Controllable Distortion-Perception Tradeoff Through Latent Diffusion for Neural Image Compression

Chuqin Zhou[1], Guo Lu[1]*, Jiangchuan Li[1], Xiangyu Chen[2], Zhengxue Cheng[1], Li Song[1], Wenjun Zhang[1]

[1]Shanghai Jiao Tong University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom
{zhouchuqin,luguo2014,cl4p-123,zxcheng,song_li,zhangwenjun}@sjtu.edu.cn, chxy95@gmail.com

Abstract

Neural image compression often faces a challenging trade-off among rate, distortion and perception. While most existing methods typically focus on either achieving high pixel-level fidelity or optimizing for perceptual metrics, we propose a novel approach that simultaneously addresses both aspects for a fixed neural image codec. Specifically, we introduce a plug-and-play module at the decoder side that leverages a latent diffusion process to transform the decoded features, enhancing either low distortion or high perceptual quality without altering the original image compression codec. Our approach facilitates fusion of original and transformed features without additional training, enabling users to flexibly adjust the balance between distortion and perception during inference. Extensive experimental results demonstrate that our method significantly enhances the pretrained codecs with a wide, adjustable distortion-perception range while maintaining their original compression capabilities. For instance, we can achieve more than 150% improvement in LPIPS-BDRate without sacrificing more than 1 dB in PSNR.

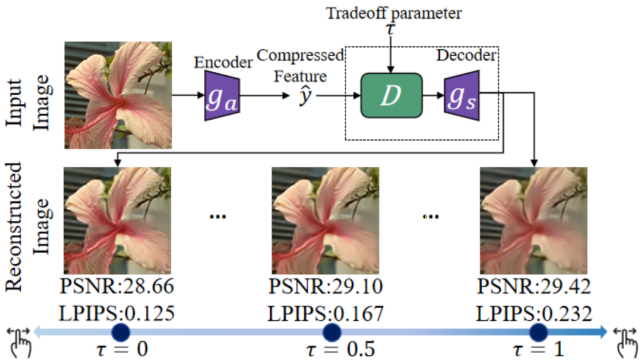

Figure 1: Overview of our proposed method. $\mathcal{D}$ represents a plug-and-play adaptive latent fusion module at decoder side for a base neural codec. We can achieve different distortion (PSNR) and perception (LPIPS) trade-offs, controlled by τ. For simplicity, quantization and entropy coding are omitted.

Introduction

As digital visual data continues to dominate Internet traffic, the development of efficient image and video codecs becomes increasingly crucial. In recent years, deep learning-based codecs have achieved significant advancements in both image domain (Cheng et al. 2020; Mentzer et al. 2020) and video domain (Hu, Lu, and Xu 2021; Li, Li, and Lu 2023; Lu et al. 2019, 2021, 2024a). These codecs have demonstrated a superior compression performance compared to traditional codecs (Bellard 2018; Bross et al. 2021).

Current learning-based image codecs primarily rely on the transform coding paradigm and variational autoencoders (VAEs) (Ballé et al. 2018). Most of these models use a rate-distortion loss function, directly optimizing for low distortion performance. However, distortion-oriented codecs often exhibit mode averaging behavior at low bitrates (Zhao, Song, and Ermon 2017), resulting in blurring that significantly degrades visual quality for human observers.

Recent studies demonstrate that optimizing for perceptual quality can lead to greater compression gains by allowing for imperceptible distortions, thereby reducing the bitrate. For example, HiFiC (Mentzer et al. 2020) proposes to use Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) to optimize neural image codecs. In contrast, CDC (Yang and Mandt 2023) employs a diffusion-based (Ho, Jain, and Abbeel 2020) decoder to improve perceptual quality. However, both approaches struggle to achieve high pixel-level fidelity, as they may introduce high-frequency noise and unrealistic textures. Blau and Michaeli (2019) explains this phenomenon by highlighting a fundamental trade-off between perceptual quality and distortion. They suggest that these two goals cannot be fully achieved simultaneously within a given architecture. We argue that both pixel-level fidelity and image-level realism are crucial for neural image compression frameworks. The inability to achieve both metrics simultaneously is a significant limitation in current image codecs. Given this limitation, an ideal codec should have the flexibility to traverse between different distortion-perception trade-offs at a given bitrate.

