Generative Image Coding with Diffusion Prior

Jianhui Chang
China Telecom Cloud Computing Research Institute
changjh1@chinatelecom.cn

***Abstract*—As generative technologies advance, visual content has evolved into a complex mix of natural and AI-generated images, driving the need for more efficient coding techniques that prioritize perceptual quality. Traditional codecs and learned methods struggle to maintain subjective quality at high compression ratios, while existing generative approaches face challenges in visual fidelity and generalization. To this end, we propose a novel generative coding framework leveraging diffusion priors to enhance compression performance at low bitrates. Our approach employs a pre-optimized encoder to generate generalized compressed-domain representations, integrated with pretrained model's internal features via a lightweight adapter and an attentive fusion module. This framework effectively leverages existing pretrained diffusion models and enables efficient adaptation to different pretrained models for new requirements with minimal retraining costs. We also introduce a distribution renormalization method to further enhance reconstruction fidelity. Extensive experiments show that our method: (1) outperforms existing methods in visual fidelity across low bitrates, (2) improves compression performance by up to 79% over H.266/VVC, and (3) offers an efficient solution for AI-generated content while being adaptable to broader content types.**

***Index Terms*—Generative coding, image compression, latent diffusion models, AI-generated content**

I. Introduction

With the rapid advancement of generative technologies, visual content has evolved from predominantly natural images to a diverse blend of natural and generated images, driving the need for more efficient image coding techniques that optimize subjective perceptual quality. Transform-based traditional coding standards (*e.g.*, HEVC [1], VVC [2]) and end-to-end learned coding approaches [3], [4] tend to face challenges in maintaining subjective reconstruction quality at high compression ratios, primarily due to data degradation and insufficient utilization of external priors.

Recent generative image coding approaches [5], [6] leverage the powerful ability of generative models to capture complex data distributions, enabling high reconstruction quality at ultra-low bitrates. While promising, current techniques mainly based on GANs [7] are limited by their generative capacity, particularly in preserving realistic textures and structural details. Moreover, the generalizability of current generative coding methods [6], [8] is limited, as compression performance is closely tied to the model's training scope. These methods typically require end-to-end rate-distortion (RD) optimization for specific scene data (*e.g.,* AI-generated content or natural scenes), making training resource-intensive. Additionally, for different scenarios, retraining the entire model is often necessary, which incurs significant overhead and decoding challenges across devices.

Diffusion models have recently achieved notable success in image generation, particularly in text-to-image tasks. Their ability to produce visually rich and well-structured images demonstrates substantial prior knowledge about visual content construction, spanning from low-level textures to high-level semantics. Existing works on diffusion-based compression follow different design paradigms. One line of research [9], [10] leverages diffusion models as post-processors atop existing codecs, focusing on enhancing decoded image quality rather than generative coding. Another line [11], [12] employs diffusion models as decoders conditioned on image-derived representations. For instance, text-conditioned diffusion decoders [11], [12] have been explored for exceptionally low bitrates (< 0.01 bpp), excelling at preserving high-level semantic consistency, but often sacrificing visual fidelity. Additionally, the concurrent work PerCo [12] utilizes vector-quantized codebooks to represent images alongside textual information. However, such methods typically require separate training of the auxiliary encoder with codebooks of different sizes, along with finetuning of the diffusion model for varying target rates and content, limiting their flexibility and efficiency in adapting to diverse compression needs.

To this end, this paper proposes a generative coding framework that leverages powerful diffusion priors to enhance compression performance, focusing primarily on improving perceptual fidelity across a broad range of low bitrates, typically around 0.01 to 0.2 bpp. We leverage pretrained latent diffusion models [13] as generative decoders, while optimizing the encoder via an pretext end-to-end learned compression task, independent of specific generative models. A lightweight adapter is introduced to ensure effective domain adaptation, providing a flexible and compatible solution across different models. To further improve reconstruction fidelity and compression efficiency, we incorporate a cross-attention mechanism for better alignment between compressed latents and internal features, and a distribution renormalization method to mitigate reconstruction distortion. The primary contributions of this work are as follows:

- We propose a generative coding framework leveraging powerful diffusion priors to achieve efficient human-centered compression. The framework includes a lightweight adapter and an attentive fusion module that enable effective domain adaptation and ensure compatibility across various pretrained latent diffusion models.

This work is partially sponsored by the National Key Research and Development Program of China (2024YFB4505704).

