG2000 [2] and JPEG [1], as well as recent learning-based encoders such as FTIC (ICLR'24) [32], Fu2023 [7], ELIC (CVPR'22) [33], Cheng2020 (CVPR'20) [34], Minnen2018 (NeurIPS'18) [16], and Ballé2018 (ICLR'18) [31].

Fig. 2a shows a comparison of PSNR distortion curves across the dataset. As can be clearly seen from the figure, our proposed method maintains competitive performance across the entire bitrate range. Comparison results with several representative baseline methods indicate that the text-guided compression framework improves reconstruction quality while maintaining encoding efficiency.

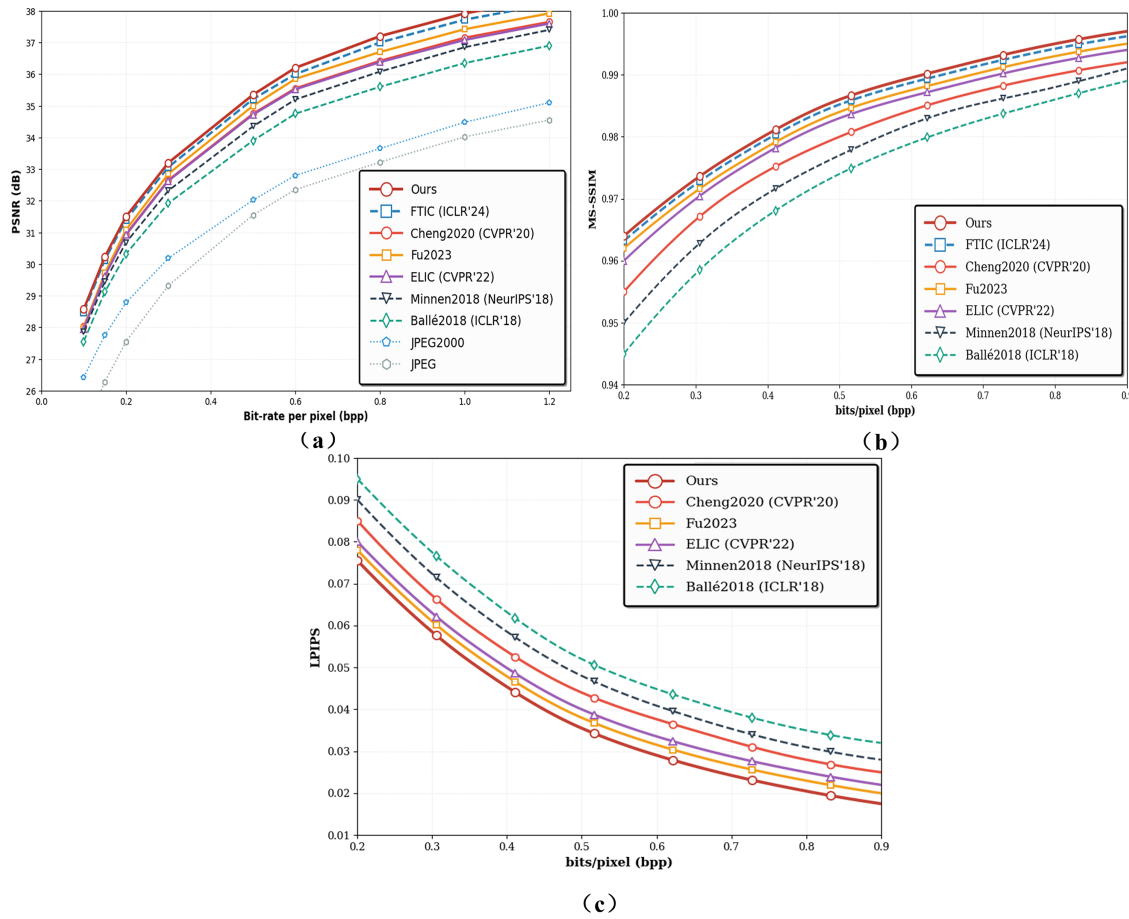


Figure 2: Comparison of average PSNR, MS-SSIM, and LPIPS on the Kodak and Tecnick datasets. (a) PSNR; (b) MS-SSIM; (c) LPIPS.

In the low-to-medium bitrate range of 0.2–0.6 bpp, our method demonstrates a consistent advantage in PSNR compared to recent learning-based compression methods such as Fu2023 [7], ELIC [33], and FTIC (ICLR'24) [32]. FTIC (ICLR'24) [32] enhances the modeling capabilities of frequency components in different directions and latent variable dependencies through a frequency-aware transform structure and an enhanced entropy model; its overall rate-distortion performance is very close to that of our method. However, relying solely on frequency-domain feature modeling and statistical redundancy compression still makes it difficult to adequately capture differences in semantic importance across image content. In contrast, the proposed method injects text-semantic conditions into both shallow local texture features and deep global structural representations, and further combines a dual-channel rate control strategy integrating content awareness and historical feedback to jointly regulate the quantization strength and rate-distortion optimization process, thereby enabling the limited bit budget to be more effectively allocated to key object contours, salient textures, and structural regions. Therefore, under the same bitrate constraint, the proposed method achieves more effective semantic preservation and detail reconstruction, thereby delivering consistently superior PSNR performance. Fu2023 [7] adopts a relatively complex transform network architecture and demonstrates strong rate-distortion competitiveness; however, its curve remains consistently lower than that of the proposed method across the entire tested bitrate range, indicating that the proposed method can achieve better reconstruction performance with lower complexity. ELIC [33] achieves performance close to ours in

the low-bitrate range, but the performance gap gradually widens as the bitrate increases. This phenomenon indicates that, at higher bitrates, the advantage of text guidance becomes more evident, because the model has a larger bitrate budget to implement the semantic-aware bit allocation strategy more finely.

Fig. 2b presents the MS-SSIM curves of different compression methods as a function of bitrate, where this metric reflects the preservation quality of structural information. It can be observed that the proposed method consistently outperforms all comparison methods across the entire tested bitrate range. In the critical bitrate range of 0.3–0.6 bpp, our method still maintains an advantage of approximately 0.003 over FTIC (ICLR'24) [32], indicating that the proposed method has a stronger capability for structural preservation under low-to-medium bitrate conditions. This advantage mainly benefits from the text-semantics-guided multi-scale feature modulation and the bitrate control mechanism that integrates content awareness with historical feedback. The former enhances the model's ability to represent key semantic regions and important structural details, while the latter enables the limited bit budget to be more reasonably allocated to edges, textures, and salient object regions, thereby improving the overall structural reconstruction quality while suppressing redundant costs in smooth regions. Therefore, compared with methods such as FTIC (ICLR'24) [32] and ELIC [33], the proposed method exhibits a more stable advantage in terms of the MS-SSIM metric.

Fig. 2c shows the comparison results of different compression methods in terms of the LPIPS metric. Our method achieves the lowest LPIPS values across the entire bitrate range, demonstrating superior perceptual quality. At 0.3 bpp, our LPIPS is 0.059, which is 4% lower than the 0.061 achieved by Cheng2020 [34] and 24% lower than the 0.078 achieved by Ballé2018 [31]. These results demonstrate that text-semantic guidance effectively improves the perceptual quality of reconstructed images. For text prompts describing key semantic entities in an image, the feature-dimensional linear modulation module generates channel modulation parameters γ and offsets β that act on visual latent features at different levels. This enhances the representational capacity of regions relevant to the target semantics while suppressing redundant responses in low-relevance background regions, thereby achieving more rational bit allocation under limited bitrate constraints.

To further evaluate the cross-dataset generalization capability of the proposed method, additional experiments are conducted on the Challenge on Learned Image Compression (CLIC) validation set, Urban100, and the DIVERse 2K-resolution image dataset (DIV2K). The proposed method is compared with several recent representative learned image compression methods, including Rethinking-LV (WACV'26) [35], Text-Adaptive Compression (TACO) [17], S2LIC: Learned Image Compression with the SwinV2 Block, Adaptive Channel-wise and Global-inter Attention Context (S2LIC) [36], Hierarchical Progressive Context Modeling Base (HPCM-Base) [37], CMamba (2025) [38], WACV'25 [39], and Quantization-Aware ResNet VAE++ (QARV++) [40]. These three datasets provide complementary evaluations for examining the robustness of the proposed text-semantic guidance mechanism and content-aware bitrate control mechanism under more diverse content distributions. Specifically, CLIC contains diverse natural images with varying resolutions and contents, Urban100 emphasizes dense structural details, repeated patterns, and high-frequency textures, while DIV2K further evaluates the generalization ability on high-resolution natural images.