Recent research has explored flexible distortion-perception tradeoffs in image compression. A notable example is MRIC (Agustsson et al. 2023), which introduces a hyperparameter in the loss function to balance perception and distortion. This hyperparameter also serves as a condition for the decoder to adjust its reconstruction. However,

*Corresponding author.

this approach requires training the entire model from scratch and suffers from the inherent instability of GAN training.

In this paper, we propose a novel compression pipeline that allows for a controllable trade-off between distortion and perception for a fixed pretrained codec, as illustrated in Fig. 1. Specifically, we introduce a plug-and-play adaptive latent fusion module at the decoder side, which transforms the decoded latent representations using a latent diffusion process (Rombach et al. 2022). This process allows representations originally optimized for low distortion to be converted to prioritize high perceptual quality, and vice versa.

Assuming the base neural image codec is distortion-oriented, we first develop an auxiliary encoder, used only in the training stage, to generate guiding information optimized for perceptual quality. We then train the adaptive latent fusion module using perceptual loss while keeping the base codec's parameters fixed. In the inference stage, we fuse the original decoded feature with outputs from diffusion step based on user's preference, and the fused features are decoded by the original decoder. When integrated with existing variable bit rate schemes, our proposed model facilitates a trade-off among rate, distortion, and perception within a unified framework. Extensive experiments demonstrate the effectiveness of our framework. For distortion-oriented codecs, we achieve a more than 150% improvement in LPIPS-BDRate with less than 1 dB sacrifice in PSNR.

Our main contributions are as follows:

- We introduce an adaptive latent fusion module that enables controllable reconstruction at the decoder side, offering varying distortion-perception trade-offs.
- Our method serves as a plug-and-play module for fixed pretrained neural image codecs and is compatible with various compression frameworks.

Related Work

Image Compression. Image compression aims to reduce storage size by exploiting intra-image redundancy. Traditional image coding standards, such as JPEG (Wallace 1991) and BPG (Sullivan et al. 2012), employ manually designed modules like DCT to enhance compression performance, guided by the rate-distortion principle. In this context, distortion is typically measured using mean square error.

Recently, VAE-based Neural Image Compression (NIC) methods (Minnen, Ballé, and Toderici 2018; Cheng et al. 2020; He et al. 2021; Zhu, Yang, and Cohen 2022; Zheng and Gao 2024) have experienced significant advancements, surpassing the current state-of-the-art traditional image codecs like VVC (Bross et al. 2021). One of the representative works is the hyperprior-based method (Ballé et al. 2018), which models latents using extracted hyperprior information for enhanced compression. Subsequent research has further improved compression performance through more sophisticated architecture (Cheng et al. 2020; Zhu, Yang, and Cohen 2022) or entropy model (He et al. 2021). However, it is noteworthy that most existing works optimize codecs based on the rate-distortion strategy, potentially introducing blur artifacts at low bitrate settings.

To address these issues and enhance the realism of compressed images, researchers have introduced the perception loss to optimize the image codecs. At this juncture, the majority of mature technologies in this field are GAN-based. For example, HiFiC (Mentzer et al. 2020) achieves much more realistic results at low-bitrate settings. MS-ILLM (Muckley et al. 2023) introduces a non-binary discriminator, further enhancing the perceptual quality of reconstructed images.

Diffusion Probabilistic Models. Denoising diffusion probabilistic models (DDPMs) (Ho, Jain, and Abbeel 2020) generate data through a series of iterative stochastic denoising steps. The joint distribution of the data x_0 and the latent variable $x_{1:T}$ is learned through the model, i.e., $p_\theta(x_0) = \int p_\theta(x_{0:T})dx_{1:T}$. The goal of DDPMs is using network $\epsilon_\theta(x_t, t)$ to predict noise ϵ from a noisy image x_t at a noise level t, where the noise ϵ is used to perturb a particular image x_0 through $x_t(x_0) = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon$.