- We introduce a cross-attention mechanism to refine the integration of compressed latents and the pretrained model's internal features, along with a distribution renormalization method to enhance reconstruction fidelity, both of which further boost compression performance.
- Extensive experiments show that our method achieves up to a 79% improvement in compression performance compared to VVC, while demonstrating versatility across natural and AI-Generated scenarios with different pretrained diffusion models.

II. Proposed Method

The main framework of the proposed generative image coding method is shown in Fig. 1. It encompasses several key components: (1) an image encoder $\mathcal{E}$ and an entropy model $\mathcal{H}$, which collaborate to encode images into rate-constrained latents; (2) a latent adapter $\mathcal{F}$ equipped with an attentive fusion module, tasked with transforming the compressed latents into control signals; and (3) a latent diffusion model [13] (LDM) $\mathcal{G}$, which serves as a robust prior for realistic image reconstruction. Our objective is to harness the capabilities of the pretrained model $\mathcal{G}$ for image reconstruction by integrating compressed latents with its internal priors through a lightweight $\mathcal{F}$ and attentive fusion, thus enabling accurate guidance of the reconstruction process by compressed signals.

At the encoder side, to ensure flexibility and generalization, the pre-optimized encoder $\mathcal{E}$ is used to transform the input image $\mathbf{x}$ into a unified compressed-domain latent representation $\mathbf{y}$, defined as $\mathbf{y} = \mathcal{E}(\mathbf{x})$. After transformation, the latents $\mathbf{y}$ are quantized and entropy-encoded into bitstreams. On the decoder side, the decoded latents $\hat{\mathbf{y}}$ are transformed by the latent adapter $\mathcal{F}$ into a feature set $\mathbf{f}$. This set is then aligned and fused with the intermediate representations $\mathbf{c}_t$ of the denoising U-Net at each time step t within the latent diffusion model. This integration of encoded information steers the generative process, enabling the synthesis of a high-fidelity reconstructed image $\hat{\mathbf{x}}$. In the latent diffusion model, the generative process begins with random noise $\mathbf{z}$ sampled from a prior Gaussian distribution $\mathcal{N}(0, 1)$ and is conditioned on the compressed information $\hat{\mathbf{y}}$, expressed as:

$$\hat{\mathbf{x}} = \mathcal{G}(\mathbf{z}, \hat{\mathbf{y}}), \hat{\mathbf{x}} \sim P(\hat{\mathbf{x}} \mid \hat{\mathbf{y}}). \tag{1}$$

Since the compressed information $\hat{\mathbf{y}}$ provides strong spatial guidance, the target distribution $P(\hat{\mathbf{x}})$ is primarily determined by $\hat{\mathbf{y}}$ and the internal diffusion priors, allowing us to achieve high-fidelity reconstruction results for the compression task.

A. Latent Adapter and Attentive Fusion

Various approaches have been explored to adapt pretrained models to different tasks or domains in computer vision, though they may not be ideally suited for image compression. For instance, ControlNet [14] employs a trainable copy of a large-scale pretrained backbone to generate additional feature maps, incurring significant computational overhead. Another representative approach, T2I-Adapter [15], controls the color and structure of generated images but does not provide the required granularity to achieve the high perceptual reconstruction fidelity targeted by our method.

To address these challenges, this paper designs a latent adapter and an attentive latent fusion module for high visual fidelity generative coding. The adapter aligns the compressed-domain latents $\hat{\mathbf{y}}$ with features from both the downsampling and upsampling modules of the U-Net, defined as $\mathbf{f} = \mathcal{F}(\hat{\mathbf{y}})$. Each feature extraction module includes a convolutional layer and two residual modules, producing features f_i aligned with the corresponding U-Net features $c_t^{(i)}$ at each time step t.

