Towards Extreme Image Compression with Latent Feature Guidance and Diffusion Prior

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, Jingwen Jiang

***Abstract*—Compressing images at extremely low bitrates (below 0.1 bits per pixel (bpp)) is a significant challenge due to substantial information loss. Existing extreme image compression methods generally suffer from heavy compression artifacts or low-fidelity reconstructions. To address this problem, we propose a novel extreme image compression framework that combines compressive VAEs and pre-trained text-to-image diffusion models in an end-to-end manner. Specifically, we introduce a latent feature-guided compression module based on compressive VAEs. This module compresses images and initially decodes the compressed information into content variables. To enhance the alignment between content variables and the diffusion space, we introduce external guidance to modulate intermediate feature maps. Subsequently, we develop a conditional diffusion decoding module that leverages pre-trained diffusion models to further decode these content variables. To preserve the generative capability of pre-trained diffusion models, we keep their parameters fixed and use a control module to inject content information. We also design a space alignment loss to provide sufficient constraints for the latent feature-guided compression module. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches in terms of both visual performance and image fidelity at extremely low bitrates.**

***Index Terms*—Image compression, diffusion models, content variables, extremely low bitrates.**

I. Introduction

With the explosive growth of image data on the internet, how to achieve efficient data transmission and storage has become increasingly important. Lossy image compression is a popular solution for saving storage and transmission bandwidth by exploring the spatial and perceptual redundancy of images. Traditional compression standards, such as JPEG2000 [1], BPG [2], and VVC [3], are widely used in practice. However, these algorithms produce severe blocking artifacts at extremely low bitrates when applied to scenarios with limited bandwidth.

Recent learning-based image compression methods have attracted significant interest and also have great potential to outperform traditional codec methods. The main success of these methods is due to designing kinds of non-linear transforms [4], [5] and entropy models [6], [7]. However, these methods are usually optimized towards the rate-distortion trade-off, which generates over-smooth results at low bitrates. To achieve high-quality reconstruction at extremely low bitrates, some methods [8], [9], [10], [11] introduce Generative Adversarial Networks (GANs) [12] to optimize the rate-distortion-perception trade-off and achieve significant improvements in the perceptual quality. However, these methods inevitably introduce unpleasant visual artifacts, especially at extremely low bitrates.

Recently, diffusion models have shown great generation ability in image and video generation [13], [14], [15] and also have great potential to rival GANs. This encourages many researchers to develop various diffusion-based generative image compression methods [16], [17], [18], [19]. For example, CDC [19] utilizes the diffusion model as a conditional decoder and shows remarkable perceptual performance at the medium and high bitrates. However, when used at extremely low bitrates, this method fails to reconstruct pleasing results with reliable details. To overcome this limitation, several works have begun to explore diffusion generative models for extreme image compression [20], [21] by using pre-trained text-to-image diffusion models. Pan et al. [20] encode images as textual embeddings with extremely low bitrates, using pre-trained text-to-image diffusion models for the reconstruction. However, finding optimal textual embeddings involves iterative learning, which is time-consuming and resource-intensive. Lei et al. [21] directly transmit short text prompts and compressed image sketches, employing the pre-trained ControlNet [14] to produce reconstructions with high perceptual quality and semantic fidelity. However, due to the lack of color, texture, and other detailed information, the reconstructions are semantically close to the originals but poor in fidelity. Therefore, *how to develop an effective diffusion-based extreme generative compression method is worth further exploration.*

We observe that the key to the success of [20] and [21] lies in utilizing pre-trained text-to-image diffusion models as prior knowledge. This observation motivates us to further explore the application of such models in extreme image compression. To this end, we develop an end-to-end **Diff**usion-based **E**xtreme **I**mage **C**ompression (**DiffEIC**) model that effectively combines compressive variational autoencoders (VAEs) [23] with pre-trained text-to-image diffusion models. This model enables visually pleasing reconstructions that resemble the originals, even at extremely low bitrates. Similar to Lei et al. [21], we propose a conditional diffusion decoding module. This module employs the well-trained stable diffusion as a fixed decoder and injects external condition information via a trainable control module. This strategy fully exploits the powerful generative prior encapsulated in stable diffusion. However, rather than explicitly using short text prompts and image sketches as compressed representations, we develop a novel latent feature-guided compression module based on

This work are supported by the National Natural Science Foundation of China (62376208, 62088102HZ), the Natural Science Foundation of Shaanxi Province (No.2021JZ-04). (*Corresponding author: Chenyang Ge.*)

The authors are with the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: lizhiyuan2839@163.com; zhouyh@mail.xjtu.edu.cn; haowei@stu.xjtu.edu.cn; cyge@mail.xjtu.edu.cn; jiangjingwen@stu.xjtu.edu.cn)

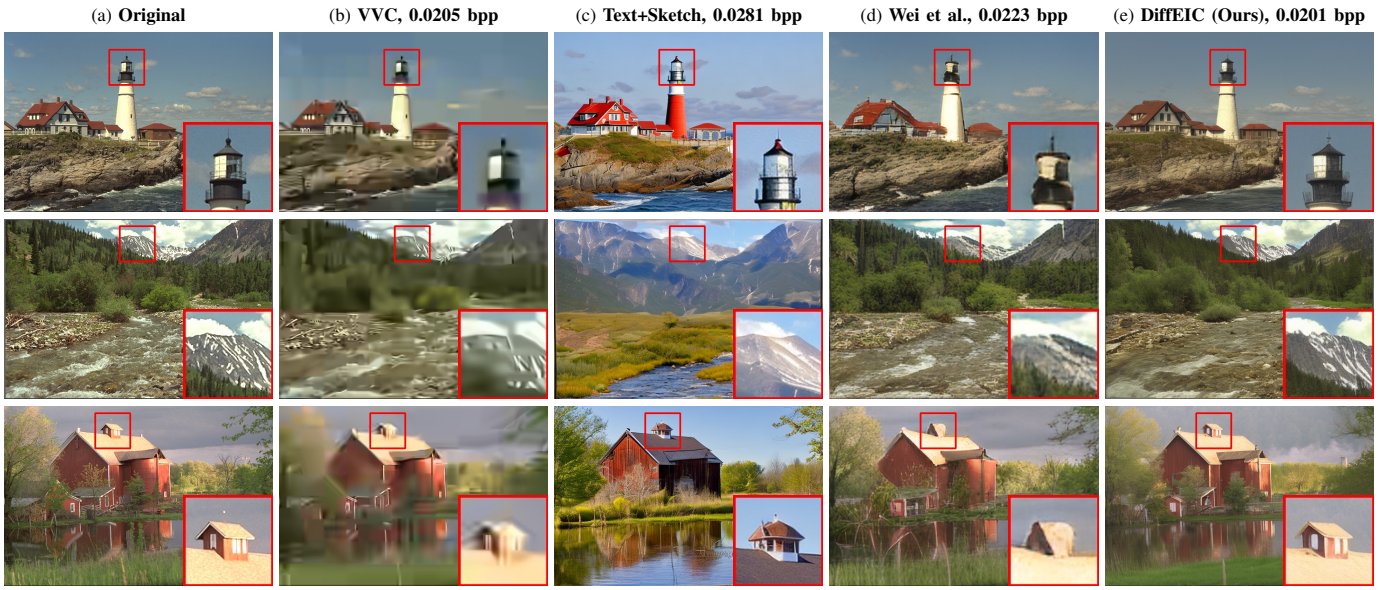


Fig. 1. Visual examples of the reconstructed results on the Kodak [22] dataset. The proposed DiffEIC produces much better results in terms of perception and fidelity. For example, the small attic is well reconstructed.

compressive VAEs to adaptively select information essential for reconstruction. This module follows the transform coding paradigm and initially decodes compressed information into content variables. In addition, we leverage spatial feature transform [24] layers to fuse external guidance for higher reconstruction fidelity. To achieve end-to-end training, we design a simple yet effective space alignment loss function to provide sufficient constraints for the latent feature-guided compression module. With the help of the mentioned components, the proposed extreme image compression framework achieves favorable results compared to state-of-the-art approaches, as demonstrated in Fig. 1.

