

Decouple Distortion from Perception: Region Adaptive Diffusion for Extreme-low Bitrate Perception Image Compression

Jinchang Xu¹ Shaokang Wang¹ Jintao Chen² Zhe Li² Peidong Jia¹
 Fei Zhao¹ Guoqing Xiang³ Zhijian Hao⁴ Shanghang Zhang^{1*} Xiaodong Xie¹

¹State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University ²School of Software & Microelectronics, Peking University,

³Alibaba Cloud Computing, Beijing, China ⁴State Key Discipline Laboratory of Wide Band-Gap Semiconductor Technology, School of Microelectronics, Xidian University

{jinchang-xu, feizhao}@pku.edu.cn, {skwang6272, cjt, 2401210335, 2101111573}@stu.pku.edu.cn
 xiangguoqing.xgq@alibaba-inc.com, haozhijian@xidian.edu.cn, {shanghang, donxie}@pku.edu.cn



Figure 1. **Visual comparison of decoded images.** (a) Original images with the user-specified RoI area highlighted by a green dashed box. (b) VAE compression method cannot achieve extreme-low bitrate considering distortion optimization. (c) PerCo sacrifices the consistency of key areas to obtain realism at extreme-low bitrate. (d) Traditional encoders have severe blocking effects at extreme-low bitrate. (e) The proposed method maintains the overall advanced perceptual quality and maintains the consistency of key areas.

Abstract

Leveraging the generative power of diffusion models, generative image compression has achieved impressive per-

*ceptual fidelity even at extremely low bitrates. However, current methods often neglect the non-uniform complexity of images, limiting their ability to balance global perceptual quality with local texture consistency and to allocate coding resources efficiently. To address this, we introduce the **Map-guided Masking Realism Image Diffusion Codec (MRIDC)**, designed to optimize the trade-*

* Corresponding Author. This work was supported by the National Natural Science Foundation of China (62476011), and the Alibaba Innovative Research Program.

off between local distortion and global perceptual quality in extreme-low bitrate compression. MRIDC integrates a vector-quantized image encoder with a diffusion-based decoder. On the encoding side, we propose a **Map-guided Latent Masking (MLM)** module, which selectively masks elements in the latent space based on prior information, allowing adaptive resource allocation aligned with image complexity. On the decoding side, masked latents are completed using the **Bidirectional Prediction Controllable Generation (BPCG)** module, which guides the constrained generation process within the diffusion model to reconstruct the image. Experimental results show that MRIDC achieves state-of-the-art perceptual compression quality at extremely low bitrates, effectively preserving feature consistency in key regions and advancing the rate-distortion-perception performance curve, establishing new benchmarks in balancing compression efficiency with visual fidelity. Our code can be found at <https://github.com/xjc97/mridc>.

1. Introduction

With advances in image capture and generation technologies, the volume of digital image content has surged, underscoring the need for efficient image compression techniques. Traditional codecs [1–3] and Variational Autoencoder (VAE)-based end-to-end compression approaches [4–7] minimize compression Distortion (D) under Rate constraints (R_c) by enhancing handcrafted operators and model architectures:

$$\min D, \quad \text{s.t. } R \leq R_c. \quad (1)$$

However, with the exponential growth of data volume and the emergence of extremely bandwidth-constrained applications, such as satellite communications [8, 9], the demand for extreme-low bitrate compression has become increasingly urgent. Traditional codecs and VAE-based methods struggle to achieve such low bitrates (e.g., 0.01 bpp) without incurring severe distortion.

Under such extreme bitrate constraints, the Rate-Distortion-Perception (RDP) theory [10] highlights that there is a bottleneck in minimizing the distortion of the image, which can be turned into maximizing the subjective quality of the image. More specifically, minimizing distortion means reducing the error between the reconstruction and the corresponding input image, while maximizing perceptual quality means that the distribution of the reconstructed image is close to the input image so that the user cannot tell which is which. Recent advancements in generative models have significantly improved the ability to generate high-realistic images [11–18]. Given the ability of these models to capture intricate textures and enhance overall perceptual quality, recent image compression work combining Generative

Adversarial Networks (GANs) [19–22] or Diffusion Models (DMs) [23–26] has achieved significant optimization of the perceptual quality of reconstructed images. By exploiting the rich image modeling priors inherent in generative models, these approaches allow for the reconstruction of images with minimal artifacts using extremely few code-words [19, 27, 28].

To achieve the balance of Distortion-Perception (DP) under bitrate constraints, existing work improves the flexibility of generative compression through fixed-step refinement [24], parameter-conditioned decoding [29, 30], and dual-branch mixing [22], which is essentially to implicitly or explicitly interpolate low-distortion and high-realism images. However, distortion can be measured not only on the whole image but also within a single image while considering the importance of different regions, which means refining the granularity of the distortion measure and thus decoupling it from the constraints of perceptual quality. Intuitively, keeping better consistency between the reconstruction and the input image in the Regions of Interest (RoI) region while relaxing the constraints in other regions by generating realistic content can obtain more satisfactory results. Especially at extreme-low bitrates, the adaptive allocation of limited bitrate resources (R) by generative compression is expected to obtain better DP frontiers.