The transformed latents $\mathbf{f}$ and internal features $\mathbf{c}_t$ are then fused and integrated into the generative diffusion process for image reconstruction. The commonly used additive fusion [14], [15] method assumes spatial alignment and overlooks contextual information, limiting reconstruction capability. Thus, we introduce an attentive fusion module to enhance fusion accuracy for improving image reconstruction fidelity based on spatial cross-attention. As shown in Fig. 2, the U-Net intermediate features $c_t^{(i)}$ and transformed latents f_i are summed to form the base feature $\hat{c}_t^{(i)} = c_t^{(i)} + f_i$, enhancing spatial correlation. The transformed latents f_i, derived from compressed information, act as context vectors. By reshaping their dimensionality to $HW \times C$, we achieve a linearization of the spatial dimensions, allowing each feature vector to contribute independently to the attention mechanism. Three 1×1 convolution layers then map $\hat{c}_t^{(i)}$ to $\mathbf{Q}$, and reshape f_i to $\mathbf{K}$ and $\mathbf{V}$. With FC denoting a fully connected layer, the fused feature $\hat{\mathbf{f}}_i$ is computed as:

$$\hat{\mathbf{f}}_i = \mathbf{V} + FC(Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V})), \tag{2}$$

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{C}}\right) \cdot \mathbf{V}. \tag{3}$$

The cross-attention mechanism facilitates fusion by capturing correlations between compressed information and the generative model's internal features at each spatial location. Consequently, it enables the latent diffusion model to obtain more precise spatial details of the transformed latents derived from target images, effectively improving the accuracy of reconstructed images.

B. Optimization Strategy

Pretrained generative models usually provide strong priors for visual reconstruction, but retraining them is costly and risks degrading these priors. To avoid this, we keep the pretrained latent diffusion model $\mathcal{G}$ frozen and employ a two-stage optimization strategy: first optimizing the encoder $\mathcal{E}$ and entropy estimation model $\mathcal{H}$ through a pretext task, then freezing them and fine-tuning the lightweight adapter $\mathcal{F}$ and the attentive fusion module.

(1) **Encoder Optimization.** During the first stage, a learned image compression task [16] is employed for pretext optimization. This involves an encoder $\mathcal{E}$, an auxiliary decoder, and an entropy model $\mathcal{H}$. This auxiliary decoder is introduced specifically for the pretext optimization and will be discarded

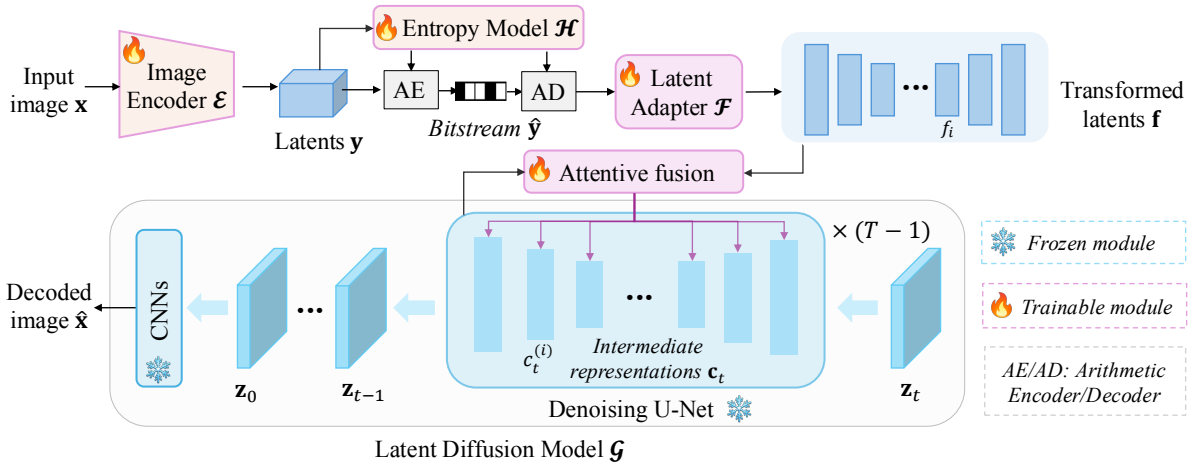


Fig. 1: Overview of the proposed generative image coding framework.

after the optimization is completed. To achieve variable-rate coding with a unified encoder and entropy model, we utilize channel-wise scaling and quantization techniques, as detailed in [17]. For rate-distortion optimization [18], [19], each Lagrange multiplier λ_s is associated with a set of quantization parameters for rate level s. The rate constraint $\mathcal{L}^s_{rate}$ is determined by $\mathcal{H}$. Using MS-SSIM [20] to measure the reconstruction distortion $\mathcal{L}^s_{dist}$, the pretext optimization objective is formulated as follows:

$$\mathcal{L}_{pretext} = \sum_{s=0}^{L-1} \mathcal{L}^s_{rate} + \lambda_s \mathcal{L}^s_{dist}. \tag{4}$$

During training with Eq. (4), the auxiliary decoder is updated alongside $\mathcal{E}$ and $\mathcal{H}$ via stochastic gradient descent. Upon completion of the pretext optimization, only the parameters of $\mathcal{E}$ and $\mathcal{H}$ are retained and remain fixed for the subsequent training phase with other modules.