In summary, the main contributions of this work are as follows:

1) To the best of our knowledge, we propose the first extreme image compression framework that combines compressive VAEs with pre-trained text-to-image diffusion models in an end-to-end manner.

2) We develop a latent feature-guided compression module to adaptively select information essential for reconstruction. By introducing external guidance, we effectively improve reconstruction fidelity at extremely low bitrates.

3) We propose a conditional diffusion decoding module that fully exploits the powerful diffusion prior contained in the well-trained stable diffusion to facilitate extreme image compression and improve realistic reconstruction. In addition, we design a simple yet effective space alignment loss to enable end-to-end model training.

The remainder of this paper is organized as follows. The related works are summarized in Section II. The proposed method is described in Section III. The experiment results and analysis are presented in Section IV and Section V, respectively. Finally, we conclude our work in Section VI.

II. Related Work

A. Lossy image compression

Lossy image compression plays a crucial role in image storage and transmission. Traditional compression standards, such as JPEG [25] and JPEG2000 [1], are widely used in practice. However, they tend to introduce block artifacts due to the lack of consideration of spatial correlation between image blocks. In recent years, learned image compression has made significant progress and achieved impressive rate-distortion performance [26], [19], [27]. The main success of these methods is attributed to the development of various transform networks and entropy models. Liu et al. [28] introduce a non-local attention module to improve transform networks. In [29], He et al. employ invertible neural networks (INNs) to mitigate the information loss problem. Zhu et al. [30] construct nonlinear transforms using swin-transformers, achieving superior compression performance compared to CNNs-based transforms. In [19], Yang et al. innovatively use conditional diffusion models as decoders. Furthermore, several methods [31], [6], [7] enhance performance by improving entropy models. For example, Minnen et al. [31] combine hierarchical priors with autoregressive models to reduce spatial redundancy within latent features. In [27], He et al. assume the redundancy in spatial dimension and channel dimension is orthogonal and propose a multi-dimension entropy model. Qian et al. [32] utilize a transformer to enable entropy models to capture long-range dependencies.

B. Extreme image compression

In some practical scenarios, such as underwater wireless communication, the bandwidth is too narrow to transmit the images or videos. To overcome this dilemma, extreme image compression towards low bitrates (e.g., below 0.1bpp) is urgently needed. Several algorithms [10], [11], [33], [34]

leverage GANs for realistic reconstructions and bit savings. In [10], Agustsson et al. incorporate a multi-scale discriminator to synthesize details that cannot be stored at extremely low bitrates. Mentzer et al. [34] explore normalization layers, generator and discriminator architectures, training strategies, as well as perceptual losses, achieving visually pleasing reconstructions at low bitrates. However, these approaches suffer from the unstable training of GANs and inevitably introduce unpleasant visual artifacts.

In addition, some approaches incorporate prior knowledge to achieve extreme image compression. Gao et al. [35] leverage invertible networks, significantly reducing information loss during the compression process. Similarly, Wei et al. [36] employ invertible prior and generative prior, indirectly achieving extreme compression by rescaling images with extreme scaling factors (i.e., $16\times$ and $32\times$). In [37], Li et al. employ physical prior to achieve extreme compression of underwater images. Jiang et al. [38] utilize the text descriptions of images as prior to guide image compression. However, the prior knowledge used in these methods has yet to meet the demands of extreme image compression effectively. To address this problem, some methods [20], [21] use more powerful pre-trained text-to-image diffusion models as prior knowledge. In [20], Pan et al. encode images into short text embeddings and then generate high-quality images with pre-trained text-to-image diffusion models by feeding the text embeddings. However, finding compressed textual embedding requires time-consuming iterative learning, which is not acceptable in practical applications. Lei et al. [21] directly compress the short text prompts and binary contour sketches in the encoded side, and then use them as input to the pre-trained text-to-image diffusion model for reconstruction in the decoded side. Due to the limited information provided by text and sketches, the reconstruction results show poor image fidelity compared to the original input. Hence, fully exploiting the potential of pre-trained diffusion models for extreme image compression remains challenging and investigated.

Unlike previous methods, we propose DiffEIC, a framework that efficiently incorporates compressive VAEs with pre-trained text-to-image diffusion models in an end-to-end manner. Leveraging the nonlinear capability of compressive VAEs and the powerful generative capability of pre-trained text-to-image diffusion models, our DiffEIC achieves both high perceptual quality and high-fidelity image reconstruction at extremely low bitrates.

C. Diffusion models

Inspired by non-equilibrium statistical physics [39], diffusion models convert real data distributions into simple, known distributions (e.g., Gaussian) through a gradual process of adding random noise, known as the diffusion process. Subsequently, they learn to reverse this diffusion process and construct desired data samples from noise (i.e., the reverse process). Denoising diffusion implicit models (DDPM) [16] improves upon the original diffusion model and has profoundly influenced subsequent research. The latent diffusion model (LDM) [13] significantly reduces computational costs by performing diffusion and reverse steps in the latent space. Stable diffusion is a widely used large-scale implementation of LDM. Owing to their flexibility, tractability, and superior generative capability, diffusion models have achieved remarkable success in various vision tasks.

Due to the complexity of the diffusion process, training diffusion models from scratch is computationally demanding and time-consuming. To address this problem, some algorithms [14], [40], [41] introduce additional trainable networks to inject external conditions into fixed, pre-trained diffusion models. This strategy simplifies the exhaustive training from scratch while maintaining the robust capability of pre-trained diffusion models. In [14], Zhang et al. employ pre-trained text-to-image diffusion models (e.g., stable diffusion) as a strong backbone with fixed parameters and reuse their encoding layers for controllable image generation. Similarly, Mou et al. [40] introduce lightweight T2I-Adapters to provide extra guidance for pre-trained text-to-image diffusion models. In [41], Lin et al. use the latent representation of coarse restored images as conditions to help the pre-trained diffusion models generate clean results. We note that the main success of these algorithms on image generation and restoration is due to the use of pre-trained diffusion models. The robust generative capability of such models motivates us to explore effective approaches for extreme image compression at low bitrates.

III. Methodology

In this section, we propose the DiffEIC for extreme image compression. As shown in Fig. 2, the proposed DiffEIC consists of two primary stages: image compression and image reconstruction. Specifically, the former stage aims to compress images and generate content-related variables. The latter stage is designed for decoding the content variables into reconstructed images. Furthermore, a space alignment loss is introduced to enable end-to-end model training and alleviate the misalignment between internal knowledge in stable diffusion and the content variables.

A. Image Compression with Compressive VAEs

As shown in Fig. 2(a), we propose a latent feature-guided compression module (LFGCM) based on compressing VAEs [23]. The encoding process, decoding process, and network details of LFGCM are introduced below.