As shown in Fig. 3a, the proposed method achieves the best PSNR performance on the CLIC validation set across most of the tested bitrate range. In particular, the advantage is more evident in the low-to-medium bitrate range, where bit allocation is more sensitive to semantic importance and local texture complexity. This result indicates that the proposed text-semantic guidance is not limited to the Kodak24 and Tecnick benchmarks, but can also generalize to more diverse natural image contents. By injecting BLIP- and CLIP-based semantic representations into multi-scale latent features, the model can better preserve semantically

important structures under constrained bit budgets. Fig. 3b reports the PSNR comparison on Urban100. Urban100 contains more repeated patterns, dense edges, and high-frequency structural details than natural image datasets, making it more challenging for learned image compression models. The proposed method still achieves competitive performance and remains among the top-performing methods across the bitrate range. Although QARV++ shows slightly higher PSNR at several bitrate points, our method consistently outperforms most other baselines. Compared with TACO, the proposed method obtains higher PSNR over the tested bitrate range, indicating that the combination of text-semantic guidance and content-aware bitrate control provides more stable rate-distortion optimization in structure-rich scenarios. Fig. 3c further presents the results on DIV2K. Since DIV2K contains high-resolution natural images with diverse textures and image contents, it provides an additional evaluation of cross-dataset generalization. The proposed method maintains strong and stable PSNR performance on DIV2K and remains close to or above the strongest competing methods over most bitrate points. This demonstrates that the proposed framework can generalize not only to complex structural images such as Urban100, but also to high-resolution natural image distributions.

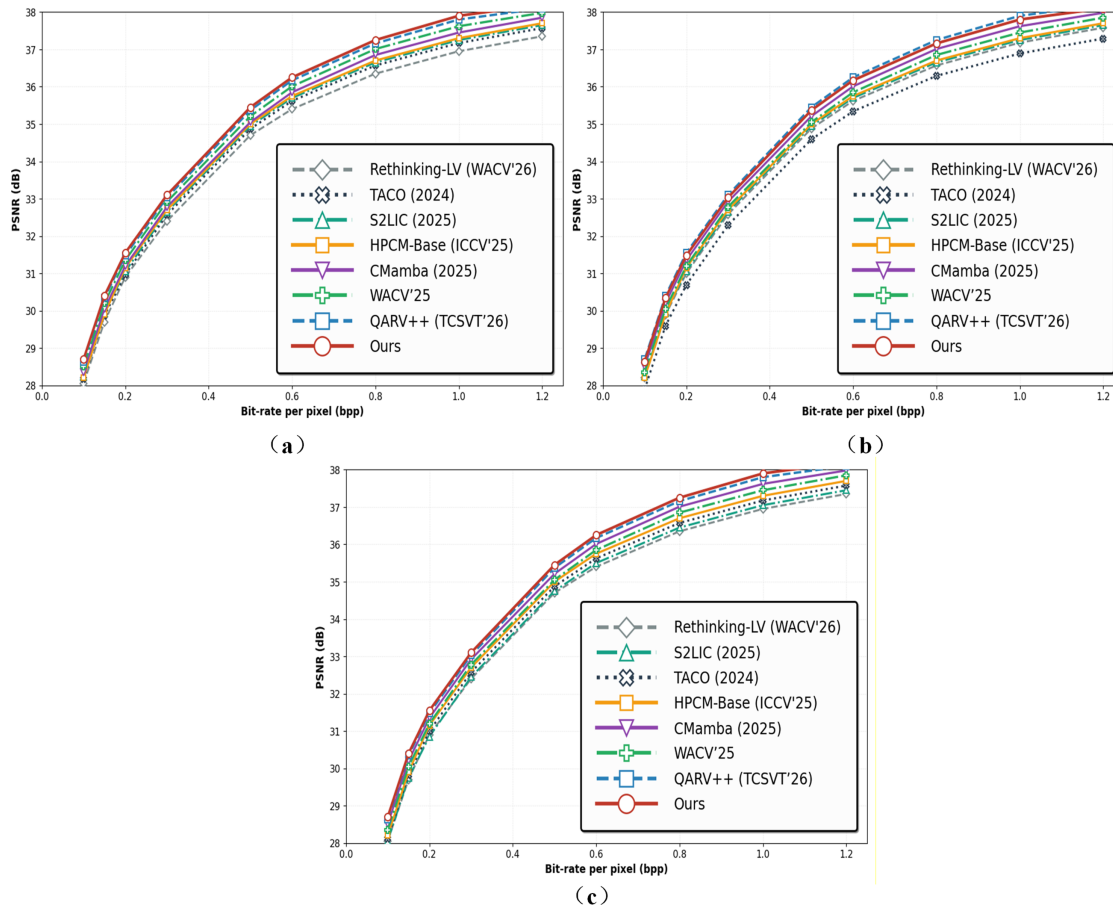


Figure 3: Cross-dataset PSNR performance comparison on the CLIC validation set, Urban100, and DIV2K. (a) CLIC validation set; (b) Urban100; (c) DIV2K.

The experimental results on CLIC, Urban100, and DIV2K validate the cross-dataset robustness of the proposed framework. The consistent performance on CLIC demonstrates its generalization capability on diverse natural images, the competitive results on Urban100 confirm its effectiveness in handling dense

structural details and high-frequency textures, and the results on DIV2K further verify its robustness on high-resolution natural images. These results indicate that the improvements brought by text-semantic guidance and content-aware bitrate control are not limited to the original Kodak24 and Tecnick evaluation settings.

4.3 Computational Complexity Analysis

We conducted a systematic analysis of the computational complexity of the model using four metrics: encoding time, decoding time, GPU used for inference, GPU memory usage, and forward FLOPs. These metrics provide a comprehensive reflection of the model's deployment costs from various perspectives, including runtime efficiency, resource consumption, and computational scale. As shown in Table 1, our method achieves BD-Rate reductions of 15.58% and 16.86% on the Kodak and Tecnick datasets, respectively, demonstrating rate-distortion performance superior to all comparison methods.

Table 1: Comparison of rate-distortion (R-D) performance and computational complexity of various learned image compression models on the Kodak and Tecnick datasets, evaluated on an NVIDIA RTX 4090 GPU. Lower Bjøntegaard Delta Rate (BD-Rate) values indicate better rate-distortion performance. “-” denotes that the result is unavailable.

Method	Enc (s)	Dec (s)	Mem (G)	FLOPs (G)	Params (M)	BD-Rate (PSNR)	
						Kodak	Tecnick
Ballé2018 [31]	0.073	0.067	–	244.44	12.40	0.00%	0.00%
Minnen2018 [16]	0.038	0.117	0.072	179.83	20.15	–1.92%	–
Cheng2020 [34]	0.041	0.153	0.127	258.53	28.60	–3.35%	–5.93%
ELIC [33]	0.237	0.120	0.373	332.42	33.29	–7.02%	–7.64%
FTIC [32]	0.125	0.242	–	–	70.97	–14.50%	–15.10%
Ours	0.326	0.151	0.708	394.20	66.10	–15.58%	–16.86%

From the perspective of computational complexity, our method requires 394.2 G FLOPs for forward computation, with 66.1 M parameters and an inference memory footprint of 0.708 G. Compared with methods such as Cheng2020 [34] and ELIC [33], our model shows increases in both parameter count and computational cost, indicating that the performance improvement is achieved on the basis of stronger feature modeling capability and higher computational investment. Despite the increased model complexity, our method still maintains reasonably competitive practical inference efficiency. Specifically, the decoding time is 0.151 s, which is significantly lower than that of the two autoregressive baselines, Minnen2018 [16] and Cheng2020 [34], and is also close to the 0.120 s decoding time of ELIC [33], indicating that the performance gain does not introduce excessive decoding latency. On the other hand, the encoding time of our method is 0.326 s, which is higher than that of some lightweight baselines, reflecting the additional encoder-side computational cost introduced by text-semantic extraction and adaptive bitrate control.

To further clarify the source of the additional encoding cost, we decompose the runtime of the proposed method into several major components, as shown in Table 2. This analysis is particularly important because the proposed framework introduces BLIP and CLIP for text-semantic guidance, which brings additional computation compared with pure visual learned compression models.

Table 2: Runtime decomposition of the proposed method with separate BLIP and CLIP timing.