(2) **Adapter Optimization.** After the first-stage training, the second stage targets the optimization of the latent adapter $\mathcal{F}$ and the attentive fusion module with $\mathcal{E}$, $\mathcal{H}$ and $\mathcal{G}$ kept fixed. In the diffusion process, latents $\mathbf{z}_0$ from the input image $\mathbf{x}$ are perturbed with noise at each timestep t to produce $\mathbf{z}_t$. Image reconstruction is performed iteratively in the reverse diffusion process by denoising $\mathbf{z}_t$ using a U-Net noise estimator ϵ_θ, conditioned on the transformed compressed feature set $\mathbf{f}$. The denoising loss function is defined as follows:

$$\mathcal{L}_{adp} = \mathbb{E}_{\mathbf{z}_0,t,\mathbf{f},\epsilon\sim\mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{f})\|_2^2 \right]. \tag{5}$$

In the training phase, input images are encoded by $\mathcal{E}$, transmitted, and transformed by $\mathcal{F}$ to $\mathbf{f}$, which is then integrated into the denoising process via the fusion module. With other modules held fixed, $\mathcal{F}$ and the fusion module undergo fine-tuning, enabling the utilization of the pre-trained model $\mathcal{G}$ for the image compression task upon completion.

C. Fidelity Enhancement

The proposed generative coding method successfully reconstructs images while maintaining consistent visual semantics and structure. However, distortions in color distribution highlight the need for improved fidelity. In image processing,

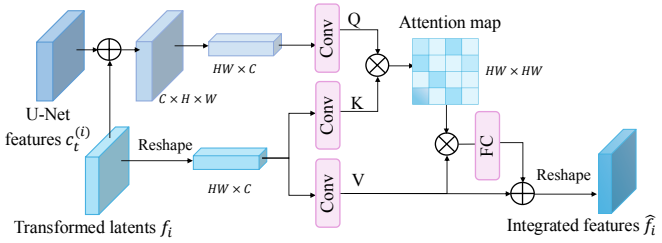


Fig. 2: Attentive latent fusion based on spatial cross-attention.

color moments—mean (μ) and standard deviation (σ)—are fundamental metrics for characterizing color distribution, with the mean capturing brightness and the standard deviation reflecting contrast through pixel value variation. Style transfer research [21] demonstrates that tuning instance normalization in DNN feature spaces retains content while transferring color styles. Extending this insight, we propose a method to correct color deviations in the synthesized images' *pixel domain* by aligning their channel-wise mean and standard deviation with the original image, thereby enhancing reconstruction fidelity.

Specifically, to minimize reconstruction distortion, the color distribution parameters of the reconstructed image should align with those of the original. The mean (μ) and standard deviation (σ) are computed for each channel, with $\mu = \mu_R, \mu_G, \mu_B$ and $\sigma = \sigma_R, \sigma_G, \sigma_B$. For efficient transmission, these parameters are quantized to $\hat{\mu}$ and $\hat{\sigma}$ with a step size Δ. On the decoder side, the initial reconstruction $\hat{\mathbf{x}}$ is renormalized for fidelity enhancement using the transmitted parameters $\hat{\mu}$ and $\hat{\sigma}$ as follows:

$$\hat{\mathbf{x}}_{norm} = \left(\frac{\hat{\mathbf{x}} - \hat{\mu}}{\hat{\sigma}}\right)\sigma + \mu, \tag{6}$$

where $\hat{\mathbf{x}}_{norm}$ represents the final decoded image after improving color fidelity. The proposed approach significantly improves color accuracy in reconstructed images. Moreover, we find that using statistics from smaller regions leads to more accurate color correction. This approach is further extended to block-level correction, where statistical parameters are transmitted and applied for each image block, ensuring more precise color alignment with negligible bitrate cost.