1) Encoding Process: Given an input image x, we first obtain external guidance z_g with stable diffusion's encoder $\mathcal{E}$ as follows:

$$z_g = \mathcal{E}(x). \tag{1}$$

Then z_g is used to guide the extraction of the latent representation y and the side information z, sequentially, which can be expressed as:

$$y = \mathcal{N}_e(x, z_g),\ z = \mathcal{N}_{he}(y, z_g), \tag{2}$$

where $\mathcal{N}_e$ denotes the encoder network and $\mathcal{N}_{he}$ denotes the hyper-encoder network. Then we apply a hyper-decoder to draw a parameter ψ from the quantized side information $\hat{z}$:

$$\hat{z} = \mathcal{Q}(z),\ \psi = \mathcal{N}_{hd}(\hat{z}), \tag{3}$$

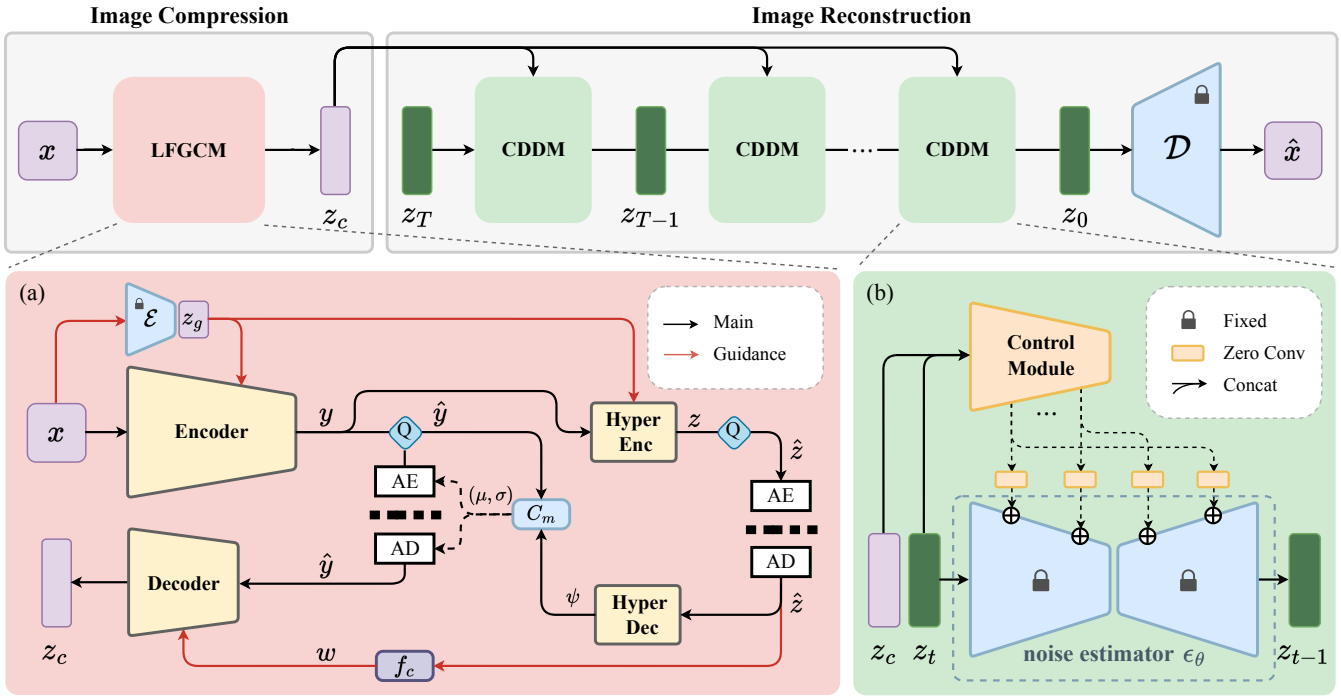


Fig. 2. The two-stage pipeline of the proposed DiffEIC. **Image Compression:** Initially, we leverage the latent feature-guided compression module (LFGCM) to adaptively select information essential for reconstruction and obtain z_c. The LFGCM maps images onto content variables using the framework of compressive VAEs. In addition, it employs the latent representations of images in the diffusion space as external guidance to dynamically modify intermediate features. **Image Reconstruction:** We leverage the conditional diffusion decoding module (CDDM) for realistic image reconstruction and obtain $\hat{x}$. The CDDM contains a trainable control module and a fixed noise estimator. Note that the control module and noise estimator are connected with zero convolutions (zero-initialized convolution layers).

where $\mathcal{N}_{hd}$ denotes the hyper-decoder and $\mathcal{Q}(\cdot)$ denotes the quantization operation, i.e., adding uniform noise during training and performing rounding operation during inference. Finally, the context model $\mathcal{C}_m$ uses ψ and the quantized latent representation $\hat{y} = \mathcal{Q}(y)$ to predict the Gaussian entropy parameters (μ, σ) for approximating the distribution of $\hat{y}$.

2) Decoding process: Given the quantized $\hat{y}$ and $\hat{z}$, we first use the information extraction network f_c to extract a representation w from $\hat{z}$, which can be expressed as:

$$w = f_c(\hat{z}). \tag{4}$$

The external guidance information, originally contained in z_g, is captured in w. This effectively compensates for the unavailability of z_g during the decoding process. Instead of directly reconstructing the original input image, we initially decode $\hat{y}$ into a content variable z_c:

$$z_c = \mathcal{N}_d(\hat{y}, w), \tag{5}$$

where $\mathcal{N}_d$ denotes the decoder network. The content variable z_c is further decoded in the subsequent image reconstruction stage using diffusion prior.

3) Network Details: Fig. 3 illustrates the network architecture of LFGCM, which integrates guidance components (the elements denoted by red arrows, SFT and SFT $Resblk$) to dynamically modulate intermediate feature maps. In our implementation, the information extraction network f_c has the same network structure as the hyper-decoder $\mathcal{N}_{hd}$, and we adopt context model $\mathcal{C}_m$ proposed by He et al. [27]. The guidance components first use a series of convolutions to resize the external feature G to the appropriate dimensions. Then the SFT layers [24] are employed to inject the network with external guidance information. Specifically, given an external feature G and an intermediate feature map F, a pair of affine transformation parameters (i.e., α for scaling and β for shifting) is generated as follows:

$$\alpha, \beta = \Phi_\theta(G), \tag{6}$$

where Φ_θ denotes a stack of convolutions. Then the tuned feature map F' can be generated by:

$$F' = SFT(F, G) = \alpha \otimes F + \beta, \tag{7}$$

where $\otimes$ denotes element-wise product.

B. Image Reconstruction with Diffusion Prior

As shown in Fig. 2(b), we propose a conditional diffusion decoding module (CDDM) to further decode the content variables. In this section, we introduce stable diffusion and the proposed CDDM, sequentially.

1) Stable Diffusion: Stable diffusion first employs an encoder $\mathcal{E}$ to encode an image x into a latent representation $z_0 = \mathcal{E}(x)$. Then z_0 is progressively corrupted by adding Gaussian noise through a Markov chain. The intensity of the added noise at each step is controlled by a default noise schedule β_t. This process can be expressed as follows:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \ t = 1, 2, \cdots, T, \tag{8}$$