Component	Time Per Image	Percentage	Location
BLIP caption generation	0.096 s	29.4%	Encoder-side only
CLIP text embedding	0.028 s	8.6%	Encoder-side only
Semantic extraction subtotal	0.124 s	38.0%	Encoder-side only
Visual analysis transform	0.073 s	22.4%	Encoder
FiLM modulation and feature fusion	0.019 s	5.8%	Encoder
Hyperprior and entropy coding	0.110 s	33.7%	Encoder
Total encoding time	0.326 s	100.0%	Encoder
Decoder synthesis transform	0.151 s	–	Decoder
Total decoding time	0.151 s	–	Decoder

Note: All runtime results are averaged over the Kodak24 dataset with batch size 1 on an NVIDIA RTX 4090 GPU. Warm-up iterations are performed before measurement, and GPU synchronization is used to obtain stable timing results. The percentage is calculated only with respect to the total encoding time. Therefore, the decoder-side percentage is not reported in this table. Boldface entries indicate subtotal or total runtime values.

Although the overall computational cost is higher than that of several existing learned compression baselines, the additional complexity is mainly concentrated on the encoder side. BLIP and CLIP are used to generate and encode semantic guidance only during encoding. The decoder does not require BLIP, CLIP, or any external text input. Therefore, the decoding latency remains comparable to several existing learned compression models.

As shown in Table 2, BLIP caption generation takes 0.096 s per image, and CLIP text embedding takes 0.028 s per image. Together, these two semantic extraction components account for 0.124 s, corresponding to 38.0% of the total encoding time. If the complete encoding–decoding pipeline is considered, the BLIP and CLIP components account for approximately 26.0% of the total runtime, calculated as $0.124 / (0.326 + 0.151)$. This confirms that the use of vision–language models introduces non-negligible encoder-side overhead.

The visual analysis transform and hyperprior entropy coding take 0.073 and 0.110 s, respectively, and remain the major compression-related components of the encoder. In contrast, FiLM modulation and feature fusion introduce only 0.019 s of additional latency, indicating that the feature-wise semantic modulation itself is relatively lightweight. Therefore, the increased complexity mainly comes from encoder-side semantic extraction and entropy coding rather than from the FiLM operation. The complexity analysis indicates that the proposed method achieves improved rate–distortion performance at the cost of additional encoder-side semantic extraction. This trade-off is acceptable for applications where reconstruction fidelity and semantic consistency are prioritized over strict real-time encoding.

In practical applications, the additional encoder-side cost is acceptable in several scenarios where images are encoded once but decoded, stored, transmitted, or analyzed multiple times. For example, in cloud-side image storage and archival compression, the encoding process can be performed offline or on server-side devices, while users mainly require high-quality reconstruction during later browsing, retrieval, or distribution. In bandwidth-constrained remote image transmission, such as UAV imaging, remote sensing image delivery, and edge-to-cloud visual monitoring, preserving object contours, salient regions, and semantically important structures is often more important than minimizing the one-time encoding latency. Similarly, in semantic-preserving visual communication and high-quality image sharing, slight increases in encoding time can be acceptable when they lead to improved structural fidelity and perceptual consistency.

at the same bitrate. Therefore, the proposed method is more suitable for quality-oriented or analysis-oriented compression scenarios, rather than strict real-time encoding applications such as ultra-low-latency video conferencing.

To further provide an indicative comparison with traditional codec baselines, we report the coding runtime of Versatile Video Coding (VVC) and High Efficiency Video Coding (HEVC) reference software in Table 3. Different from learned image compression models, traditional codecs do not involve neural inference, model parameters, FLOPs, or GPU memory usage. Therefore, we report encoding and decoding runtime instead of inference time. Since the Versatile Video Coding Test Model (VVC-VTM) and the High Efficiency Video Coding Test Model (HEVC-HM) are CPU-based reference implementations, while the proposed learned compression model is evaluated on an NVIDIA RTX 4090 GPU, this comparison is not intended as a strictly hardware-equivalent benchmark, but rather as an indicative coding-runtime reference.

Table 3: Indicative coding-runtime comparison with traditional VVC and HEVC reference codecs.

Methods	Platform	Enc. Time	Dec. Time	BD-Rate
VVC	CPU	57.4 s	0.09 s	0.0
HEVC	CPU	1.53 s	0.06 s	24.2%
Ours	RTX 4090 GPU	0.33 s	0.15 s	−4.9%

As shown in Table 3, VVC achieves stronger compression performance than HEVC, but it introduces substantially higher encoder-side runtime. Compared with HEVC-HM, VVC-VTM reduces the BD-Rate but increases the encoding time from 1.53 to 57.4 s per image. This indicates that the improved compression efficiency of VVC is obtained at the cost of much higher encoding complexity.

Compared with the reported VVC reference software, the proposed method achieves a lower BD-Rate of −4.9% and requires only 0.33 s for encoding per image. This suggests that the proposed model provides a favorable trade-off between compression performance and encoder-side runtime. However, the decoding time of our method is 0.15 s, which is higher than the reported HEVC and VVC decoding times. This is mainly because the proposed decoder relies on a neural synthesis transform, whereas HEVC/VVC decoding is performed by conventional codec pipelines. This comparison indicates that the proposed method achieves competitive compression performance with lower encoding time than the CPU-based VVC reference software under the reported setting. Considering the different implementation platforms, these results should be interpreted as an indicative coding-runtime comparison rather than a fully hardware-equivalent speed benchmark.

4.4 Image Quality Assessment

We further selected representative examples from the image dataset and conducted a subjective evaluation against Ballé2018 [31] at the same bitrate. The overall subjective quality comparison is shown in Fig. 4, and enlarged visual examples are further presented in Fig. 5. Fig. 5a shows a scene featuring a hat. At 0.07 bpp, the reconstructed image from Ballé2018 [31] exhibits noticeable blurring in the cloud patterns and the text on the hat brim, whereas our method preserves clear cloud textures and accurate text while maintaining the same bitrate. The textual description involves local entities with explicit semantic references, such as the logo on the cap brim. The feature-wise linear modulation module conditionally guides and selectively enhances the corresponding feature channels, enabling the high-frequency details associated with the text region to be better preserved in the low-bitrate representation, and thus improving its reconstruction quality and semantic discriminability. In addition, because the cloud region exhibits abrupt high-texture

complexity, Ballé2018 [31] produces block artifacts and ringing in this region. In contrast, our method adaptively adjusts the current scaling factor and Lagrange multiplier by tracking the deviation between the historical bitrate $\bar{r}_t$ and the target bitrate r^* , thereby avoiding detail loss caused by over-compression in the cloud region.