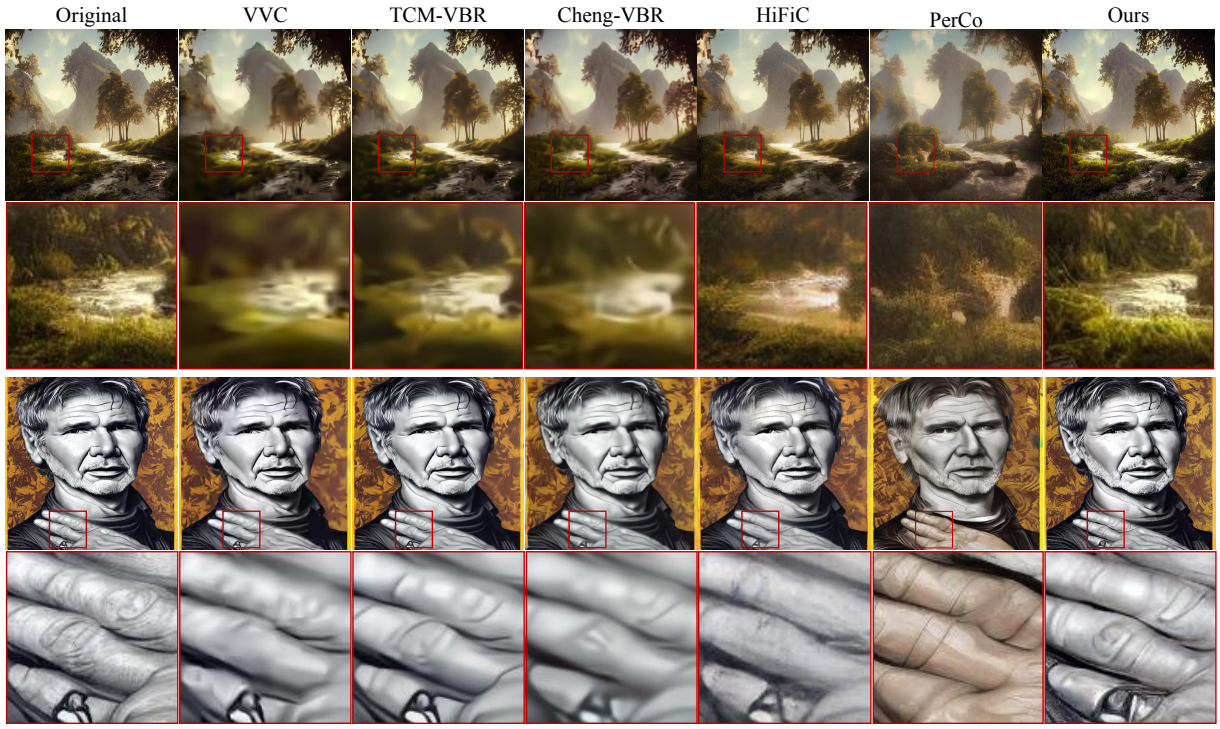


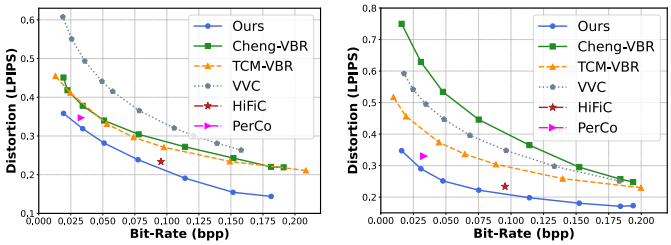
Fig. 3: Qualitative comparisons of VVC, Cheng-VBR, TCM-VBR, HiFiC, PerCo and ours on the AIGC dataset.

III. Experiments

A. Experimental Settings

a) Datasets: Artificial Intelligence Generated Content (AIGC) has recently become a key AI research focus. The statistics of AIGC data align well with generative model priors, making it well-suited for generative coding. Thus, this study utilizes the DiffusionDB dataset [22], containing 14 million high-quality, predominantly anime-style images generated from real user prompts. For experiments, 50,634 images are randomly selected for training and 200 for testing at a resolution of 512×512. Besides, to evaluate the method's compatibility with other pretrained diffusion models on natural scenes, 100,000 images of OpenImages dataset[1] are used to fine-tune the latent adapter. The Kodak dataset[2] is used to assess compression performance on real-world images.