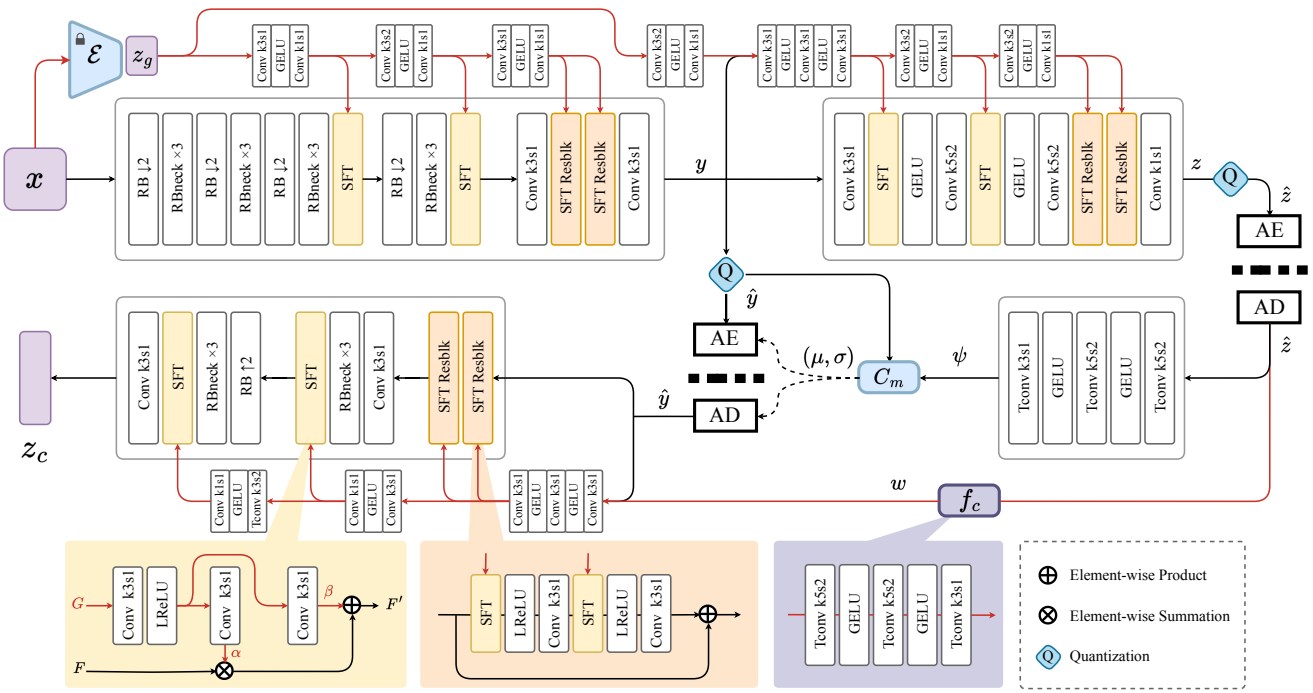


Fig. 3. The architecture of the proposed LFGCM. $Conv\ k3s1$ denotes convolution with 3×3 filters and stride 1. $Tconv\ k3s1$ denotes transposed convolution with 3×3 filters and stride 1. RB denotes residual block [42]. $RBneck$ denotes residual bottleneck block [42]. AE and AD denote arithmetic encoder and decoder, respectively. C_m denotes context model. The black and red arrows denote main and guidance flow, respectively.

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is a sample from a standard Gaussian distribution, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. The corrupted representation z_t approaches a Gaussian distribution as t increases. To iteratively convert z_T back to z_0, a noise estimator ϵ_θ with U-Net [43] architecture is learned to predict the added noise ϵ at each time step t:

$$\mathcal{L}_{sd} = \mathbb{E}_{z_0,c,t,\epsilon} \|\epsilon - \epsilon_\theta(z_t, c, t)\|^2, \tag{9}$$

where c denotes control conditions such as text prompts and images. After completing the iterative denoising process, a decoder $\mathcal{D}$ is used to map z_0 back into pixel space.

2) Conditional Diffusion Decoding Module: Given an input image x, a content variable z_c containing image information is obtained in the image compression stage. We then utilize the pre-trained stable diffusion as prior knowledge to further decode the content variable z_c. To inject the content information contained in z_c into stable diffusion, we introduce a control module with the same encoder and middle blocks as in the noise estimator ϵ_θ. However, we decrease the channel number of the control module to 20% of its original capacity for an effective balance between performance and efficiency. In addition, we increase the channel number of the first convolution layer to 6 to accommodate the concatenated input of the content variable z_c and the latent noise z_t. Through the control module, we can obtain a series of conditional features that contain content information and align with the internal knowledge of stable diffusion. Then, we add these conditional features to the encoder and decoder of the noise estimator ϵ_θ using 1×1 convolutions. Leveraging the powerful generative capability encapsulated in pre-trained stable diffusion, we can obtain a high perceptual quality reconstruction $\hat{x}$ even at extremely low bitrates. This reconstruction $\hat{x}$ contains the combined influence of both the content variable z_c and the initial noise latent z_T. Notably, we keep stable diffusion fixed during the training process to preserve its prior knowledge intact referring to [14].

C. Model Objectives

1) Noise Estimation Loss: Due to the external condition z_c introduced by proposed CDDM, Eq. (9) is modified as:

$$\mathcal{L}_{ne} = \mathbb{E}_{z_0,c,t,\epsilon,z_c} \|\epsilon - \epsilon_\theta(z_t, c, t, z_c)\|^2, \tag{10}$$

where text prompt c is set to empty.

2) Rate loss: We employ the rate loss $\mathcal{L}_{rate}$ to optimize the rate performance as:

$$\mathcal{L}_{rate} = R(\hat{y}) + R(\hat{z}), \tag{11}$$

where $R(\cdot)$ denotes the bit rates.

3) Space Alignment Loss: The space alignment loss forces the content variables to align with the diffusion space, providing robust constraints for LFGCM:

$$\mathcal{L}_{sa} = \|z_c - \mathcal{E}(x)\|^2. \tag{12}$$

In summary, the total loss of DiffEIC is defined as:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{rate} + \lambda_{sa} \mathcal{L}_{sa} + \lambda_{ne} \mathcal{L}_{ne}, \tag{13}$$

where λ_{sa} and λ_{ne} denote the weights for space alignment loss and noise estimation loss, respectively. λ is used to achieve a trade-off between rate and distortion.