With this in mind, we propose the *Map-guided-Masking Realism Image Diffusion Codec* (MRIDC) to achieve endogenous-adaptive perceptual image compression at extreme-low bitrate. In our framework, images are first encoded into codebook indices through latent coding, where the predefined vector quantization space ensures a controllable extreme-low bitrate. By observing the correlation between the importance of regions in the original image and the redundancy of elements in the latent space, we introduce the *Map-guided Latent Masking* (MLM) module on the coding side. The MLM module adaptively masks latent representations based on the prior RoI information and allocates limited bit resources to key areas, thereby maintaining the consistency of details of the corresponding area in the reconstructed image and the original image. On the decoding side, we propose the *Bidirectional Prediction Controllable Generation* (BPCG) module, which reconstructs the masked indices by leveraging information from neighboring elements. The explicit prediction completes the latent information of the non-RoI region and smoothes the coherence between it and the RoI region. The reconstructed latent constrains the denoising process of the diffusion model, resulting in a decoded image with low distortion of key details and optimal global perceptual quality. Experimental results demonstrate that MRIDC achieves state-of-the-art perceptual compression quality at extremely low bitrates, effectively preserving feature consistency in key regions while allocating latent representation resources

in a training-free manner. By leveraging the MLM and BPCG modules, MRIDC enables conditional compression based on various priors, ensuring higher consistency in regions of interest (RoI) while delivering exceptional perceptual quality, as illustrated in Figure 1.

In summary, our contributions are as follows:

- We introduce a novel diffusion-based image compression framework, MRIDC, which combines latent vector quantization with the powerful generative capabilities of diffusion models. MRIDC achieves high perceptual quality in compressed images at extremely low bitrates.
- On the encoding side, we design the Map-guided Latent Masking (MLM) module to adaptively allocate latent representation resources in a *training-free* manner, guided by various forms of Region of Interest (RoI) maps. MLM effectively preserves key area information, supporting low-distortion reconstruction within limited resources.
- On the decoding side, we develop the Bidirectional Prediction Controllable Generation (BPCG) module, which constrains the diffusion model to reconstruct masked latent elements, producing decoded images with high perceptual quality.

2. Related works

Generative Image Compression. Traditional image codecs typically rely on transformation, quantization, and entropy coding [1, 2]. Building on these principles, early studies applied VAE-based, end-to-end image compression optimized through a rate-distortion loss function [4, 31, 32], achieving improved rate-distortion (RD) performance over traditional codecs [7, 33–38]. Recent advances in generative models capable of producing highly realistic images have further enhanced perceptually optimized compression. For instance, some works leverage the discriminator in GAN-based architectures to supervise the perceptual quality of reconstructed images, driving the compression algorithm to retain more perceptual information during reconstruction [20, 22, 39]. Others use the denoising properties of diffusion models to post-process decoded images and enhance perceptual quality [40]. Additionally, several studies recover images directly from compressed latent features, incorporating guidance mechanisms to control the consistency of generated images [23, 41, 42]. The strong distributional modeling capabilities of generative models enable high perceptual quality image reconstructions even at extremely low bitrates [19, 28, 43]. However, existing methods largely overlook the adaptive allocation of coding resources under extreme compression conditions. Greater quality improvements can be achieved by targeting limited coding resources to Regions of Interest (RoI) within the image, especially at these lower bitrates.

Quality Balance in Generative Compression. According to Rate-Distortion-Perception (RDP) theory, optimiz-

ing distortion and perceptual quality under a fixed bitrate is inherently contradictory. To address this, existing methods introduce external parameters in the generation process to balance distortion against subjective quality. By considering the joint distribution of distortion and perception, loss-conditional training allows the decoder to reconstruct images from the bitstream with variable quality based on conditional parameters [29, 30]. Additionally, inspired by the concept of interpolating between distortion-optimized and perception-optimized images, some studies implement control parameters as interpolation weights [22] or through varying the number of perception refinement iterations [24, 26, 44].

Further research [45–49] suggests that balancing distortion and perceptual quality should account not only for the image as a whole but also for the varying complexity of specific regions within the image. However, these approaches are not well-suited to scenarios requiring extremely low bitrates. Adaptive optimization based on RoI can drive joint improvement in both distortion and perceptual quality under extreme compression, a critical factor overlooked in existing generative compression methods.

3. Method

MRIDC is designed to optimize the trade-off between local distortion and global perceptual quality for extreme-low bitrate image compression. Section 3.1 introduces the task definition, followed by descriptions of the RoI-guided Encoder and Diffusion-based Decoder in Sections 3.2 and 3.2.2. Finally, Section 3.3 details our training approach.

3.1. Preliminary

Based on RDP theory [10], lossy compression at extreme-low bitrates requires a balance between distortion and perception quality loss:

$$\min \mathcal{L}_{dp} = D + \lambda P, \quad \text{s.t. } R \leq R_c, \quad (2)$$

where R is the rate of the bitstream and R_c is the bitrate constraint, which is 0.05 bpp in extreme-low bitrate scenarios. D is the distortion of an image sample, P is the perception quality calculated based on the statistical distribution of the dataset, and λ is the trade-off parameter. Specifically, the distortion D is defined as:

$$D = \mathbb{E}[\Delta(X, \hat{X})], \quad (3)$$

where $X \sim p_X$ and $\hat{X} \sim p_{\hat{X}}$ are the set of input and reconstruction image pairs, and Δ is the full-reference distortion metric defined on image pairs (PSNR, MS-SSIM, LPIPS, etc.). The perception quality loss P is defined as:

$$P = d(p_X, p_{\hat{X}}), \quad (4)$$

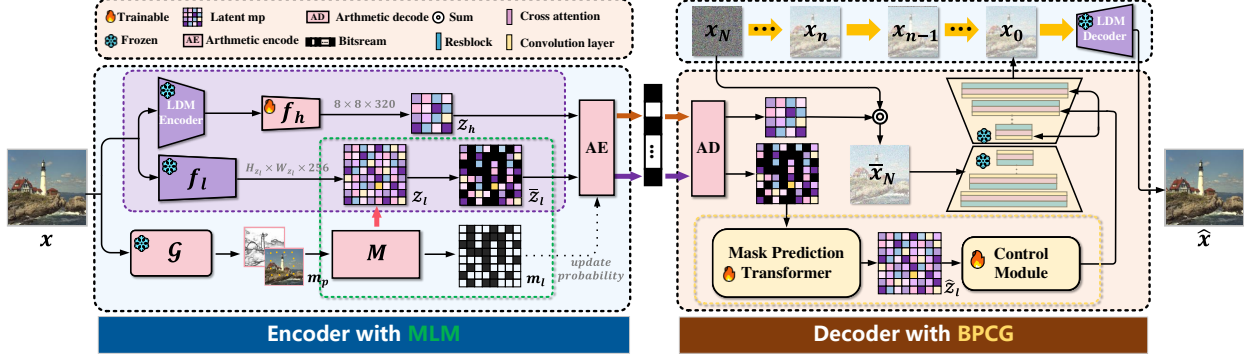


Figure 2. **Overview of the proposed diffusion-based generative image compression framework.** The framework consists of an encoding and decoding stage. In the encoding stage, the LDM encoder and f_h maps the RGB image into a VQGAN latent space, producing z_h . f_l map the input image into a latent diffusion space, producing z_l . The ROI map generator $\mathcal{G}$ produces control data (e.g., edge maps, bounding boxes) to generate a mask, applied to z_l to retain RoI information. z_h and $\bar{z}_l$ are then encoded via arithmetic coding. In decoding, the arithmetic decoder reconstructs z_h , and initial random noise x_N is added as input for diffusion denoising. MPT predicts masked elements from $\bar{z}_l$, outputting $\hat{z}_l$, while a control model guides denoising. After N steps, the noise-free latent x_0 is reconstructed into an image using the LDM decoder.

where $d(\cdot, \cdot)$ measures the divergence between image distributions (KID, FID, etc.).

Given input image X , existing generative lossy image compression obtains the reconstructed image $\hat{X}$ through an encoder $f(\cdot)$ and a decoder $g(\cdot)$. The bitrate R in Equation (2) is expressed by $f(X)$, which satisfies the constraint of R_c through the entropy model or Vector Quantization (VQ) coding. The quality optimization is determined by λ , which defines the tendency of the loss function and then trains the model parameters. However, existing compression considers the image sample $x \in X$ as a whole when considering Equation (3), ignoring the changes in image content and the importance of pixel-level distortion. Decoupling the distortion to finer granularity is expected to further improve the consistency in key areas while maintaining the global perceptual quality in Equation (4).

3.2. MRIDC: Adaptive Generative Compression

Motivated by the uneven importance of image regions, we propose MRIDC to further optimize the fidelity of the target region while maintaining the overall high perceptual quality according to a given RoI map M . Therefore, combining Equation (3) and (4), our target is formalized as:

$$\mathcal{L}_{dp|m} = \mathbb{E}[\Delta(X, \hat{X})|M] + \lambda_m d(p_X, p_{\hat{X}}), \text{ s.t. } R \leq R_c. \quad (5)$$

As shown in Figure 2, we use a low-level encoder f_l and a high-level encoder f_h to capture local details $z_l = f_l(X)$ and global semantic information $z_h = f_h(X)$ of the image at the encoding end. The local details are then adaptively masked conditioning to the given RoI map M , forcing more information to be stored in key areas and avoiding the allocation of redundant resources in non-RoI areas. At

the decoding end, the prediction model $\mathcal{T}$ first restores the masked elements in $f_l(X)$ to obtain coherent texture control information for the foreground and background. Then, the global semantic information $f_h(X)$ is used as a starting point to reconstruct the image using a diffusion generation model $g_m(\cdot)$ controlled by $f_l(X)$, that is:

$$\hat{X} = g_m(f_h(X), \mathcal{T}(f_l(X|M))). \quad (6)$$

3.2.1 RoI-guided Encoder with Latent Masking

On the encoding side, we first define the use of VQ to express the latent feature after image transfer. With a predefined size of the latent grid, we can achieve precise control of the bitrate after compression, thereby naturally satisfying the constraint of the extreme-low bit rate R_c in Equation (2). Specifically, we define the codebook as $\mathcal{C} := \{(k, e(k))\}_{k \in [K]}$, consisting of a finite set of pairs where each code k is associated with an embedding $e(k) \in \mathbb{R}^{n_z}$. Here, K denotes the codebook size, and n_z represents the dimensionality of each embedding.

For local detail information, we utilize the pre-trained VQ encoder [50] as f_l to map the input image to $z_l \in \mathbb{R}^{H_{z_l} \times W_{z_l} \times 256}$. Each vector $z_l^{i,j} \in \mathbb{R}^{n_{z_l}}$ in z_l is replaced by the nearest embedding in the corresponding codebook $\mathcal{C}_l$ of the pre-trained encoder through a vector quantization operation $\mathcal{Q}(\cdot)$:

$$\mathcal{Q}(z_l^{i,j}; \mathcal{C}_l) = \arg \min_{k \in [K]} \|z - e_k\|_2^2, \quad (7)$$

where i and j represent the position of the vector in the $H_{z_l} \times W_{z_l}$ dimension space. Based on the representation ability of the pre-trained model, z_l expresses the spatial information of the image in a lower dimension, and its distribution is consistent with the pixel domain information.

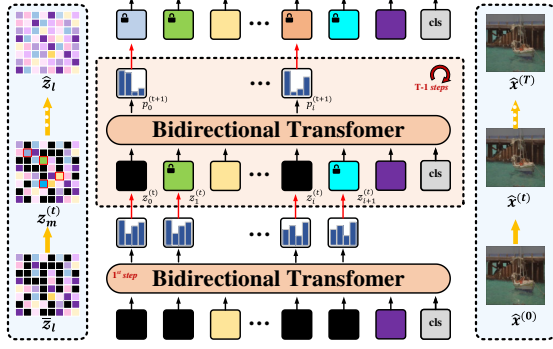


Figure 3. **MPT predict process.** MPT predicts the masked tokens iteratively based on the prediction probabilities at each step, with only those of high confidence being retained. The right images visualize the results decoded from the intermediate steps.