b) Implementation Details: All experiments are conducted using PyTorch on four NVIDIA Tesla 32G-V100 GPUs. The Adam optimizer is used with a learning rate of $1e-5$. The first-stage model follows Cheng et al. [16] with 128 latent channels, is trained for 600 epochs with Lagrange multipliers {50.0, 16.0, 3.0, 1.0, 0.5, 0.25, 0.1, 0.05, 0.01, 0.005} and a batch size of 96, starting from pretrained CompressAI[3] parameters. The pretrained latent diffusion model is sourced from Stable Diffusion v1.4[4] with an input of 512×512 and latent dimensions of $4 \times 64 \times 64$. The latent adapter network and fusion module, shared across bitrates, are fine-tuned for 10 epochs with a batch size of 8. During inference, images are generated using the DDIM deterministic sampling schedule [23] with 10 sampling steps. The test results are obtained using a global random seed of 42, ensuring deterministic initialization of latent noise $\mathbf{z}$ for the denoising process. Varying the random seed produces a set of reconstructions that may reflect the uncertainty inherent in specific compressed latents [12]. The standard deviation of LPIPS values across different random initializations is empirically below 0.01 in our results, indicating its impact on reconstruction results is negligible. For fidelity enhancement, a block size of 64×64 is used, with parameters quantized to 6 bits. Encoding these parameters adds an average of 0.01 bpp.



(a) R-D results on DiffusionDB. (b) R-D results on Kodak.

Fig. 4: The R-D comparisons on Kodak and the AIGC dataset, DiffusionDB. Lower LPIPS scores indicate higher fidelity.

c) Evaluation Metrics: Traditional coding methods typically focus on preserving signal fidelity, using metrics like PSNR and SSIM to evaluate pixel-level distortions. However, these metrics often struggle to reflect the perceptual quality. Likewise, dataset-level measures like FID and KID

[1]https://storage.googleapis.com/openimages/web/index.html
[2]https://r0k.us/graphics/kodak/
[3]https://interdigitalinc.github.io/CompressAI/
[4]https://huggingface.co/CompVis/stable-diffusion-v1-4

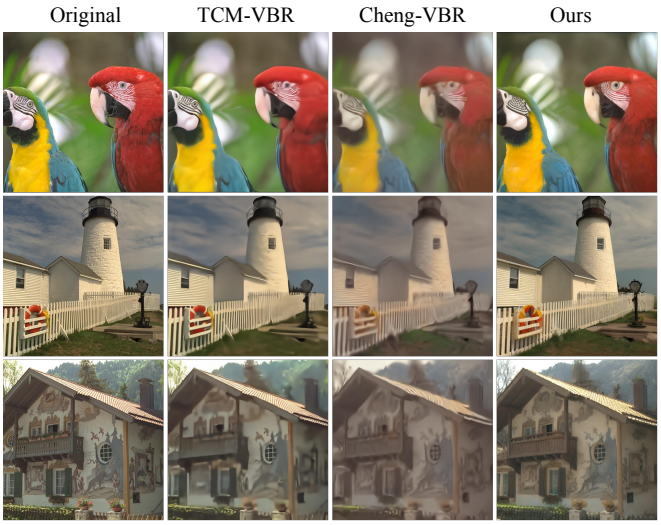


Fig. 5: Qualitative comparisons on the Kodak dataset.

assess distribution discrepancies but are insufficient for evaluating image-level perceptual fidelity. Therefore, we adopt the Learned Perceptual Image Patch Similarity (LPIPS) [24], which measures feature-domain distortion between paired images, better aligns with human perception, and is widely recognized in the field [5], [6]. Additionally, bits per pixel (bpp) is employed to evaluate rate performance.

B. Compression Performance Comparison

a) Compared Methods: The proposed method is compared with several representative baselines: (1) **VVC**, the Versatile Video Coding standard [2], using VTM-11.0 with intra-coding under common test conditions; (2) **Cheng-VBR**, a deep learning-based end-to-end compression method [16], used for pretext optimization in this study; (3) **TCM-VBR**, one of the state-of-the-art methods [4] utilizing a learnable quantization scale matrix and training strategy similar to this work for variable bitrate coding; and (4) **HiFiC**, a generative compression method [5], trained at a fixed low bitrate. (5) **PerCo**[5] [12], one of the latest diffusion-based generative coding methods. The released checkpoints use Stable Diffusion v2.1, which is more advanced than ours.