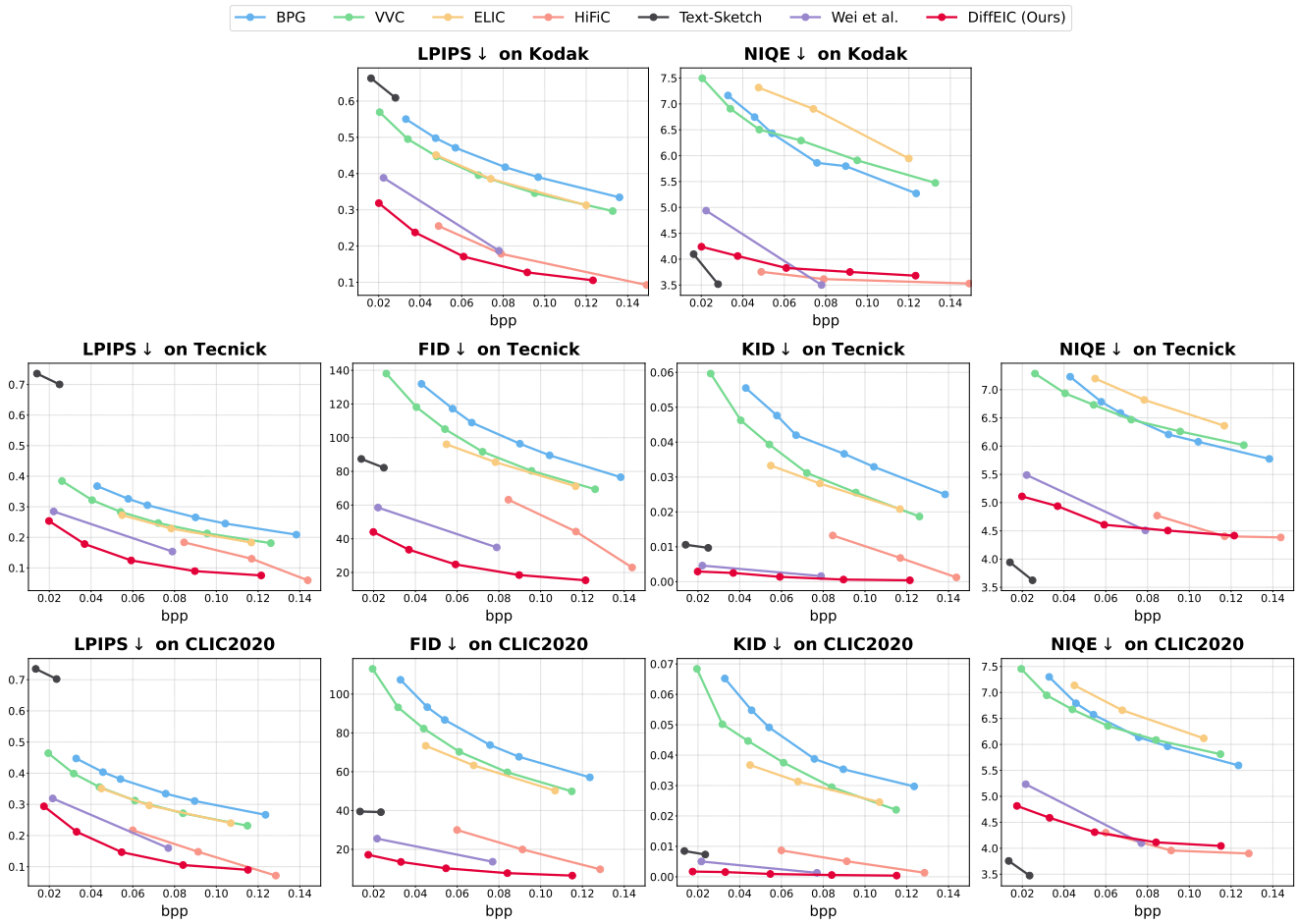


Fig. 4. Quantitative comparisons with state-of-the art methods in terms of perceptual quality (LPIPS↓ / FID↓ / KID↓ / NIQE↓) on the Kodak [22], Tecnick [44], and CLIC2020 [45] datasets.

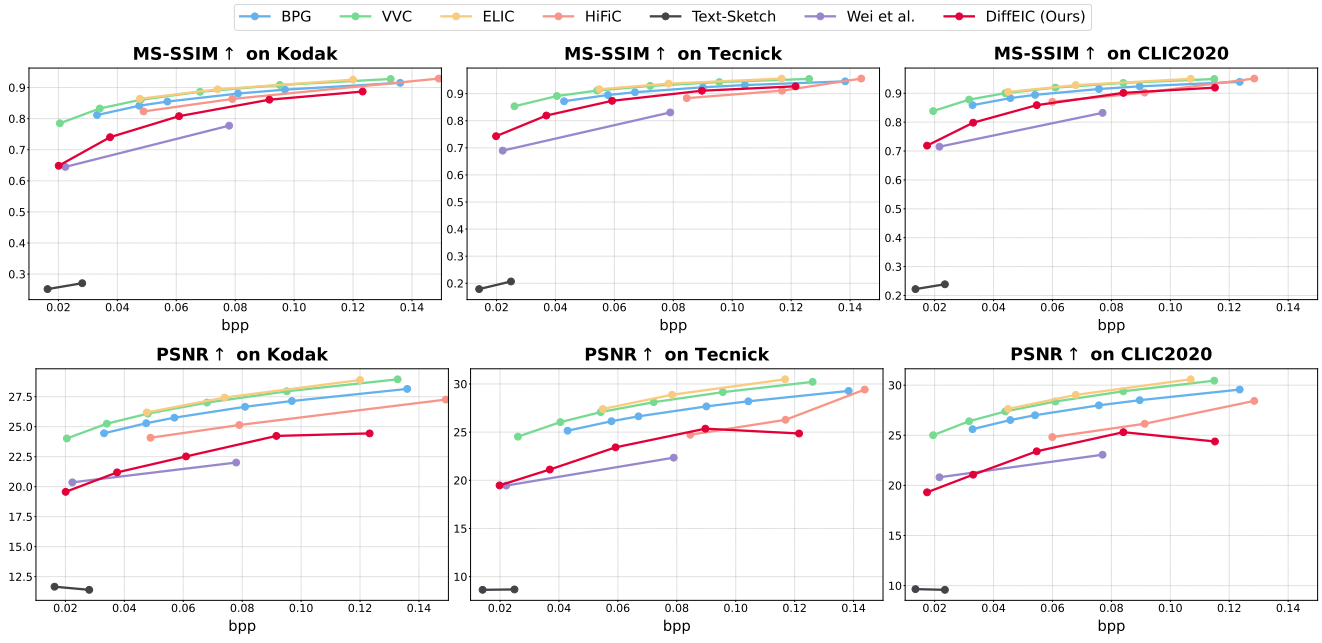


Fig. 5. Quantitative comparisons with state-of-the art methods in terms of pixel fidelity (MS-SSIM↑ / PSNR↑) on the Kodak [22], Tecnick [44], and CLIC2020 [45] datasets.

TABLE I
BD-RATE AND BD-METRIC ON BENCHMARK DATASETS: KODAK [22], TECNICK [44], AND CLIC2020 [45], WHERE LPIPS IS USED AS THE DISTORTION METRIC IN BD-METRIC. WE ADOPT BPG [2] AS THE ANCHOR. **BOLD** INDICATES THE BEST.

Methods	Kodak		Tecnick		CLIC2020	
	BD-rate	BD-LPIPS	BD-rate	BD-LPIPS	BD-rate	BD-LPIPS
VVC [3]	-27.93%	-0.0479	-30.93%	-0.0475	-32.00%	-0.0508
ELIC [27]	-25.64%	-0.0455	-34.31%	-0.0545	-31.99%	-0.0521
HiFiC [34]	-76.75%	-0.2400	-49.86%	-0.0979	-66.17%	-0.1607
Wei et al. [36]	-77.16%	-0.2459	-69.11%	-0.1482	-73.89%	-0.1885
DiffEIC (Ours)	**-87.03%**	**-0.2822**	**-78.70%**	**-0.1857**	**-82.07%**	**-0.2232**

IV. EXPERIMENTS

A. Experimental Settings

1) Model Training: We train DiffEIC on the **LSDIR** [46] dataset, which contains 84,991 high-quality training images. The images are randomly cropped to 512×512 resolution. In our experiments, we use Stable Diffusion 2.1-base[1] as the diffusion prior. We set the number of channels in the control module to be 20% of those in stable diffusion. We train our model in an end-to-end manner using Eq. (13), where λ_{sa} and λ_{ne} are set as 2 and 1, respectively. To achieve different coding bitrates, we choose λ from $\{1, 2, 4, 8, 16\}$. For optimization, we utilize Adam [47] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and set the learning rate to 1×10^{-4}. The training batch size is set to 4. Inspired by previous work [6], we first train the proposed DiffEIC with $\lambda = 1$ for 300K iterations, and then adapt them using target λ for another 200K iterations. We set the learning rate to 2×10^{-5} during the fine-tuning process. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. *The source code and trained models will be available on the authors' homepage after acceptance of the manuscript*.

2) Test Data: For evaluation, we use three commonly used benchmarks: Kodak [22], Tecnick [44], and CLIC2020 [45] datasets. The **Kodak** dataset contains 24 natural images with a resolution of 768×512. The **Tecnick** dataset contains 140 images with 1200×1200 resolution. The **CLIC2020** dataset has 428 high-quality images. For the Tecnick and CLIC2020 datasets, we center-crop the images to a size of 768×768 referring to [19]. Note that if the height or width of original images is smaller than 768, we resize them using bicubic interpolation.