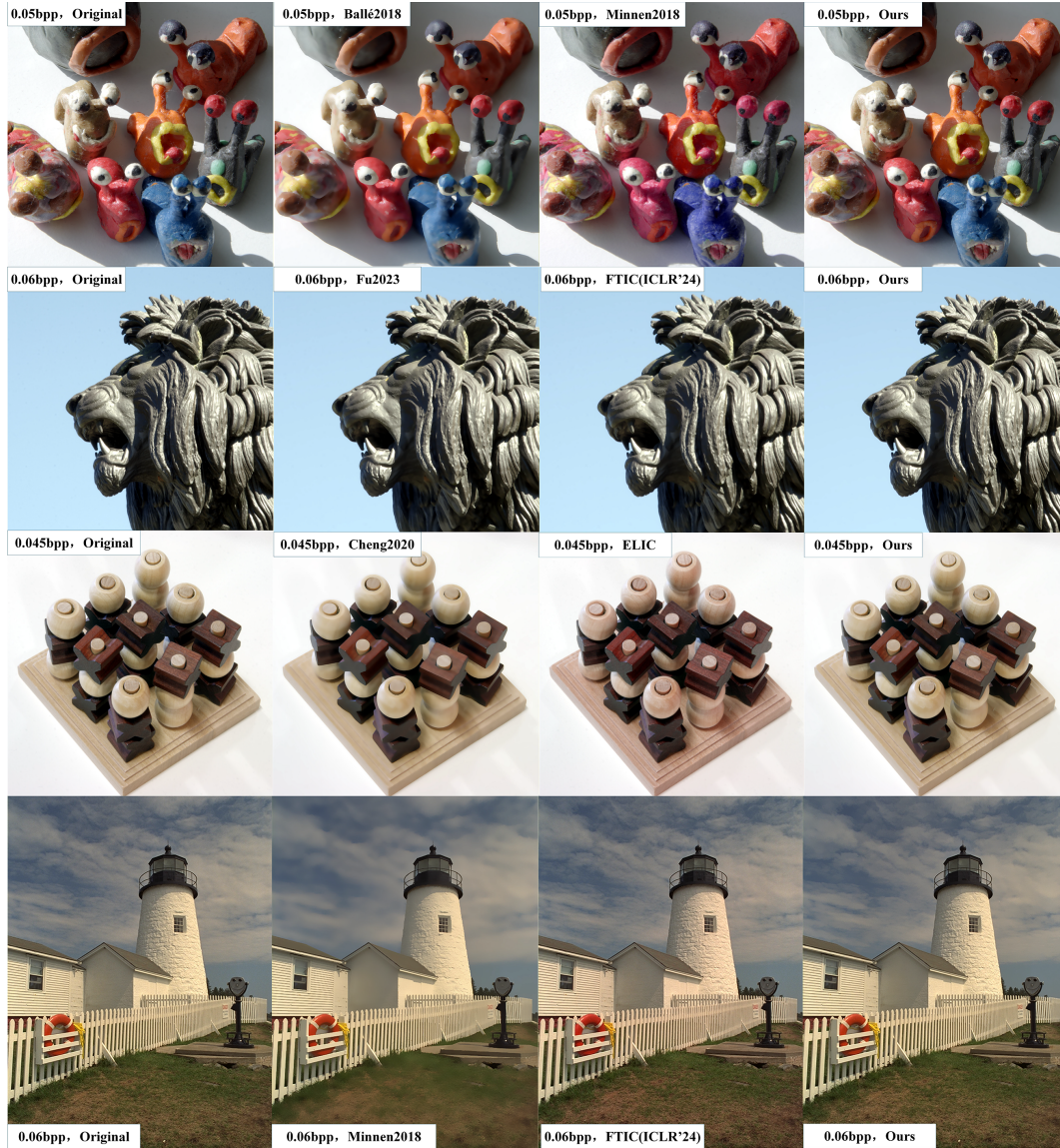


Figure 4: Subjective quality comparison on the Kodak and Tecnick datasets.

Fig. 5b shows the statue scene. At the same bitrate of 0.065 bpp, Ballé2018 [31] exhibits obvious smoothing artifacts around the ear and facial regions of the figure, and the overall image is also accompanied by slight blocking artifacts. In contrast, our method can better preserve the ear contour and high-frequency facial details, while recovering richer texture layers on the background stone wall.

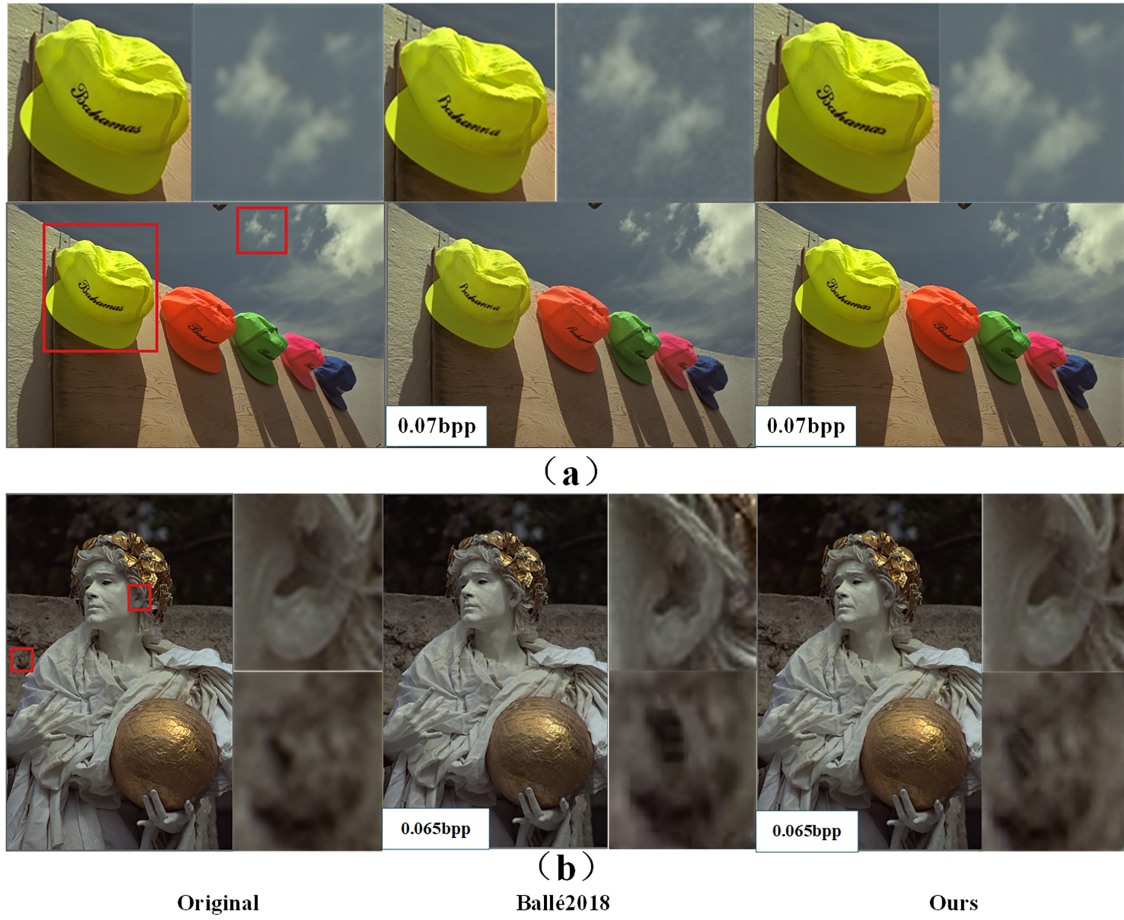


Figure 5: Visual examples of images generated after compression at extremely low bit rates: (a) is the hat brim image, and (b) is the statue image.

4.5 Ablation Experiment

This section first discusses the effect of the text-semantics-guided feature-wise linear modulation module on model performance, and then evaluates the impact of the bitrate control strategy integrating content awareness and historical feedback on the compression performance of the model.

We conduct an ablation study by removing the text branch and the feature-wise linear modulation module, while retaining the image encoders g_{a1} , g_{a3} , the bitrate control module, and the remaining coding structures as the baseline model. The ablation results are shown in Fig. 6. Across the tested bitrate range, removing the semantic guidance module consistently degrades the reconstruction quality. In the representative bitrate range of 0.3–0.6 bpp, the average PSNR drops by approximately 0.7–0.8 dB. Meanwhile, the LPIPS value also increases noticeably. At approximately 0.30 bpp, LPIPS increases from 0.059 for the full model to 0.074 for the baseline, corresponding to a relative deterioration of about 25.4%. Under the premise that the basic coding framework and bitrate control mechanism remain unchanged, these results indicate that the text branch and feature-wise linear modulation play a key role in selectively enhancing and preserving perceptually and structurally important features, thereby verifying the value of semantic guidance for overall compression performance.