b) Quantitative Evaluation: Fig. 4 shows that the proposed generative coding method significantly outperforms VVC, Cheng-VBR [16], HiFiC, PerCo and TCM-VBR [4] in RD performance using LPIPS as the metric. Tab. I further quantifies this improvement with the Bjontegaard metric [25], showing bitrate reductions of 67.8% compared to VVC, 50.1% compared to Cheng-VBR, and 40.1% compared to TCM-VBR on the AIGC dataset. In terms of BD-LPIPS, the proposed method enhances reconstruction quality by 27.0% over Cheng-VBR, 24.5% over TCM-VBR, and 37.8% over VVC. TCM-VBR combines CNNs and transformers in its transform and entropy models to enhance both local and global feature capture, achieving better RD performance than Cheng-VBR and VVC. HiFiC's performance is constrained by the suboptimal generative capabilities of GANs, while PerCo introduces textual information that does not contribute to improving visual fidelity. In contrast, the proposed method leverages powerful diffusion priors and an attentive fusion mechanism, enabling more perceptually faithful reconstructions from compact latent representations. These results highlight the method's strong RD performance across a wide compression ratio range (100× to 2000×), demonstrating its effectiveness.

TABLE I: BD-Rate and BD-metric results relative to VVC, Cheng-VBR and TCM-VBR respectively. LPIPS is used as the distortion metric.

Dataset	Metric	VVC	Cheng-VBR	TCM-VBR
AIGC	BD-Rate	-67.75%	-50.14%	-40.08%
	BD-LPIPS	-37.76%	-27.03%	-24.46%
Kodak	BD-Rate	-79.24%	-81.62%	-69.93%
	BD-LPIPS	-39.42%	-45.80%	-29.34%

c) Qualitative Evaluation: Fig. 3 presents the subjective reconstruction results of the proposed generative compression method alongside those of VVC, Cheng-VBR [16], TCM-VBR [4], HiFiC [5] and PerCo [12]. VVC exhibits noticeable block artifacts, while TCM-VBR and Cheng-VBR suffer from excessive smoothing and blurring. HiFiC introduces distinct generative artifacts. In particular, textures such as hair, grass, leaves, and mountains appear blurred, with competing methods failing to preserve sharp edges, leading to degraded visual quality. While PerCo produces visually realistic results, its ability to preserve fine-grained details (*e.g.*, color, edges, shapes) is limited. In contrast, the proposed method, leveraging generative diffusion priors, not only produces sharp and realistic texture edges but also achieves higher visual fidelity in reconstructions. These results demonstrate its superior visual quality, validating its effectiveness in improving human-centered compression performance.

C. Generalizability Evaluation

The LDM used in this study is renowned for visual art generation, making the proposed generative coding method particularly suitable for AIGC artwork coding. To evaluate the method's compatibility with various pretrained models, we apply another pretrained diffusion model, Realistic Vision V6.0, sourced from Civitai[6], for compressing general photographic content. We finetune only the latent adapter $\mathcal{F}$ and the associated fusion module using the loss function in Eq. (5) for 2 epochs, while keeping all other modules fixed. The proposed method is mainly compared against Cheng-VBR [16] and TCM-VBR [4] on the Kodak dataset. Notably, the proposed method and Cheng-VBR shared the same encoder.

As the RD results shown in Fig. 4 (b) and Tab. I, the proposed method significantly outperforms HiFiC, PerCo, Cheng-VBR and TCM-VBR, achieving compression efficiency improvements of 81.6% over Cheng-VBR and 69.9% over TCM-VBR for in terms of BD-rate natural scene images. Additionally, our method achieves a reconstruction quality improvement of 45.8% over Cheng-VBR and 29.3% over TCM-VBR measured by BD-LPIPS. For subjective quality

[5]https://github.com/Nikolai10/PerCo

[6]https://civitai.com/models/4201/realistic-vision-v60-b1

TABLE II: Complexity analysis of the proposed method's modules and VVC reference software.

Method	VVC		Ours		
Module	Encoder	Decoder	Encoder	Adapter	LDM
Inference Time (ms)	3940.6	187.0	59.2	2.6	76.5

at an 800× compression ratio shown in Fig. 5, the end-to-end learned coding methods exhibit significant blurring and distortion. In contrast, the proposed method maintains high visual fidelity, producing sharp, realistic textures. These results validate the effectiveness of the proposed generative coding approach in ensuring compatibility with various pretrained diffusion models and delivering strong performance across diverse scenarios.