3) Metrics: For quantitative evaluation, we employ several established metrics to assess the perceptual quality of results, including the Learned Perceptual Image Patch Similarity (**LPIPS**) [48], Fréchet Inception Distance (**FID**) [49], Kernel Inception Distance (**KID**) [50], and Naturalness Image Quality Evaluator (**NIQE**) [51]. Meanwhile, we employ the Peak Signal-to-Noise Ratio (**PSNR**) and Multi-Scale Structural Similarity Index (**MS-SSIM**) [52] to measure the fidelity of reconstruction results. Furthermore, the bits per pixel (bpp) is used to evaluate rate performance. Note that FID and KID are calculated on 256×256 patches according to [34]. Since the Kodak dataset is too small to calculate FID and KID, we do not report FID or KID results on it.

B. Comparisons With State-of-the-art Methods

We compare our DiffEIC with state-of-the-art learned image compression methods, including ELIC [27], HiFiC [34], Text+Sketch [21], and Wei et al. [36]. In addition, we compare with traditional image compression methods BPG [2] and VVC [3]. For BPG software, we optimize image quality and compression efficiency with the following settings: "YUV444" subsampling mode, "x265" HEVC implementation, "8-bit" depth, and "YCbCr" color space. For VVC, we employ the reference software VTM-23.0[2] with intra configuration.

1) Quantitative Comparisons: In Fig. 4, we illustrate the rate-perception curves for different methods over the three datasets. Traditional standards, BPG [2] and VVC [3], as well as the MSE-optimized approach ELIC [27], exhibit notable deficiencies in perceptual quality metrics compared to our DiffEIC. HiFiC [34] performs well at higher bitrates but is less effective at extremely low bitrates. The diffusion-based Text+Sketch [21] achieves the best NIQE scores, but shows considerable limitations in other metrics when compared to our DiffEIC. Due to indirectly achieving extreme image compression through extreme image rescaling, Wei et al. [36] generally underperform compared to our DiffEIC across all metrics. In comparison, our proposed DiffEIC (represented by the red lines) achieves the lowest values in LPIPS, FID, and KID. These low values suggest that DiffEIC excels at preserving the perceptual integrity of the images, producing reconstructions with minimal perceptual differences from the originals. In addition, Fig. 5 shows the rate-distortion curves on the three datasets. Traditional standards along with ELIC [27] achieve higher PSNR and MS-SSIM values than our DiffEIC. However, trading a little fidelity for significantly improved perceptual quality is considered worthwhile at extremely low bitrates. Compared to perception-oriented methods, i.e., HiFiC [34] and Wei et al. [36], our DiffEIC shows competitive performance in distortion metrics. Due to the lack of color, texture, and other information, Text+Sketch [21] scores lower on PSNR and MS-SSIM, indicating a significant difference between its reconstructions and the original images.

To better compare the R-D performance of these methods, we demonstrate the Bjontegaard metric (BD-metric) [53] in Table I. We use BPG [2] as the anchor and LPIPS as the distortion metric for calculating the BD-metric. Notably, we omit Text+Sketch [21] from this comparison because it focuses primarily on semantic similarity with the original image. As shown in Table I, with equivalent reconstruction quality,

[1]https://huggingface.co/stabilityai/stable-diffusion-2-1-base

[2]https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/master

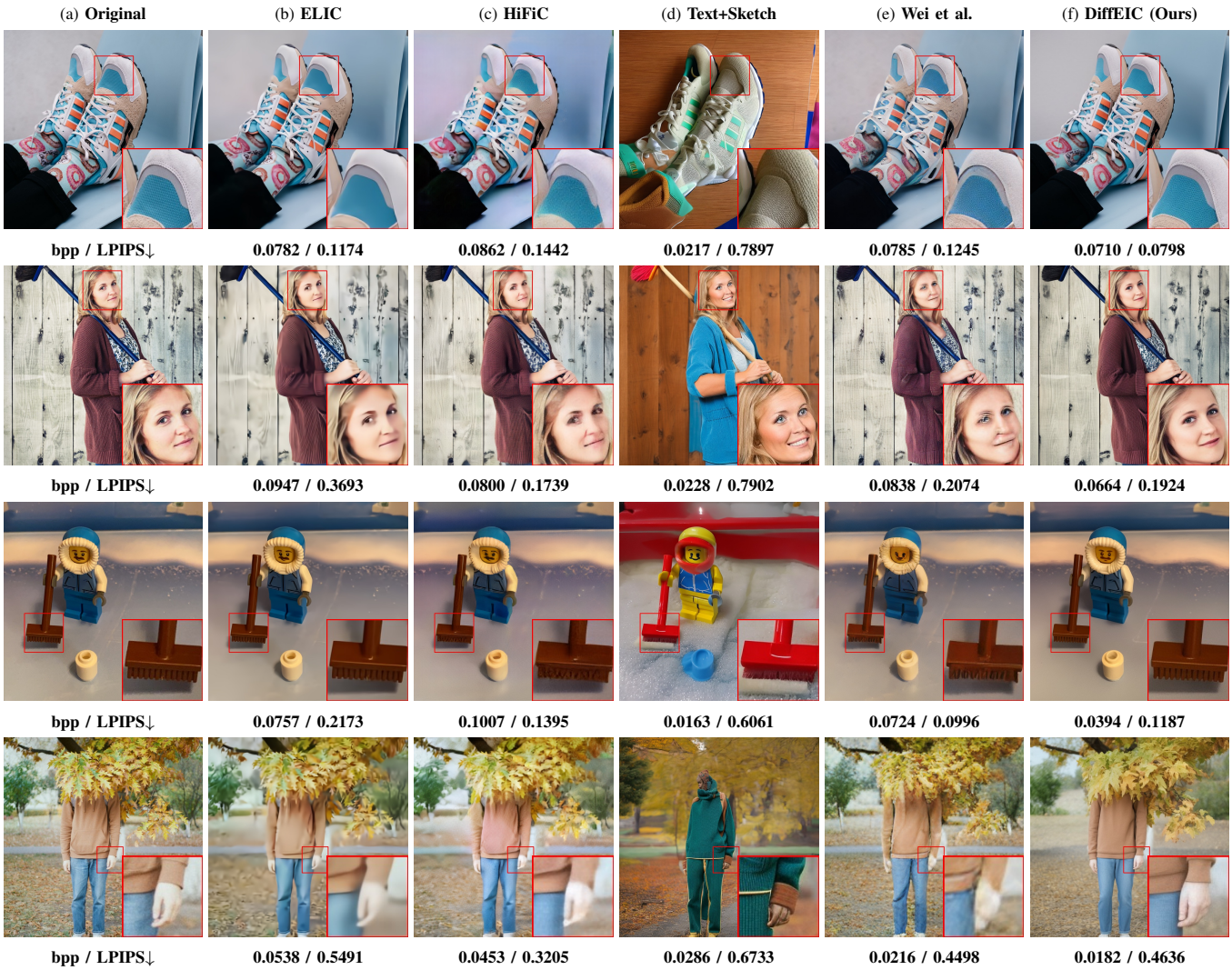


Fig. 6. Visual comparisons of the proposed DiffEIC framework with the MSE-optimized ELIC [27], the GANs-based HiFiC [34], the diffusion-based Text+Sketch [21], and Wei et al. [36] on the CLIC2020 [45] dataset. For each method, the bpp and LPIPS values are shown beneath images. Compared to other methods, our method produces more realistic and faithful reconstructions with lower bpp.

our DiffEIC achieves the highest bitrates saving of 87.03%, 78.70%, and 82.07% on the three datasets, respectively. Meanwhile, at equivalent bitrates, our DiffEIC achieves the greatest LPIPS improvement of 0.2822, 0.1857, and 0.2232 on the three datasets, respectively.