Therefore, for a user-defined RoI map generator $\mathcal{G}$, we design an MLM module to adaptively allocate latent representation resources. This allows the pixel-domain RoI map $m = \mathcal{G}(x)$ to further constrain the expression of z_l after being resized to the latent space, $m_l \in 0, 1^{H_{z_l} \times W_{z_l}}$, in a training-free manner. The embedding of the RoI region, indicated by $m_l^{i,j} = 1$, is retained to keep more latent information. The remaining regions are masked to avoid the impact of their redundant information on the reconstruction quality of the RoI region and further reduce the codewords required to express z_l :

$$\bar{z}_l = \mathbb{1}_{\{m_l^{i,j}=1\}} z_l + \alpha \times \mathbb{1}_{\{m_l^{i,j}=0\}} [\text{mask}], \quad (8)$$

where $\mathbb{1}_{\text{judge}}$ is the indicator function and is equal to 1 when judge is true, otherwise equal to 0. α is a ratio parameter for the non-RoI region masking. $[\text{mask}]$ is a special token to replace the masked tokens.

For global semantic information, we implement a lightweight VQ encoder as the f_h to generate the high-level latent representation $z_h \in \mathbb{R}^{8 \times 8 \times 320}$ from the Latent Diffusion Model (LDM) space. Based on the perceptual information captured by the LDM encoder, our f_h flexibly condenses the global information of different images and maps it to the fixed-size latent space. The parameters of f_h and the corresponding codebook $\mathcal{C}_g$ are trained through:

$$\mathcal{L}_{\text{vq}} = \|\text{sg}[z_h] - e_g\|_2^2 + \|\text{sg}[e_g] - z_h\|_2^2, \quad (9)$$

where $\text{sg}[\cdot]$ represents the stop-gradient operation, and e_g represents the embedding in the $\mathcal{C}_g$. The global information z_h and the masked local detail information $\bar{z}_l$ are losslessly encoded into the bitstream via arithmetic coding. The base bitrate of z_h and $\bar{z}_l$ is determined by their sizes, while the resulting bitrate can be controlled by adjusting the ratio α .

3.2.2 Diffusion-based Decoder with Latent Prediction

On the decoding side, $\bar{z}_l$ and z_h are first decoded from the bitstream and used to generate the reconstructed image $\hat{x}$.

Algorithm 1 MPT iterative prediction algorithm.

Input: $\bar{z}_l$, **Parameter:** T

Output: $\hat{z}_l$

- 1: $z_m^{(1)} \leftarrow \bar{z}_l$
 - 2: **for** $t = 1$ to T **do**
 - 3: **for** $i = 0$ to $\text{sizeof}(z_m^{(t)})$ **do**
 - 4: $z_i^{(t)} \leftarrow \arg \max p_i^{(t)}, \quad c_i \leftarrow p_i^{(t)}(z_i^{(t)})$
 - 5: **end for**
 - 6: $m_i^{(t+1)} \leftarrow \begin{cases} 1, & c_i < \text{sorted}_j(c_j)[n^{(t)}], \forall i \\ 0, & \text{otherwise} \end{cases}$
 - 7: Update $z_m^{(t+1)}$ using $m^{(t+1)}$ as per Equation (8)
 - 8: **end for**
 - 9: **return** $\hat{z}_l \leftarrow z_m^{(T)}$
-

To leverage $\bar{z}_l$ better, we apply a BPCG module, consisting of the Mask Prediction Transformer (MPT) $\mathcal{T}$ and the Control Module (CM). MPT predicts masked elements in $\bar{z}_l$ to produce a completed $\hat{z}_l$, ensuring regional continuity. As detailed in Algorithm 1, MPT uses a bidirectional transformer to iteratively predict the probabilities $p_i^{(t)}$ for each masked token by attending to other tokens, as follows:

$$p_i^{(t)} = p(z_i^{(t)} | \bar{z}_m^{(t)}) \quad (10)$$

where $\bar{z}_m^{(t)}$ be the masked latent map at iteration t , $\bar{z}_m^{(1)}$ equals to $\bar{z}_l$ and $\bar{z}_m^{(T)}$ equals to $\hat{z}_l$, with i as the position of a masked token. After producing $p_i^{(t)}$, MPT samples the predicted token $z_i^{(t)}$ as $\arg \max p_i^{(t)}$. Only high-confidence tokens are retained, while others are re-masked for subsequent iterations. The MPT loss function $\mathcal{L}_{\text{pred}}$ is defined as the cross-entropy between the ground-truth one-hot token and the predicted token:

$$\mathcal{L}_{\text{pred}} = -\mathbb{E} \left[\sum_{\forall i \in [1, \text{sizeof}(\bar{z}_l)]} \log p(z_i^{(t)} | \bar{z}_m^{(t)}) \right], \quad (11)$$

where $\text{sizeof}(\bar{z}_l) = W_{z_l} \times H_{z_l}$.