D. Complexity Analysis

Tab. II presents the inference time for each module of the proposed method, with the LDM reflecting single-step inference. For reference, the encoding and decoding times of the VVC reference software VTM-11.0 on the CPU platform for 512×512 resolution images are also provided. Notably, the adapter contributes minimally to the overall inference time. As a result, fine-tuning a lightweight adapter for different pretrained models introduces negligible computational overhead, offering an adaptable and computationally efficient solution for incorporating various generative diffusion priors. The reported compression performance of our method is based on 10 iterations of LDM inference. Despite increased inference time with more iterations, accelerated sampling research shows reduced complexity and fewer steps can yield quality images, enabling practical generative coding deployments.

IV. Conclusion

This paper presents a generative coding framework that achieves high perceptual quality at low bitrates. By leveraging diffusion priors through a pre-optimized encoder, lightweight adapter, and fusion module, our method ensures compatibility with various pretrained diffusion models. The integration of attentive feature fusion and distribution renormalization further enhances reconstruction fidelity, improving compression efficiency. Experimental results demonstrate that the proposed method outperforms H.266/VVC by up to 79%, showcasing its effectiveness and versatility across natural and AI-generated content. These findings highlight the method's potential as an efficient solution for AI-generated content and a flexible approach to various generative coding applications.

Acknowledgment

We thank Jie Wu, Hongbin Liu, Hao Yang and Siwei Ma for their insightful discussions and computational support.

References

[1] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[3] Johannes Ballé, Philip A Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici, "Nonlinear Transform Coding," *IEEE Journal of Selected Topics in Signal Processing*, pp. 339–353, 2020.

[4] Jinming Liu, Heming Sun, and Jiro Katto, "Learned Image Compression with Mixed Transformer-CNN Architectures," in *CVPR*, 2023, pp. 14388–14397.

[5] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," in *NeurIPS*, 2020, pp. 11913–11924.

[6] Jianhui Chang, Jian Zhang, Jiguo Li, Shiqi Wang, Qi Mao, Chuanmin Jia, Siwei Ma, and Wen Gao, "Semantic-aware visual decomposition for image coding," *International Journal of Computer Vision*, vol. 131, no. 9, pp. 2333–2355, 2023.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in *NeurIPS*, 2014, pp. 2672–2680.

[8] Jianhui Chang, Zhenghui Zhao, Chuanmin Jia, Shiqi Wang, Lingbo Yang, Qi Mao, Jian Zhang, and Siwei Ma, "Conceptual Compression via Deep Structure and Texture Synthesis," *IEEE Transactions on Image Processing*, vol. 31, pp. 2809–2823, 2022.

[9] Emiel Hoogeboom, Eirikur Agustsson, Fabian Mentzer, Luca Versari, George Toderici, and Lucas Theis, "High-fidelity image compression with score-based generative models," *arXiv preprint arXiv:2305.18231*, 2023.

[10] Noor Fathima Ghouse, Jens Petersen, Auke Wiggers, Tianlin Xu, and Guillaume Sautière, "A Residual Diffusion Model for High Perceptual Quality Codec Augmentation," *arXiv preprint arXiv:2301.05489*, 2023.

[11] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti, "Text+Sketch: Image Compression at Ultra Low Rates," in *ICML 2023 Workshop on Neural Compression*, 2023.

[12] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *ICLR*, 2023.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023, pp. 3836–3847.

[15] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan, "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI*, 2024, vol. 38, pp. 4296–4304.

[16] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, "Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules," in *CVPR*, 2020, pp. 7939–7948.

[17] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai, "Asymmetric Gained Deep Image Compression with Continuous Rate Adaptation," in *CVPR*, 2021, pp. 10532–10541.

[18] Claude Elwood Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[19] Johannes Ballé, Valero Laparra, and Eero Simoncelli, "End-to-end optimized image compression," in *ICLR*, 2017.

[20] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. IEEE, 2003, vol. 2, pp. 1398–1402.

[21] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017, pp. 1501–1510.

[22] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau, "DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models," in *ACL*, 2023, pp. 893–911.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising Diffusion Implicit Models," in *ICLR*, 2020.

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *CVPR*, 2018, pp. 586–595.

[25] Bjontegaard, Gisle, "Calculation of average PSNR differences between RD-curves," *ITU-T VCEG-M33*, 2001.