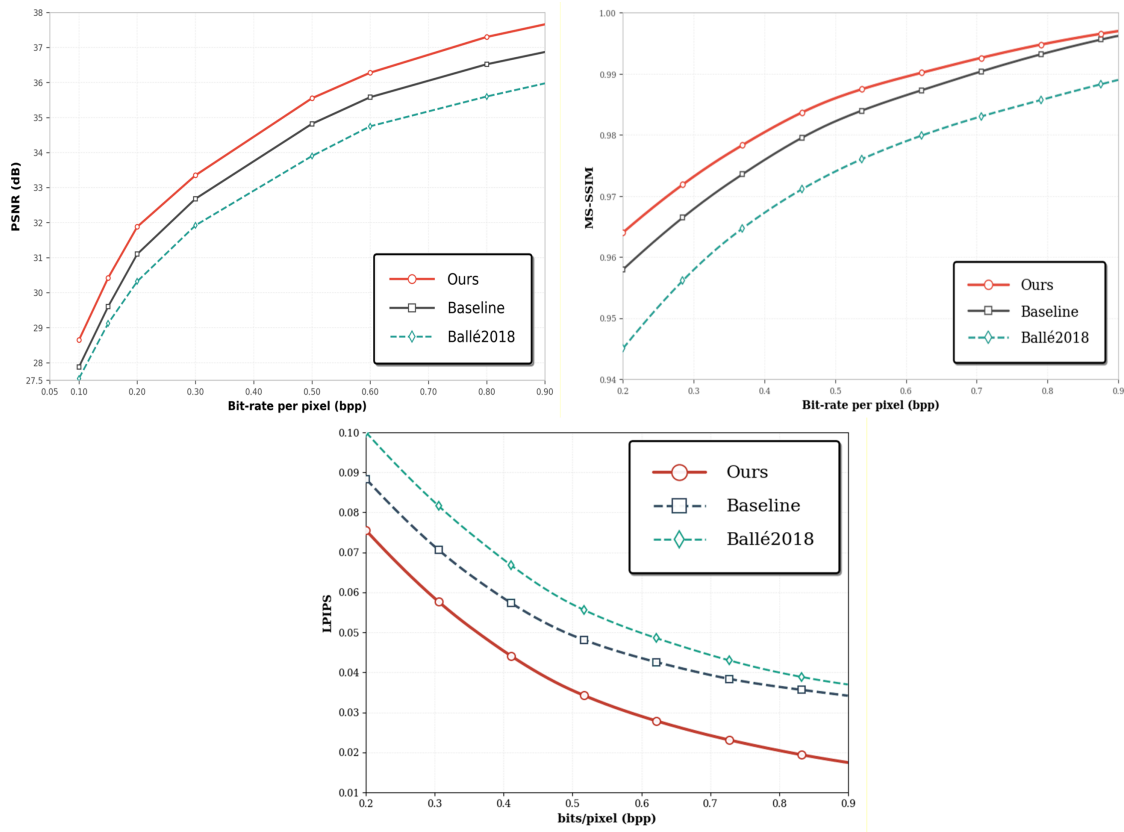


Figure 6: Comparison of the baseline model without the text branch and feature-wise linear modulation module, the full model, and Ballé2018 in terms of PSNR and LPIPS.

The results show that CLIP with a fixed generic prompt provides limited but consistent semantic gain, while BLIP-generated image-specific captions lead to more noticeable improvement. Compared with the baseline without text guidance, the CLIP w/fixed prompt variant improves PSNR from 32.6 to 32.84 dB, indicating that even generic textual priors can provide auxiliary semantic information. When BLIP-generated image-specific captions are used without the CLIP text encoder, the PSNR further increases to 33.02 dB, suggesting that image-dependent textual descriptions are more effective for guiding latent feature modulation than generic textual prompts. The BLIP + CLIP variant achieves 33.18 dB PSNR and lower LPIPS, demonstrating that BLIP and CLIP provide complementary semantic information: BLIP supplies image-specific descriptive cues, while CLIP maps these cues into a more discriminative semantic embedding space.

Compared with the BLIP + CLIP variant without bitrate control, the full model further improves PSNR from 33.18 to 33.38 dB while reducing the actual bitrate from 0.312 to 0.301 bpp. This indicates that the proposed bitrate control module does not improve reconstruction quality by simply increasing the bit budget. Instead, it improves the efficiency and accuracy of bit allocation under the target bitrate constraint. The full model achieves a PSNR gain of 0.75 dB over the baseline, together with consistent improvements in MS-SSIM and LPIPS, verifying the effectiveness of both text-semantic guidance and content-aware bitrate control. At approximately 0.30 bpp, the MS-SSIM increases from 0.965 for the baseline without text guidance to 0.974 for the full model, corresponding to an absolute gain of 0.009. Since MS-SSIM values are close to one, we further measure the reduction in structural dissimilarity using $1 - \text{MS-SSIM}$, which decreases from 0.035 to 0.026, corresponding to a relative reduction of approximately 25.7%. Moreover, the actual bitrate of the full

model is 0.301 bpp, which is almost identical to the 0.300 bpp of the baseline. Therefore, the improvement cannot be simply attributed to a higher bit budget, but is mainly due to BLIP- and CLIP-based semantic representations, feature-wise linear modulation, and content-aware bit allocation. These results are consistent with the component-level ablation results in Table 4 and further demonstrate that text-semantic guidance provides measurable structural-similarity improvement over the baseline without text guidance.

Table 4: Component-level ablation of text-semantic guidance and bitrate control at approximately 0.30 bpp on Kodak24 and Tecnick datasets.

Variant	BLIP Caption	CLIP Text Encoder	FiLM	Bitrate Control	PSNR	MS-SSIM	LPIPS	Actual bpp	Enc. Time
Baseline w/o text	×	×	×	✓	32.63	0.965	0.074	0.300	0.226 s
CLIP w/fixed prompt	×	✓	✓	✓	32.84	0.968	0.069	0.301	0.251 s
BLIP caption w/o CLIP	✓	×	✓	✓	33.02	0.970	0.066	0.301	0.294 s
BLIP + CLIP w/o bitrate control	✓	✓	✓	×	33.18	0.972	0.062	0.312	0.327 s
Full model	✓	✓	✓	✓	33.38	0.974	0.059	0.301	0.341 s

Note: In the “BLIP caption w/o CLIP” variant, the CLIP text encoder is replaced by a lightweight trainable text projection module with the same output dimension. In the “CLIP w/fixed prompt” variant, the image-specific BLIP caption is replaced by a generic prompt, such as “a natural image”. Similar trends can also be observed at other bitrate points.

As shown in Fig. 7, when only the content-aware branch or only the historical-feedback branch is retained, the overall rate-distortion performance of the model is inferior to that of the full model. In the medium-to-high bitrate range, the PSNR of the two models with only a single branch exhibits only a slight decrease. However, at low bitrates, PSNR deteriorates significantly by about 0.5 dB due to unreasonable bit allocation.

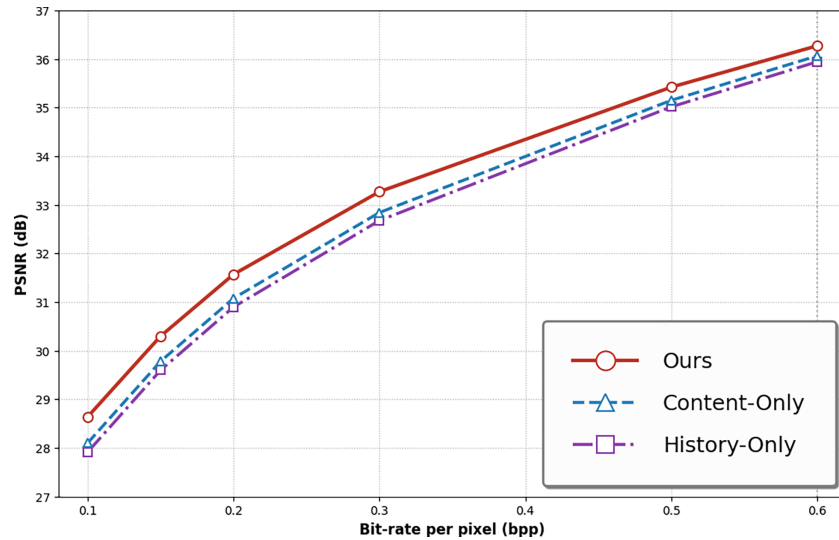


Figure 7: Comparison of the PSNR results of the ablation models that retain only the content-aware branch or the historical-feedback branch with those of the full model.

The reason for this phenomenon is that, at low bitrates, every bit is critical, and the cost of allocation errors is magnified. Using only the content-aware branch lacks the smoothing constraints provided by historical information, leading to control instability, increased sensitivity to noise in instantaneous rate

estimates, and a greater tendency to oscillate. When only the historical-feedback branch is retained, the model cannot adaptively adjust to the varying bitrate demands of different content, resulting in insufficient allocation for texture-rich regions and excessive allocation for smooth regions, which leads to severe quality non-uniformity. As the bitrate increases and a more sufficient bit budget becomes available, the encoder is better able to compensate to some extent for the limitations introduced by a single strategy. For example, although the historical-feedback strategy cannot precisely adapt to content variations, the ample bit budget is still sufficient to alleviate quality degradation in texture-rich regions. Similarly, the fluctuations in instantaneous estimation caused by the content-aware strategy can also be buffered by the sufficient bitrate, reducing the negative impact of bit-allocation errors.

Table 5 directly reveals the limitations of single-sided bitrate control. Under content-aware guidance alone, although the model can adaptively adjust according to the semantics and complexity of the current image, it lacks the ability to correct systematic bitrate bias. As a result, the AE reaches 0.0045 bpp, which is approximately $1.88\times$ that of the full model. Similarly, under historical-feedback guidance alone, the model relies mainly on the bitrate trajectory of previous samples while ignoring the content characteristics of the current image, leading to a larger AE of 0.0056 bpp and the highest cross-content standard deviation.

Table 5: Ablation results of the rate-control architecture at 0.1 bpp on the Kodak and Tecnick datasets.

Configuration	AE (bpp)	RE (%)	OS (%)	Oscillation Count	Cross-Content σ
Full model	0.0024	2.4	1.2	0.2	0.011
Content awareness only	0.0045	4.5	3.1	1.3	0.019
Historical feedback only	0.0056	5.6	4.2	1.7	0.026
Fixed-weight fusion	0.0033	3.3	2.0	0.8	0.015

Note: AE, absolute bitrate error; RE, relative bitrate error; OS, overshoot ratio.