In general, the diffusion model assumes that the original latent feature x_0 is gradually superimposed with Gaussian noise to form a noisy latent:

$$x_n = \sqrt{\beta_n} x_0 + \sqrt{1 - \beta_n} \epsilon_n, \quad n = 1, 2, 3, \dots, N, \quad (12)$$

where $\epsilon_n \sim \mathcal{N}(0, \mathbf{I})$, $\beta_n = 1 - \gamma_n$ is decreasing parameter, and $\bar{\beta}_n = \prod_{i=1}^n \beta_i$. The noise estimator ϵ_θ is trained to denoise using the noisy latent x_n and initial x_0 at each step n . To leverage the global semantic information in z_h , we define the starting point of denoising as:

$$\bar{x}_N = \sqrt{\bar{\beta}_N} z_h + \sqrt{1 - \bar{\beta}_N} \epsilon_N. \quad (13)$$

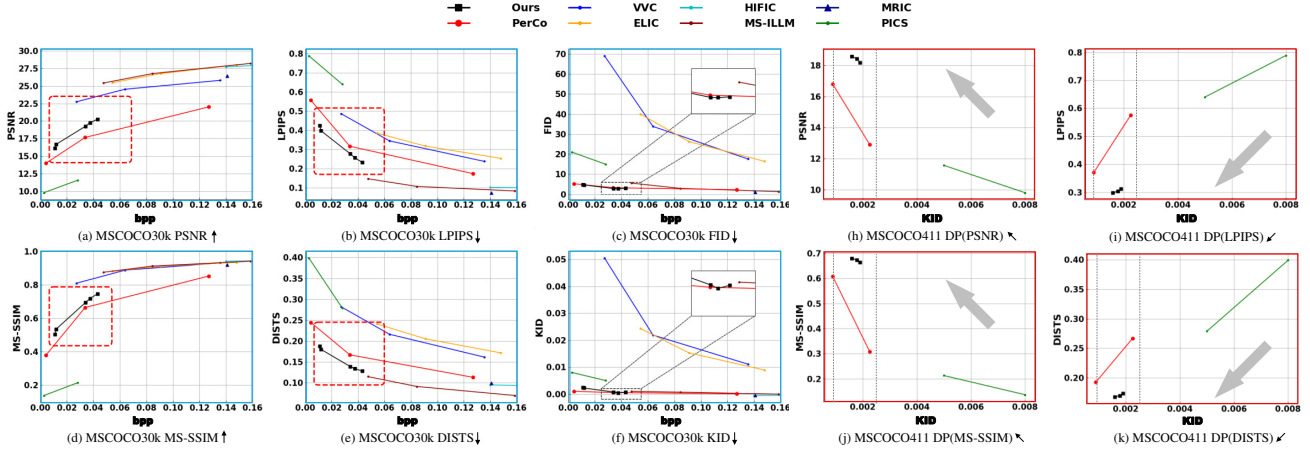


Figure 4. **Comparison with state-of-the-art methods.** The blue box compares local image distortion and global perceptual quality based on the bitrate. The red box shows the comparison of RoI region distortion based on global perceptual quality, where a lower and narrower horizontal axis range indicates better and more stable perception quality. The direction of the arrow indicates better performance.

During noise estimation, $\hat{z}_l$ is injected into the middle layer of ϵ_θ via CM, which uses ControlNet [51] as its backbone. The diffusion process is trained with a loss:

$$\mathcal{L}_{\text{dg}} = \mathbb{E}_{\hat{z}_l, z_h, n, \epsilon} \|\epsilon - \epsilon_\theta(\hat{z}_l, n, z_h)\|^2 + \lambda_{\text{lpips}} \mathcal{L}_{\text{lpips}}(x, \hat{x}), \quad (14)$$

where $\mathbb{E}_{\hat{z}_l, z_h, n, \epsilon} \|\epsilon - \epsilon_\theta(\hat{z}_l, n, z_h)\|^2$ quantifies noise estimation accuracy. LPIPS [52] is calculated between the original image x and LDM-decoded $\hat{x}$ to constrain subjective distortion, balanced by λ_{lpips} .

3.3. Training

The entire compression algorithm is trained in an end-to-end manner with an overall loss of:

$$\mathcal{L} = \mathcal{L}_{\text{vq}} + \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{dg}}, \quad (15)$$

which supervises the training of f_h , MPT, and control modules in the process, while other parts keep the pre-trained parameters frozen.

4. Experiments

4.1. Experimental Setup

Implementation Details. For implementation, we utilize the pre-trained VQ encoder [50] as f_l and implement a lightweight VQ encoder as f_h , whose detailed structure is described in the supplementary material. The bidirectional model for predicting masked latent in the decoder is fine-tuned from [50]. The generation process combines ControlNet [51] and a pre-trained Stable Diffusion v2.1 [14] latent diffusion model. During training, we use the Adam optimizer with a learning rate of 1×10^{-5} , a batch size

fixed at 16. All training samples are randomly cropped to 512×512 patches. For the definition of RoI area, we use Edge-map (Edge), Bounding box (Bbox), and global random (Rand). Edge is generated by RCF [53], and Bbox is taken from dataset annotation or user specification.

Metrics. To quantify compressed image bitrate, we use bits-per-pixel (bpp). For local image distortion, we use PSNR and MS-SSIM [54] to measure the objective distortion and use LPIPS [52] and DISTS [55] to measure the subjective distortion. For global perception quality, we employ FID [56] and KID [57]. To validate our balance in DP adaptation, we also report objective and subjective distortion metrics for each RoI region based on the overall FID and KID perception quality.

Datasets. We train our models on OpenImages-v6 [58], comprising nearly 1.7 million images. Evaluation of image quality is conducted on the MSCOCO30K [59] datasets, with all samples resized to a resolution of 512×512 , consistent with the evaluation settings used by [23]. To verify the DP adaptation effect, we randomly sampled 411 images from MSCOCO30k according to the appropriate RoI ratio of the whole image, marked as MSCOCO411. We also conducted experiments on the Kodak [60] dataset and the results are in the supplementary material.