2) Qualitative Comparisons: Fig. 6 shows visual comparisons among the evaluated methods at extremely low bitrates. The ELIC method [27] generates over-smooth results due to the rate-distortion oriented optimization. By using adversarial learning, HiFiC [34] generates realistic results but introduces visual artifacts (see Fig. 6(c)). Although the diffusion-based method Text+Sketch [21] significantly reduces the bpp, it tends to reconstruct semantically reasonable results, which sacrifices the pixel fidelity. The recent method [36] employs a pre-trained VQGAN [54] as generative prior to achieve extreme image rescaling. However, the reconstruction results in Fig. 6(e) contain obvious artifacts. In contrast, the proposed DiffEIC generate clearer and more realistic results at extreme bitrates. For example, our DiffEIC reconstructs facial details with the highest quality (see the second row of Fig 6).

V. Analysis and Discussions

To better analyze the proposed method, we perform ablation studies and discuss its limitations.

TABLE II
BD-rate and BD-metric on CLIC2020 [45] dataset, where LPIPS is used as the distortion metric in BD-metric. We adopt DiffEIC(Ours) as the anchor.

Variants	BD-rate	BD-LPIPS
W/o latent feature guidance	13.19%	0.0151
W/o diffusion prior	35.77%	0.0566

A. Effectiveness of Latent Feature Guidance

In the proposed LFGCM, the latent representation of the input image in the diffusion space serves as external guidance. This guidance is utilized to dynamically modify intermediate

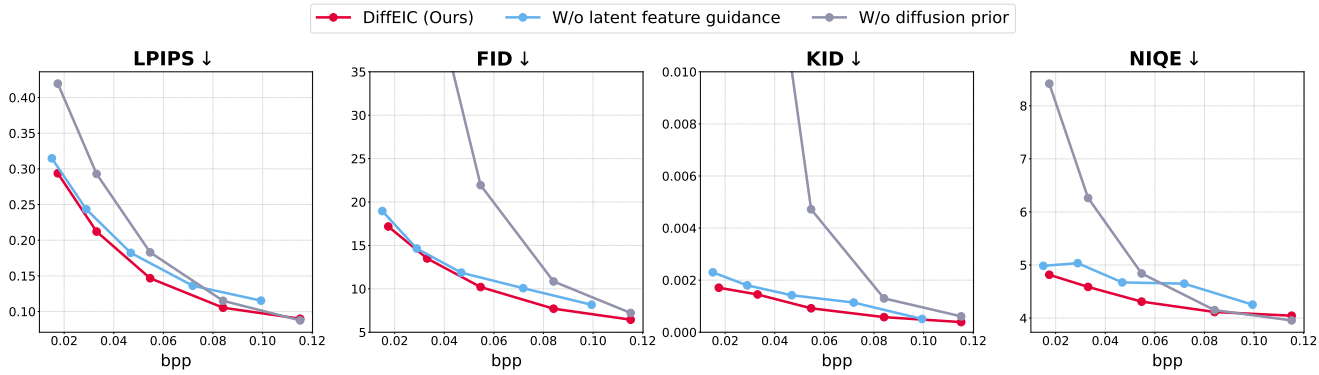


Fig. 7. Ablation of the latent feature guidance and diffusion prior on the CLIC2020 [45] dataset.

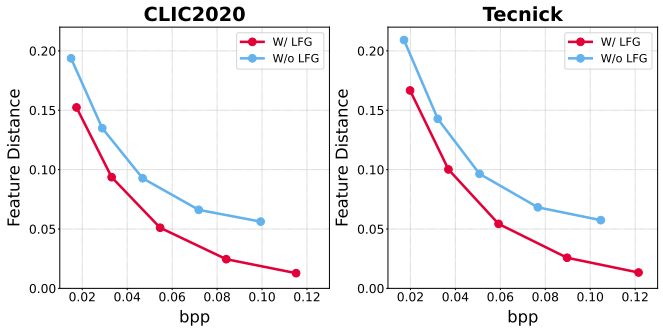


Fig. 8. Euclidean distance between content variables and their corresponding latent image representations on the CLIC2020 [45] dataset and the Tecnick [44] dataset. LFG denotes latent feature guidance.

Fig. 9. Effectiveness of the Latent Feature Guidance (LFG) for image compression.

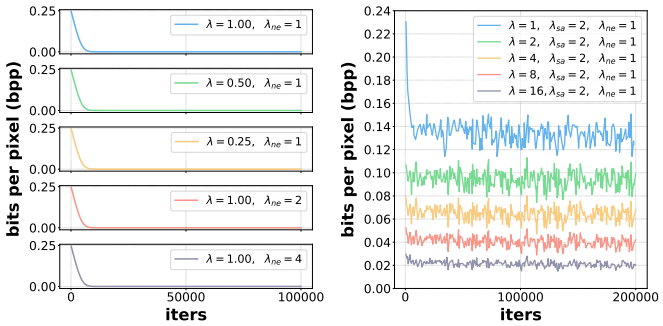
Fig. 10. Effectiveness of the diffusion prior for image compression. The bpp and LPIPS↓ scores are shown beneath images.

features through SFT layers. To validate the effectiveness of this strategy, we remove the guidance components and retrain from scratch using the same experimental settings. As shown in Fig. 7, removing the guidance components results in an obvious degradation in performance metrics. Table II demonstrates that without the guidance components, the bitrates increase by 13.19% at the same reconstruction quality and the LPIPS value increases by 0.0151 at equivalent bitrates.

Fig. 8 shows the Euclidean distance between content variables and their corresponding latent image representations in the diffusion space. It is clear from the results that using latent feature guidance (LFG) significantly reduces the Euclidean distance. This suggests that content variables are more closely aligned with the diffusion space, enabling the subsequent diffusion process to capture more accurate details. For example, the facial details in Fig. 9 are reconstructed more consistently with the original image when using LFG.

B. Effectiveness of Diffusion Prior

The proposed CDDM employs pre-trained stable diffusion as prior to decode content variables. Leveraging the powerful generative capability encapsulated in stable diffusion, our method can produce highly realistic reconstructions at extremely low bitrates. To evaluate the effectiveness of the diffusion prior, we compare the proposed method with the variant model without using diffusion prior. Specifically, we directly decode content variables using the decoder $\mathcal{D}$ of stable diffusion. As shown in Fig. 7, our DiffEIC exhibits a slower degradation in performance metrics as bpp decreases. In stark contrast, the variant without diffusion prior suffers from a sharp decline in performance over all perceptual metrics. As Table II demonstrates, lacking the diffusion prior leads to a 35.77% increase in bitrates for the same reconstruction quality, and the LPIPS value increases by 0.0566 at equivalent bitrates. This indicates that the diffusion prior is crucial for high perceptual reconstruction at extremely low bitrates. Further elucidating this point, Fig. 10 compares the reconstructions by our DiffEIC and the variant without the diffusion prior at



(a) W/o space alignment loss (b) W/ space alignment loss

Fig. 11. Effectiveness of the space alignment loss for end-to-end training.

various bitrates. As bpp decreases, the reconstructions by the variant without diffusion prior become increasingly smooth. Conversely, our DiffEIC, by fully utilizing the powerful generative capability encapsulated in the diffusion prior, is able to produce reasonable and visually pleasing details even at extremely low bitrates.