DDPMs have demonstrated success in various application areas, including image and video generation (Podell et al. 2024; Lu et al. 2024b), super-resolution (Wang et al. 2023; Luo et al. 2024), and restoration (Yu et al. 2024). Recent research has extended their application to image compression. CDC (Yang and Mandt 2023) uses the diffusion model as a decoder, employing compressed features as conditions to guide the diffusion process in the image domain. HFD (Hoogeboom et al. 2023) implements the diffusion model as an image restorer, using the low-quality decoded image to condition the diffusion process. Latent Diffusion Models (LDMs) proposed by Rombach et al. (2022) execute the diffusion process in latent domain of VAE. While Careil et al. (2023); Pan, Zhou, and Tian (2022) have applied LDM paradigm to image compression, they all focus on ultra-low bitrates, where reconstruction preserves only semantic information, not pixel-level fidelity. We aim to explore LDM applications at a more general bitrate setting, seeking to maintain low distortion while achieving high perceptual quality.

Given the fundamental trade-off between perceptual quality and distortion, all of the aforementioned compression works focus on achieving either lower distortion or better perceptual quality. Approaches focusing on reducing distortion tend to result in blurring at low bitrates, while those prioritizing perceptual quality often introduce unrealistic noise.

Distortion-Perception Trade-off in Compression. Since it is challenging to optimize both simultaneously, there are attempts to allow users to choose between the two. Zhang et al. (2021) propose achieving this trade-off through different decoders. Yan, Wen, and Liu (2022) introduce an approach using two decoders with interpolation in the image domain. MRIC (Agustsson et al. 2023) introduces a hyperparameter β at training and inference stages to control the weight of distortion and perception. Building on this, Iwai, Miyazaki, and Omachi (2024) proposes controlling the quantization process with hyperparameters q, β to achieve tunable rate, distortion, and realism. DIRAC (Ghouse et al. 2023) adopts an approach similar to HFD (Hoogeboom et al. 2023), where the output of the diffusion model is a residual added to the original low-quality image. This method can