Comparison Methods. We compare our method across three categories of image compression techniques: traditional codecs, VAE-based image compression, and generative image compression. From traditional codecs, we include VTM [61] for its advanced codec performance. The VAE-based image compression algorithm is represented by ELIC [33]. In generative image compression, we compared GAN-based and diffusion models. GAN-based methods include HiFiC [20], MS-ILLM [39], and MRIC [29], among which MRIC is the current representative of DP balancing

Table 1. **Ablation results of z_h .** MS stands for MS-SSIM. FID and KID are only calculated based on the entire image. KID is $\times 10^3$ for clarity of precision.

Abla.	Region	Metrics					
		PSNR $\uparrow$	MS $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	KID $\downarrow$
w/o z_h	Image	18.0	0.68	0.32	0.17	55.1	1.08
	RoI	17.0	0.64	0.35	0.19	-	-
w/ z_h	Image	20.3	0.72	0.24	0.13	43.9	0.14
	RoI	18.9	0.69	0.29	0.17	-	-

Table 2. **Ablation results of RoI map type and masking ratio on global perception quality.** *Low* represents the bitrate corresponding to the small size z_l , and *High* represents the bitrate corresponding to the large size z_l . KID is $\times 10^3$ for clarity of precision.

Rate	Metric	RoI Type and Masking Ratio							
		Bbox			Edge			Rand	
Low	FID $\downarrow$	48.2	50.7	54.8	52.1	58.6	66.8	51.0	55.3
	KID $\downarrow$	0.56	1.48	1.95	1.06	2.36	3.03	1.19	2.13
High	FID $\downarrow$	44.8	46.1	50.2	46.6	52.4	58.2	46.7	50.7
	KID $\downarrow$	1.87	1.41	1.66	2.10	2.26	3.26	1.50	1.61

based on the entire image. For diffusion model-based methods, we compare with PICS [23] and PerCo [27], among which PerCo is the representative of current extreme-low bit rate compression. Please refer to the supplementary material for the specific comparison implementation.

4.2. Comparison with State-of-the-Art

In the main results, we implemented two sizes of z_l in Equation (8), $32 \times 32 \times 256$ and $16 \times 16 \times 256$, corresponding to two basic bitrate points 0.04bpp and 0.01bpp. We further used Edge as the RoI mask, and took the mask rate α in the equation from 20%, 30%, and 40%, obtaining three variable bitrate results, which finally corresponded to the six MRIDC bitrate points shown in Figure 4 marked by the blue box. Our observations show that all three diffusion-based methods exhibit higher distortion but significant gains in perceptual quality relative to other approaches. Diffusion-based methods yield more realistic reconstructions than alternative methods. Our method, MRIDC, achieves a perceptual quality similar to PerCo while offering superior distortion quality even without enabling the RoI function.

The curve marked by the red box in Figure 4 compares the RoI area DP comparison results on MSCOCO411. With the RoI enabled, MRIDC demonstrates a marked PSNR improvement in the RoI region compared to PICS and PerCo at comparable KID, indicating that MRIDC can better balance distortion and perception than existing methods.

4.3. Ablation Study

We demonstrate the effectiveness of the design in the MRIDC encoder-decoder through ablation experiments, which are conducted on MSCOCO411. The default settings are based on a z_l of $32 \times 32 \times 256$ size and a 30% Edge mask.

Effect of Global Semantic Information. We evaluated the impact of the global semantic information z_h output by the

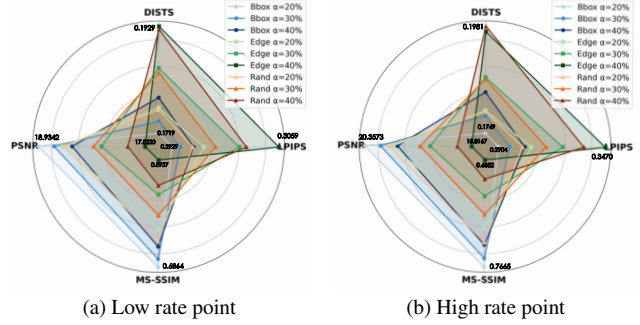


Figure 5. **Ablation results of RoI map type and masking ratio on RoI region distortion.**

Table 3. **Ablation results of MPT.** MS stands for MS-SSIM. FID and KID are only calculated based on the entire image.

Region	Status	Metrics					
		PSNR $\uparrow$	MS $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	KID $\downarrow$
w/o MPT	Image	16.4	0.56	0.41	0.19	73.2	3e-3
	RoI	15.7	0.53	0.43	0.23	-	-
w/ MPT	Image	19.3	0.67	0.27	0.14	49.4	4e-6
	RoI	17.9	0.63	0.32	0.17	-	-

encoder. As shown in Table 1, the removal of z_h leads to a significant loss in global perception, as well as distortion in both global and RoI regions. This suggests that the proposed global information plays a crucial role in providing essential starting information for the decoder denoising process. The slight bitrate overhead introduced by z_h is vital for maintaining quality and cannot be effectively compensated by the prior of the generative model.

Mask Type and Mask Ratio α . In addition to the aforementioned default configuration, we tested different mask ratios under different map forms. Table 2 illustrates the variation in global perceptual quality as the mask changes. It is evident that the Bbox has the least impact on global quality across different map forms, as it preserves the complete information of a larger area. Figure 5 further shows the distortion change trend corresponding to the RoI area, where the quality naturally decreases as the mask ratio increases.

Ablation Study on MPT. The ablation of MPT involves directly feeding the masked latent $\tilde{z}_l$ to the CM as control. As shown in Table 3, both global perception quality and local distortion are compromised, with a more pronounced decrease in global perceptual quality. This highlights that the proposed BPCG, consisting of MPT and CM, preserves global quality by maintaining coherence between the RoI and non-RoI regions, which cannot be substituted by the predictive capability of the diffusion model alone.