The fixed-weight fusion strategy partially combines content awareness and historical feedback, but it cannot dynamically balance the two sources of information under different image contents and target bitrate constraints. Therefore, its control accuracy is still inferior to that of the full model. In contrast, the full model achieves the lowest AE, RE, overshoot ratio, oscillation count, and cross-content σ , indicating that the proposed dual-channel controller provides more accurate and stable bitrate regulation.

To further evaluate bitrate control accuracy under different target bitrates, we compare the actual output bitrate with the target bitrate on the Kodak24 and Tecnick datasets, as shown in Fig. 8. The dashed line represents the ideal case where the actual bitrate is equal to the target bitrate. It can be observed that the full model consistently stays close to the ideal line on both datasets, indicating that the proposed controller can accurately approach different target bitrates.

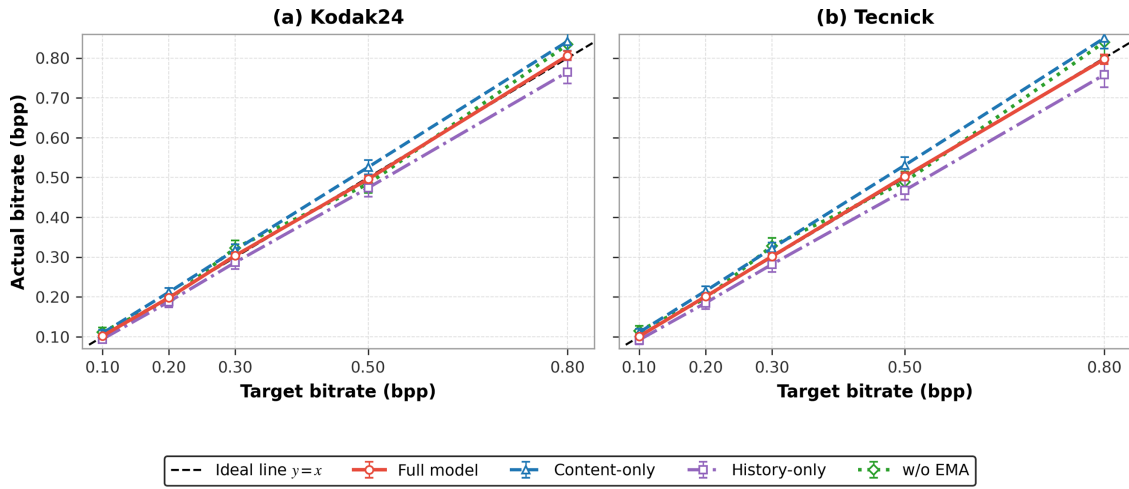


Figure 8: Actual bitrate vs. target bitrate on the Kodak24 and Tecnick datasets. The dashed line denotes the ideal line $y = x$. The closer the curve is to the dashed line, the more accurate the bitrate control is. (a) Kodak24; (b) Tecnick.

In contrast, the content-only variant tends to produce higher actual bitrates, especially at medium and high target bitrates. This suggests that relying only on content complexity may overestimate the importance of texture-rich regions and allocate excessive bits. The history-only variant generally produces lower actual bitrates, indicating that historical feedback alone may over-suppress the current image and lead to under-coding. The variant without EMA also shows larger deviation from the ideal line, demonstrating that EMA smoothing is beneficial for stabilizing bitrate regulation.

The relative bitrate deviation is defined as $\Delta_t = (\text{Actual bpp} - \text{Target bpp}) / \text{Target bpp} \times 100\%$.

Fig. 9 further shows the convergence behavior of the bitrate controller by plotting the signed relative bitrate deviation Δ_t under two representative target bitrates, 0.10 and 0.50 bpp. The gray region denotes the $\pm 3\%$ control band. Compared with the other variants, the full model exhibits faster convergence and smaller oscillation. After the initial adjustment stage, its bitrate deviation gradually approaches zero and remains within a narrow range around the target bitrate.

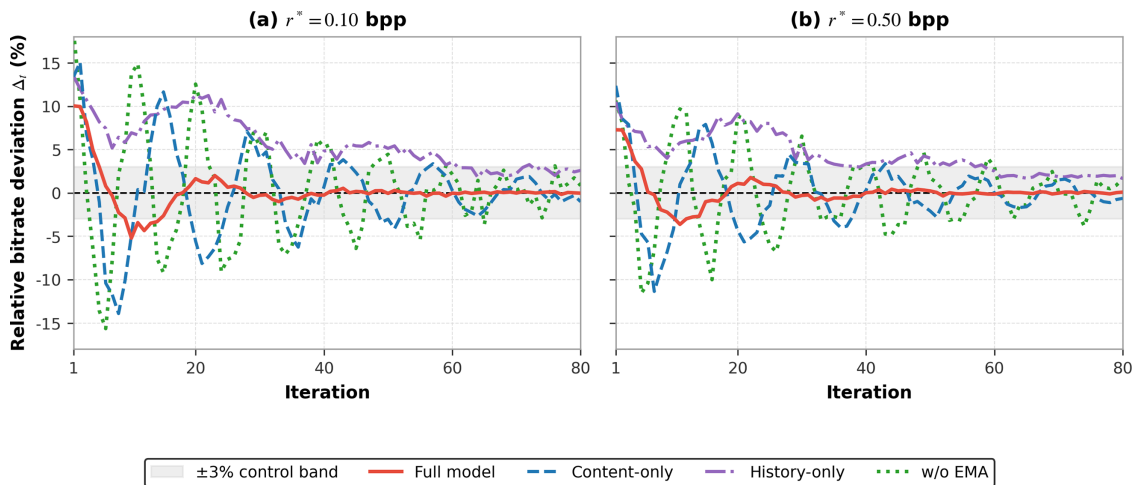


Figure 9: Trajectories of relative bitrate deviation under different target bitrates. The shaded region denotes the $\pm 3\%$ control band. (a) $r^* = 0.10$ bpp; (b) $r^* = 0.50$ bpp.

The content-only variant shows stronger oscillation because it adjusts the bitrate mainly according to the instantaneous content complexity of the current image. The history-only variant converges more slowly and tends to maintain a positive bias, indicating that historical feedback alone cannot sufficiently adapt to changes in image content. The w/o EMA variant presents more frequent fluctuations, especially during the early iterations, which confirms the importance of EMA smoothing in suppressing short-term bitrate oscillations. These results demonstrate that the combination of content awareness, historical feedback, and EMA smoothing improves both bitrate accuracy and convergence stability.

Fig. 10 presents the training convergence and the evolution of bitrate-control variables. As shown in Fig. 10a, both training loss and validation loss decrease smoothly during training, indicating that the proposed framework can be optimized stably without obvious divergence. Fig. 10b shows that the full model achieves the fastest decrease in bitrate relative error and maintains the lowest RE after convergence, which is consistent with the results in Table 5.