C. Effectiveness of Space Alignment Loss

Initially, we employ the rate loss $\mathcal{L}_{rate}$ and the noise estimation loss $\mathcal{L}_{ne}$ to train DiffEIC. Similar to Eq. (13), the loss function is formulated as $\mathcal{L}_{init} = \lambda\mathcal{L}_{rate} + \lambda_{ne}\mathcal{L}_{ne}$. As illustrated in Fig. 11(a), we observe that the bits per pixel (bpp) curves invariably converge to zero during training, regardless of the selected values for λ and λ_{ne}. This indicates that the initial loss $\mathcal{L}_{init}$ is insufficient for the end-to-end training of DiffEIC. We attribute this phenomenon to the noise estimation loss being unable to provide effective constraints for LFGCM. To address this problem, we design the space alignment loss to offer more stringent and robust constraints for LFGCM. As shown in Fig. 11(b), adding the space alignment loss enables effective end-to-end training and rate control of the model. Furthermore, this loss ensures that content variables align more accurately with the diffusion space, which effectively alleviates the burden of aligning the internal knowledge in stable diffusion with the content variables. This improved alignment contributes to enhanced reconstruction quality, as noted in Section V-A.

D. Limitation

Although the proposed DiffEIC framework achieves favorable reconstructions at extremely low bitrates, it still has some limitations. 1) While text is an important component in pre-trained text-to-image diffusion models, its application has not yet been explored within our framework. The work of Text+Sketch [21] demonstrates the powerful ability of text in extracting image semantics, encouraging us to further leverage text to enhance our method in future work. 2) Due to using a diffusion model as the decoder, the DiffEIC framework requires more computational resources and longer inference times compared to other VAEs-based compression methods. Using more advanced sampling methods may be a solution to alleviate the computing burden.

VI. Conclusion

In this paper, we propose a novel extreme image compression framework, named DiffEIC, which combines compressive VAEs with pre-trained text-to-image diffusion models to achieve realistic and high-fidelity reconstructions at extremely low bitrates (below 0.1 bpp). Specifically, we introduce a latent feature-guided compression module based on compressive VAEs. This module compresses images and initially decodes them into content variables. The use of latent feature guidance effectively improves reconstruction fidelity. We then propose a conditional diffusion decoding module to decode these content variables. Leveraging the powerful generative capability of pre-trained text-to-image diffusion models, our DiffEIC is able to produce reconstructions with rich, realistic details. In addition, a simple yet effective space alignment loss is proposed to enable end-to-end training. Extensive experiments demonstrate the superiority of DiffEIC and the effectiveness of proposed modules.

References

[1] David S Taubman, Michael W Marcellin, and Majid Rabbani. Jpeg2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11(2):286–287, 2002.

[2] Fabrice Bellard. Bpg image format. https://bellard.org/bpg/.

[3] 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*, 31(10):3736–3764, 2021.

[4] Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14388–14397, 2023.

[5] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17492–17501, 2022.

[6] David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 3339–3343. IEEE, 2020.

[7] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14771–14780, 2021.

[8] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, pages 2922–2930. PMLR, 2017.

[9] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[10] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

[11] Shubham Dash, Giridharan Kumaravelu, Vijayakrishna Naganoor, Suraj Kiran Raman, Aditya Ramesh, and Honglak Lee. Compressnet: Generative compression at extremely low bitrates. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2314–2322. IEEE, 2020.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[15] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] Lucas Theis, Tim Salimans, Matthew D Hoffman, and Fabian Mentzer. Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*, 2022.

[18] 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.

[19] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *arXiv preprint arXiv:2209.06950*, 2022.

[20] Zhihong Pan, Xin Zhou, and Hao Tian. Extreme generative image compression by learning text embedding from diffusion models. *arXiv preprint arXiv:2211.07793*, 2022.

[21] Eric Lei, Yigit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text+ sketch: Image compression at ultra low rates. In *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023.

[22] Eastman Kodak Company. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/.

[23] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

[24] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018.

[25] Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

[26] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[27] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5718–5727, 2022.

[28] Haojie Liu, Tong Chen, Peiyao Guo, Qiu Shen, Xun Cao, Yao Wang, and Zhan Ma. Non-local attention optimized deep image compression. *arXiv preprint arXiv:1904.09757*, 2019.

[29] Yueqi Xie, Ka Leong Cheng, and Qifeng Chen. Enhanced invertible encoding for learned image compression. In *Proceedings of the 29th ACM international conference on multimedia*, pages 162–170, 2021.

[30] Yinhao Zhu, Yang Yang, and Taco Cohen. Transformer-based transform coding. In *International Conference on Learning Representations*, 2021.

[31] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

[32] Yichen Qian, Xiuyu Sun, Ming Lin, Zhiyu Tan, and Rong Jin. Entroformer: A transformer-based entropy model for learned image compression. In *International Conference on Learning Representations*, 2021.

[33] Shoma Iwai, Tomo Miyazaki, Yoshihiro Sugaya, and Shinichiro Omachi. Fidelity-controllable extreme image compression with generative adversarial networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8235–8242. IEEE, 2021.

[34] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020.

[35] Fangyuan Gao, Xin Deng, Junpeng Jing, Xin Zou, and Mai Xu. Extremely low bit-rate image compression via invertible image generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[36] Hao Wei, Chenyang Ge, Zhiyuan Li, Xin Qiao, and Pengchao Deng. Towards extreme image rescaling with generative prior and invertible prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[37] Mengyao Li, Liquan Shen, Yufei Lin, Kun Wang, and Jinbo Chen. Extreme underwater image compression using physical priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1937–1951, 2022.

[38] Xuhao Jiang, Weimin Tan, Tian Tan, Bo Yan, and Liquan Shen. Multimodality deep network for extreme learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1033–1041, 2023.

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[40] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

[41] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[44] Nicola Asuni and Andrea Giachetti. Testimages: a large-scale archive for testing visual devices and basic image processing algorithms. In *STAG*, pages 63–70, 2014.

[45] George Toderici, Lucas Theis, Nick Johnston, Eirikur Agustsson, Fabian Mentzer, Johannes Ballé, Wenzhe Shi, and Radu Timofte. Clic 2020: Challenge on learned image compression, 2020. http://www.compression.cc.

[46] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1775–1787, 2023.

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[49] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[50] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[51] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[52] 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*, volume 2, pages 1398–1402. Ieee, 2003.

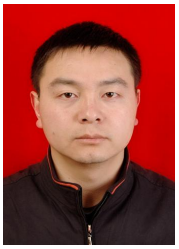
[53] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001.

[54] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

Zhiyuan Li is currently pursuing a master's degree with the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. He received his bachelor's degree from Xidian University in 2022. His research interests include image compression, image rescaling, and other visual problems.

Yanhui Zhou received the M. S. and Ph. D. degrees in electrical engineering from the Xi'an Jiaotong University, Xi'an, China, in 2005 and 2011, respectively. She is currently an associate professor with the School of Information and telecommunication Xi'an Jiaotong University. Her current research interests include image/video compression, computer vision and deep learning.

Hao Wei is currently a Ph.D. candidate with the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. He received his B.Sc. and M.Sc. degrees from Yangzhou University and Nanjing University of Science and Technology in 2018 and 2021, respectively. His research interests include image deblurring, image compression, and other low-level vision problems.



Chenyang Ge is currently an associate professor at Xi'an Jiaotong University. He received the B.A., M.S., and Ph.D. degrees at Xi'an Jiaotong University in 1999, 2002, and 2009, respectively. His research interests include computer vision, 3D sensing, new display processing, and SoC design.

Jingwen Jiang is currently pursuing a master's degree at Xi'an Jiaotong University. He received his bachelor's degree from Sichuan University in 2023. His research interests include image compression and video compression.