4.4. Complexity Analysis

Considering the complexity of iterative diffusion, we compared the encoding and decoding times of the proposed MRIDC with existing diffusion-based image compression

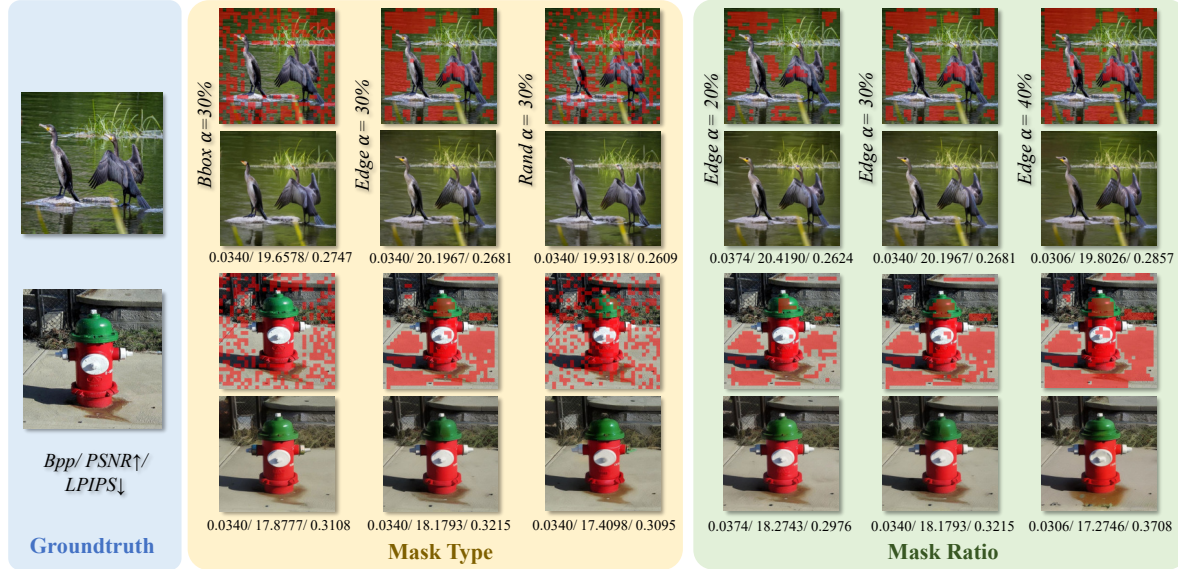


Figure 6. Visualization results of different masking methods and ratios.

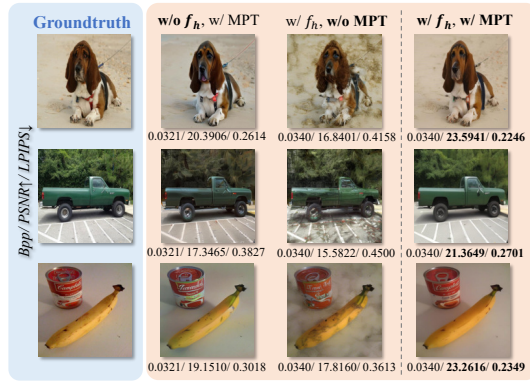


Figure 7. The subjective results of ablating f_h and MPT. Without the proposed module, the color, texture, and details of the image are significantly degraded, which validates the effectiveness of the proposed modules in preserving global image quality.

algorithms. The encoding complexity of MRIDC primarily arises from the latent masking operation, while the decoding complexity is mainly due to mask prediction. Despite this, the overall processing time remains shorter than that of PICS as shown in Table 4, as PICS relies on text guidance and requires traversal of the optimal text expression. Although the complexity of MRIDC is slightly higher than that of PerCo, it achieves superior image quality.

4.5. Qualitative Comparisons

Figure 1 illustrates the quality difference between the proposed MRIDC and the comparison method on the decoded image, with particular emphasis on the details of the RoI region. Figure 6 provides a visualization of the intuitive effects and decoded images corresponding to different masking methods and ratios. Figure 7 demonstrates the impact

Table 4. Comparison of diffusion-based image compression encode-decode time. Test on the Kodak dataset and the average time is reported. The test hardware is a single NVIDIA A800.

Method	Diffusion Steps	Time (s)	
		Encode	Decode
PerCo	20	0.04	0.98
PICS	25	221.10	42.80
MRIDC (Mask 0%)	20	0.078	1.28
MRIDC (Mask 30%)	20	0.083	1.35

of ablating the high-level encoders f_h and MPT on image quality. These results collectively highlight the effectiveness of the proposed framework and its specific modules. For additional visualization results, please refer to the supplementary material.

5. Conclusion

Considering the potential conflict between extremely low bitrate compression and consistency of decoded image details, we propose an adaptive compression framework MRIDC based on the concept of RoI, which performs generative compression based on vector quantization space. Given the RoI map, the MLM module designed at the encoder uses it to guide the distribution of bits in the latent space in a training-free manner, thereby preserving more details of the RoI. At the decoder, we introduce a BPCG module for controllable generation, based on mask prediction, to produce outputs with optimal perceptual quality. Experimental results validate the effectiveness of the proposed modules in maintaining image quality and demonstrate the advancement of MRIDC in regional adaptive DP balance. Given the current limitations on the size of the latent space, we consider optimizing its flexibility in the future to support a wider range of variable bitrates.