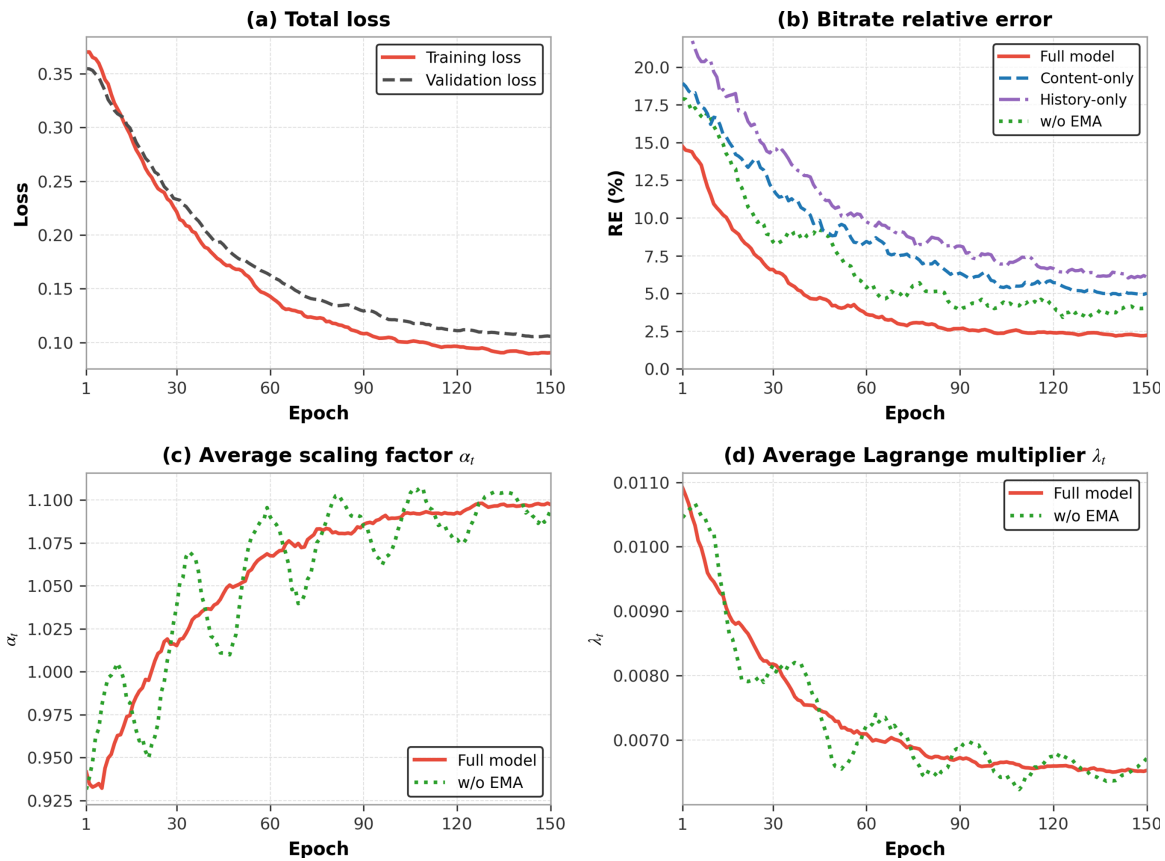


Figure 10: Training convergence and bitrate-control variable trajectories. (a) Total loss; (b) bitrate relative error; (c) average scaling factor α_t ; (d) average Lagrange multiplier λ_t .

Fig. 10c,d further illustrate the evolution of the average scaling factor α_t and the Lagrange multiplier λ_t . The scaling factor α_t gradually increases and then becomes stable, reflecting that the model progressively adjusts the amplitude of latent features to approach the target bitrate. Meanwhile, λ_t decreases and converges to a stable range, indicating that the rate-distortion trade-off is adaptively adjusted during training. Compared with the full model, the w/o EMA variant exhibits more noticeable fluctuations in both α_t and λ_t , which verifies that EMA smoothing helps stabilize the closed-loop bitrate control process.

These convergence curves demonstrate that the proposed bitrate control mechanism not only improves final bitrate accuracy but also enhances the stability of the training and control process.

5 Discussion

The framework introduces textual semantic information as a conditional guidance signal into learning-based image compression. The effectiveness of this strategy depends on the reliability and specificity of the semantic descriptions. When the image descriptions generated by BLIP are inaccurate or overly general, a semantic mismatch may arise between the text guidance and the actual image content. In such cases, the semantic branch may only provide weak, or even partially misleading, modulation information. Nevertheless, the method proposed in this paper does not rely entirely on the text branch for image reconstruction. Visual latent representations remain the primary carriers of information, while textual semantic features serve merely as soft conditioning signals through linear modulation at the feature level. Consequently, inaccurate image descriptions are generally unlikely to directly generate erroneous visual content, but they may weaken the semantic enhancement effect and reduce the benefits in terms of bitrate allocation. Therefore, future work could incorporate semantic confidence estimation, image description filtering, or multi-description ensemble strategies to improve the model's robustness in cases of semantic mismatch. The framework we propose can also be extended to cross-modal scenarios, where semantic guidance is no longer provided by automatically generated image descriptions but rather by user-defined prompts. In principle, user-defined prompts can provide more task-oriented or region-specific semantic constraints, particularly when users wish to preserve specific objects or visual attributes during the compression process. However, the effectiveness of such prompts largely depends on their consistency with the image content. Accurate prompts help the model highlight task-relevant areas, whereas vague or inconsistent prompts may introduce unreliable semantic prior knowledge. Therefore, user-defined prompts should be viewed as optional external semantic guidance rather than a required input for the model. A promising future research direction is to design a mechanism for evaluating prompt consistency, enabling the model to adaptively adjust the intensity of semantic modulation based on the degree of alignment between the prompt and the visual content. Another potential limitation arises under extremely low bitrate conditions. When the bitrate budget is severely constrained, the model may preserve high-level semantic consistency but lose detailed structural information, leading to semantic drift or structural degradation. In the method described in this paper, the text semantic branch serves only as a conditional modulation signal rather than a direct generative prior; visual latent variables remain the primary carriers of information. Therefore, text semantics is primarily used to preserve high-level semantic cues and mitigate severe semantic degradation, but it cannot fully compensate for the loss of structural information at extremely low bit rates. When it comes to abstract images and artistic content, high-level semantics are often ambiguous and difficult to accurately capture using traditional image description models. In such cases, automatically generated image descriptions may fail to capture non-object-level visual attributes, such as style, composition, color distribution, or abstract textures. Therefore, compared to natural images with clear objectives and contexts, the effectiveness of text-semantic guidance on such images may be limited. While the method described in this paper can still rely on visual features to achieve compression, the additional benefits provided by the semantic branch may be relatively limited.

Based on the above analysis, future work can further improve the robustness and practical applicability of text-guided learned image compression. More reliable semantic guidance strategies, such as semantic confidence estimation, caption filtering, and multi-caption ensemble mechanisms, are worth exploring to reduce the influence of inaccurate or overly generic text descriptions. Another promising direction is to develop lightweight vision-language feature extraction modules, which could reduce the encoder-side

computational overhead while preserving the benefits of semantic modulation. Furthermore, the bitrate control mechanism can be extended to more diverse image contents and deployment scenarios, enabling the model to maintain stable rate-distortion performance under different target bitrates and resource constraints.

6 Conclusion

In terms of practical applicability, the proposed framework is particularly suitable for bandwidth-constrained image transmission, cloud-side image storage, and semantic-preserving visual communication scenarios where maintaining key structures, salient objects, and perceptually important regions is more critical than achieving the lowest possible encoding latency. Its main advantages lie in three aspects. First, the integration of BLIP- and CLIP-based text-semantic guidance enables the encoder to distinguish semantically important regions from less informative background content, thereby improving structural preservation and perceptual consistency under limited bitrate budgets. Second, the content-aware bitrate controller with historical feedback allows the model to approach different target bitrates more stably and reduces unnecessary bit consumption caused by semantically weak high-frequency responses. Third, the proposed variable-rate design enables multiple bitrate operating points to be covered within a unified model, which improves its flexibility in practical deployment. Nevertheless, the proposed method still has several limitations. The use of BLIP and CLIP introduces additional encoder-side computational overhead, making the current framework more suitable for scenarios where encoding complexity is acceptable rather than strict real-time compression applications. In addition, the effectiveness of semantic guidance depends partly on the reliability of the generated textual description. When the caption is inaccurate, overly generic, or when the input image contains abstract or artistic content with ambiguous semantics, the semantic modulation may provide weaker guidance. Moreover, under extremely low bitrate conditions, although the model can help preserve high-level semantic consistency, fine structural fidelity may still degrade. Future work will therefore focus on lightweight semantic extraction, more robust semantic confidence estimation, and more efficient text-visual alignment mechanisms to further improve the practicality and robustness of the proposed framework.

Acknowledgement: Not applicable.

Funding Statement: This research was supported by the National Natural Science Foundation of China, Grant No. 51365019; the Construction Project of Key Laboratory of Industrial Intelligence and Systems, Department of Education of Yunnan Province, Grant No. KKPH202403003; and the Key Project of Science and Technology Plan of China National Tobacco Corporation Yunnan Provincial Company, Grant No. 2022530000241026.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Kaisen Li and Yunwei Zhang; methodology, Kaisen Li and Yunwei Zhang; software, Kaisen Li; validation, Kaisen Li and Guoying Sun; formal analysis, Kaisen Li and Guoying Sun; investigation, Kaisen Li; resources, Yunwei Zhang and Bin Li; data curation, Kaisen Li; writing—original draft preparation, Kaisen Li; writing—review and editing, Yunwei Zhang, Guoying Sun and Bin Li; visualization, Kaisen Li and Guoying Sun; supervision, Yunwei Zhang and Bin Li; project administration, Yunwei Zhang and Bin Li; funding acquisition, Yunwei Zhang and Bin Li. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study are publicly available. The MS-COCO dataset was used for training, and Kodak24, Tecnick, the CLIC validation set, Urban100, and DIV2K were used for testing. The data generated or analyzed during this study are included in this article. Further information is available from the corresponding authors upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: Yunwei Zhang and Bin Li have received research grants from China National Tobacco Corporation Yunnan Provincial Company under the Key Project of Science and Technology Plan, Grant No. 2022530000241026. Bin Li is affiliated with the Information Center, Yunnan Tobacco and Leaf Company, a for-profit tobacco enterprise. These financial and institutional relationships are disclosed as potential conflicts of interest. The remaining authors declare that they have no conflicts of interest.

References

1. Wallace GK. The JPEG still picture compression standard. *Commun ACM*. 1991;34(4):30–44.
2. Skodras A, Christopoulos C, Ebrahimi T. The JPEG 2000 still image compression standard. *IEEE Signal Process Mag*. 2000;18(5):36–58.
3. Albalawi U, Mohanty SP, Kougianos E. A hardware architecture for better portable graphics (BPG) compression encoder. In: *Proceedings of the 2015 IEEE International Symposium on Nanoelectronic and Information Systems*; 2015 Dec 21–23; Indore, India. p. 291–6.
4. Ballé J, Laparra V, Simoncelli EP. End-to-end optimized image compression. *arXiv:1611.01704*. 2016.
5. Chen T, Liu H, Ma Z, Shen Q, Cao X, Wang Y. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans Image Process*. 2021;30:3179–91. doi:10.1109/TIP.2021.3058615.
6. Li M, Zhang K, Li J, Zuo W, Timofte R, Zhang D. Learning context-based nonlocal entropy modeling for image compression. *IEEE Trans Neural Netw Learn Syst*. 2021;34(3):1132–45. doi:10.1109/tnnls.2021.3104974.
7. Fu H, Liang F, Lin J, Li B, Akbari M, Liang J, et al. Learned image compression with Gaussian-Laplacian-logistic mixture model and concatenated residual modules. *IEEE Trans Image Process*. 2023;32:2063–76. doi:10.1109/tip.2023.3263099.
8. Fu H, Liang F, Liang J, Li B, Zhang G, Han J. Asymmetric learned image compression with multi-scale residual block, importance scaling, and post-quantization filtering. *IEEE Trans Circuits Syst Video Technol*. 2023;33(8):4309–21. doi:10.1109/tcsvt.2023.3237274.
9. Zhu Y, Yang Y, Cohen T. Transformer-based transform coding. In: *Proceedings of the International Conference on Learning Representations*; 2022 Apr 25–29; Virtual.
10. Zou R, Song C, Zhang Z. The devil is in the details: window-based attention for image compression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022 Jun 19–24; New Orleans, LA, USA. p. 17492–501.
11. Xie Y, Cheng KL, Chen Q. Enhanced invertible encoding for learned image compression. In: *Proceedings of the 29th ACM International Conference on Multimedia*; 2021 Oct 20–24; Chengdu, China. p. 162–70.
12. Xu H, Hai B, Tang Y, He Z. Window-based channel attention for wavelet-enhanced learned image compression. In: *Proceedings of the Asian Conference on Computer Vision*; 2024 Dec 8–12; Hanoi, Vietnam. p. 4334–51.
13. Fu H, Liang J, Fang Z, Han J, Liang F, Zhang G. WeConvene: learned image compression with wavelet-domain convolution and entropy model. In: *Proceedings of the European Conference on Computer Vision*; 2024 Sep 29–Oct 4; Milan, Italy. Cham, Switzerland: Springer; 2024. p. 37–53.
14. Agustsson E, Tschannen M, Mentzer F, Timofte R, Gool LV. Generative adversarial networks for extreme learned image compression. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 221–31.
15. Li M, Zuo W, Gu S, Zhao D, Zhang D. Learning convolutional networks for content-weighted image compression. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2018 Jun 18–22; Salt Lake City, UT, USA. p. 3214–23.
16. Minnen D, Ballé J, Toderici GD. Joint autoregressive and hierarchical priors for learned image compression. *Adv Neural Inf Process Syst*. 2018;31:10794–803.
17. Lee H, Kim M, Kim JH, Kim S, Oh D, Lee J. Neural image compression with text-guided encoding for both pixel-level and perceptual fidelity. *arXiv:2403.02944*. 2024.
18. Qin S, Chen B, Huang Y, An B, Dai T, Xia ST. Perceptual image compression with cooperative cross-modal side information. *arXiv:2311.13847*. 2023.

19. Liu Y, Yang W, Bai H, Wei Y, Zhao Y. Region-adaptive transform with segmentation prior for image compression. In: Proceedings of the European Conference on Computer Vision; 2024 Sep 29–Oct 4; Milan, Italy. Cham, Switzerland: Springer; 2024. p. 181–97.
20. Sun Q, Guo C, Yang Y, Chen J, Xue X. Semantic-assisted image compression. arXiv:2201.12599. 2022.
21. Lei E, Uslu YB, Hassani H, Bidokhti SS. Text+ sketch: image compression at ultra low rates. arXiv:2307.01944. 2023.
22. Li C, Lu G, Feng D, Wu H, Zhang Z, Liu X, et al. MISC: ultra-low bitrate image semantic compression driven by large multimodal model. IEEE Trans Image Process. 2024;34:335–49.
23. Song M, Choi J, Han B. Variable-rate deep image compression through spatially-adaptive feature transform. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Virtual. p. 2380–9.
24. Guerin Jr ND, da Silva RC, de Oliveira MC, Jung HC, Martins LGR, Peixoto E, et al. Rate-constrained learning-based image compression. Signal Process Image Commun. 2022;101(1):116544. doi:10.1016/j.image.2021.116544.
25. Gupta R, BV S, Kapoor N, Jaiswal R, Nangi SR, Kulkarni K. User-guided variable rate learned image compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 1753–8.
26. Choi Y, El-Khamy M, Lee J. Variable rate deep image compression with a conditional autoencoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 3146–54.
27. Yang F, Herranz L, Van De Weijer J, Guitián JAI, López AM, Mozerov MG. Variable rate deep image compression with modulated autoencoder. IEEE Signal Process Lett. 2020;27:331–5. doi:10.1109/lsp.2020.2970539.
28. Kamisli F, Racapé F, Choi H. Variable-rate learned image compression with multi-objective optimization and quantization-reconstruction offsets. In: Proceedings of the 2024 Data Compression Conference (DCC); 2024 Mar 19–22; Snowbird, UT, USA. p. 193–202.
29. Ma Y, Zhai Y, Yang C, Yang J, Wang R, Zhou J, et al. Variable rate ROI image compression optimized for visual quality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 19–25; Virtual. p. 1936–40.
30. Kao CH, Weng YC, Chen YH, Chiu WC, Peng WH. Transformer-based variable-rate image compression with region-of-interest control. In: Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP); 2023 Oct 8–11; Kuala Lumpur, Malaysia. p. 2960–4.
31. Ballé J, Minnen D, Singh S, Hwang SJ, Johnston N. Variational image compression with a scale hyperprior. arXiv:1802.01436. 2018.
32. Li H, Li S, Dai W, Li C, Zou J, Xiong H. Frequency-aware transformer for learned image compression. arXiv:2310.16387. 2023.
33. He D, Yang Z, Peng W, Ma R, Qin H, Wang Y. ELIC: efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 19–24; New Orleans, LA, USA. p. 5718–27.
34. Cheng Z, Sun H, Takeuchi M, Katto J. Learned image compression with discretized Gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 14–19; Virtual. p. 7939–48.
35. Yi F, Gong Z, Zeng H. Rethinking latent variable in learned image compression. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2026 Mar 6–10; Tucson, AZ, USA. p. 8531–40.
36. Wang Y, Fu H, Cao Q, Wang S, Chen Z, Liang F. S2LIC: learned image compression with the SwinV2 block, adaptive channel-wise and global-inter attention context. Neural Netw. 2025;189:107590.
37. Li Y, Zhang H, Li L, Liu D. Learned image compression with hierarchical progressive context modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–23; Honolulu, HI, USA. p. 18834–43.
38. Wu Z, Du H, Wang S, Lu M, Sun H, Guo Y, et al. CMamba: learned image compression with state space models. arXiv:2502.04988. 2025.

39. Presta A, Tartaglione E, Fiandrotti A, Grangetto M, Cosman P. Efficient progressive image compression with variance-aware masking. In: Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2025 Feb 28–Mar 4; Tucson, AZ, USA. p. 7692–700.
40. Zhang Y, Huang Y, Zhu F. QARV++: an improved hierarchical VAE for learned image compression. IEEE Trans Circuits Syst Video Technol. 2026;36(6):7921–36.