References

- [1] Gregory K Wallace. The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, 1992. 2, 3
- [2] Fabrice Bellard. BPG image format. <https://bellard.org/bpg/>, 2018. Accessed: 2024-11-01. 3
- [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. 2
- [4] 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. 2, 3
- [5] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7948, 2020.
- [6] 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.
- [7] 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. 2, 3
- [8] Bradley Denby, Krishna Chintalapudi, Ranveer Chandra, Brandon Lucia, and Shadi Noghiabi. Kodan: Addressing the computational bottleneck in space. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pages 392–403, 2023. 2
- [9] Bill Tao, Om Chabra, Ishani Janveja, Indranil Gupta, and Deepak Vasishth. Known knowns and unknowns: Near-realtime earth observation via query bifurcation in serval. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pages 809–824, 2024. 2
- [10] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019. 2, 3
- [11] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual GAN. In *Proceedings of the European Conference on Computer Vision*, pages 205–220, 2018. 2
- [12] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.
- [13] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11304–11314, 2022.
- [14] 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. 6
- [15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.
- [16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.
- [17] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.
- [18] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [19] 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. 2, 3
- [20] 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. 3, 6
- [21] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Federico Perazzi, and Sun-Yuan Kung. Content-aware GAN compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12156–12166, 2021.
- [22] Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneider, and Björn Schuller. EGIC: Enhanced low-bit-rate generative image compression guided by semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 202–220. Springer, 2025. 2, 3
- [23] 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: From Information Theory to Applications*, 2023. 2, 3, 6, 7
- [24] Noor Fathima Ghouse, Jens Petersen, Auke Wiggers, Tianlin Xu, and Guillaume Sautiere. A residual diffusion model for high perceptual quality codec augmentation. *arXiv preprint arXiv:2301.05489*, 2023. 2, 3
- [25] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023.

- [26] Emiel Hoogetboom, 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. 2, 3
- [27] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 7
- [28] Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Towards extreme image compression with latent feature guidance and diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2024. 2, 3
- [29] Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22324–22333, 2023. 2, 3, 6
- [30] Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2900–2909, 2024. 2, 3
- [31] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *International Conference on Machine Learning*, pages 1–27, 2016. 3
- [32] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Machine Learning*, pages 1–23. PMLR, 2018. 3
- [33] 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. 3, 6
- [34] Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, and Ronggang Wang. MLIC: Multi-reference entropy model for learned image compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7618–7627, 2023.
- [35] Yibo Yang and Stephan Mandt. Computationally-efficient neural image compression with shallow decoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 530–540, 2023.
- [36] Julius Kunze, Daniel Severo, Giulio Zani, Jan-Willem van de Meent, and James Townsend. Entropy coding of unordered data structures. In *The Twelfth International Conference on Learning Representations*, 2024.
- [37] Han Li, Shaohui Li, Wenrui Dai, Chenglin Li, Junni Zou, and Hongkai Xiong. Frequency-aware transformer for learned image compression. In *The Twelfth International Conference on Learning Representations*, 2023.
- [38] Tongda Xu, Ziran Zhu, Dailan He, Yanghao Li, Lina Guo, Yuan Yuan Wang, Zhe Wang, Hongwei Qin, Yan Wang, Jingjing Liu, et al. Idempotence and perceptual image compression. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [39] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023. 3, 6
- [40] Yiyang Ma, Wenhan Yang, and Jiaying Liu. Correcting diffusion-based perceptual image compression with privileged end-to-end decoder. *arXiv preprint arXiv:2404.04916*, 2024. 3
- [41] Tom Bachard, Tom Bordin, and Thomas Maugey. CoCliCo: Extremely low bitrate image compression based on CLIP semantic and tiny color map. In *Picture Coding Symposium*, 2024. 3
- [42] Haowei Kuang, Yiyang Ma, Wenhan Yang, Zongming Guo, and Jiaying Liu. Consistency guided diffusion model with neural syntax for perceptual image compression. In *ACM Multimedia*, 2024. 3
- [43] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [44] Lucas Relic, Roberto Azevedo, Markus Gross, and Christopher Schroers. Lossy image compression with foundation diffusion models. In *Proceedings of the European Conference on Computer Vision*, pages 1–17. Springer, 2024. 3
- [45] Chunlei Cai, Li Chen, Xiaoyun Zhang, and Zhiyong Gao. End-to-end optimized roi image compression. *IEEE Transactions on Image Processing*, 29:3442–3457, 2019. 3
- [46] Myungseo Song, Jinyoung Choi, and Bohyung Han. Variable-rate deep image compression through spatially-adaptive feature transform. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2380–2389, 2021.
- [47] Yi Ma, Yongqi Zhai, Chunhui Yang, Jiayu Yang, Ruofan Wang, Jing Zhou, Kai Li, Ying Chen, and Ronggang Wang. Variable rate roi image compression optimized for visual quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1936–1940, 2021.
- [48] Meng Li, Shangyin Gao, Yihui Feng, Yibo Shi, and Jing Wang. Content-oriented learned image compression. In *Proceedings of the European Conference on Computer Vision*, pages 632–647. Springer, 2022.
- [49] Yuxi Liu, Wenhan Yang, Huihui Bai, Yunchao Wei, and Yao Zhao. Region-adaptive transform with segmentation prior for image compression. In *Proceedings of the European Conference on Computer Vision*, pages 181–197. Springer, 2024. 3
- [50] Besnier Victor and Chen Mickael. A pytorch reproduction of masked generative image transformer. *arXiv preprint arXiv:2310.14400*, 2023. 4, 6
- [51] 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. 6
- [52] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6
- [53] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Jia-Wang Bian, Le Zhang, Xiang Bai, and Jinhui Tang. Richer convolutional features for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1939–1946, 2019. 6
- [54] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 6
- [55] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020. 6
- [56] 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. 6
- [57] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *The Sixth International Conference on Learning Representations*, 2018. 6
- [58] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 6
- [59] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 6
- [60] Eastman Kodak. Kodak lossless true color image suite. <http://r0k.us/graphics/kodak/>, 1993. Accessed: 2024-10-25. 6
- [61] Bossen Frank, Suehring Karsten, and Li Xiang. Vvcsoftware. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM, 2018. Accessed: 2024-10-25. 6