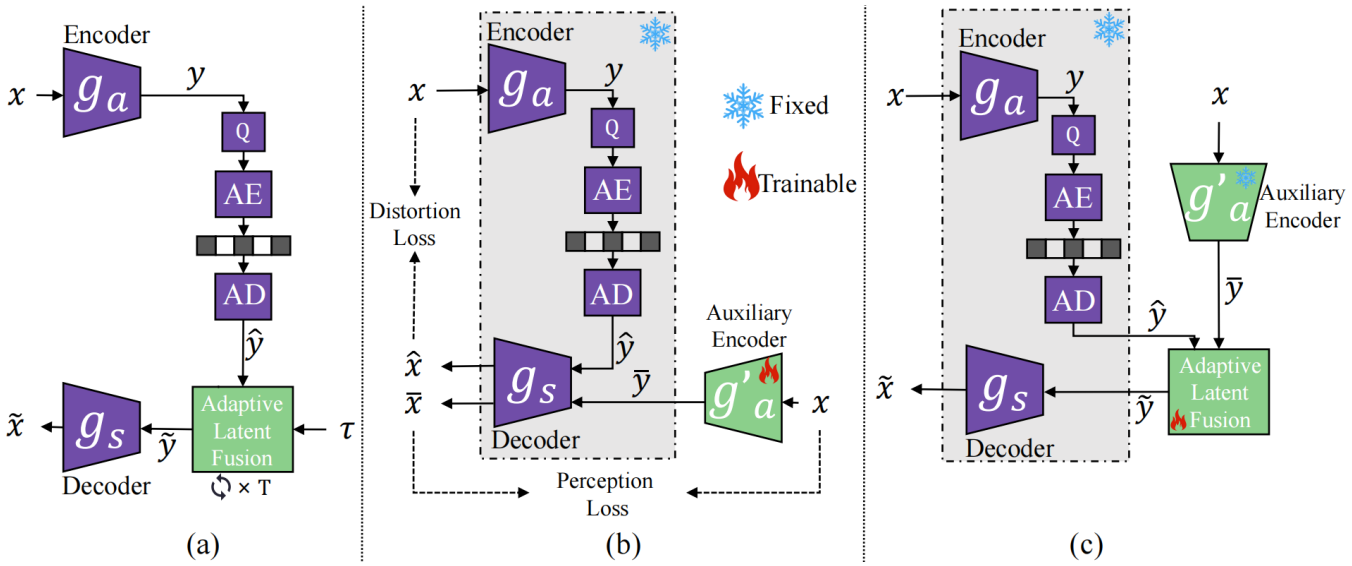

Figure 2: Illustration of the proposed method. For simplicity, we assume the base NIC is distortion-oriented. (a) represents the inference stage of our proposed pipeline. (b) and (c) represent the training procedures. We first train an auxiliary encoder g'_a for the fixed base neural codec. Then, we train a plug-and-play adaptive latent fusion module to transform the original latent representations into features optimized for perceptual quality.

control the diffusion steps to achieve a distortion-perception trade-off.

However, these methods require training the entire model from scratch, limiting their ability to leverage the strengths of pretrained codecs. In contrast, we propose a plug-and-play module, akin to the plug-and-play characteristic of the mask sampling module in Liu et al. (2023), to address the limitations of the original codec without modifying it, thereby preserving its inherent advantages.

Proposed Method

Overview

The framework of our proposed method is shown in Fig. 2 (a). To achieve a controllable distortion-perception trade-off, we introduce a plug-and-play adaptive latent fusion module on the decoder side of the existing pretrained codec. Our module can transform the original distortion-oriented features into perception-oriented features and vice versa. It also enables the fusion of these two types of features through weighted interpolation, resulting in decoded images with varying distortion-perception trade-offs.

Our approach is flexible and can be applied to various baseline codecs, whether distortion-oriented or perceptual-oriented. Here we use a basic VAE codec as an example. Specifically, the decoded latent feature $\hat{y}$, obtained by arithmetic decoding (AD), serves as the diffusion condition. Guided by the distortion-perception trade-off parameter τ, the adaptive latent fusion module generates a controllable feature $\tilde{y}$ with varying distortion-perception tradeoffs trough multiple diffusion steps. Finally, the fixed decoder g_s converts $\tilde{y}$ into a decoded image $\tilde{x}$. Our pipeline enables controllable reconstruction without modifying the existing network architecture or retraining the model.

Auxiliary Encoder for Baseline NIC

While many studies have focused on optimizing image compression architectures, a classical VAE architecture is illustrated on Fig. 2 (b). The corresponding loss function is as follows:

$$\min \mathcal{R}(Q(g_a(x))) + \beta \cdot \mathcal{L}(x, g_s(Q(g_a(x)))), \quad (1)$$

where Q denotes quantization and $\mathcal{R}(\cdot)$ represents the bitrate of the quantized latent representation. In distortion-oriented methods, $\mathcal{L}(x, \hat{x}) = ||x - \hat{x}||_2^2$ is used to measure distortion between the input and the reconstruction. For perception-oriented methods, $\mathcal{L}(x, \hat{x})$ is defined based on the perceptual metrics such as LPIPS (Zhang et al. 2018).

For clarity, we assume that our base neural codec is distortion-oriented unless otherwise specified in this paper. In the proposed framework, we aim to transform the existing quantized feature $\hat{y}$ to the controllable feature $\tilde{y}$ optimized for perceptual metrics. However, this transformation is non-trivial since the distributions of latent features optimizing for distortion or perception could be totally different, which is challenging even for the powerful diffusion methods. To address this issue, we further propose an auxiliary encoder to generate the corresponding guiding information, which is only used in the training stage.

Specifically, as shown on Fig. 2 (b), we introduce auxiliary encoder g'_a with the same structure as the original encoder g_a, directly connected to the fixed original decoder g_s during training. We optimize g'_a by minimizing perceptual loss between input and reconstruction, such as LPIPS, while keeping all other modules frozen. Then the optimized feature $\bar{y}$ will preserve more perceptual information and is

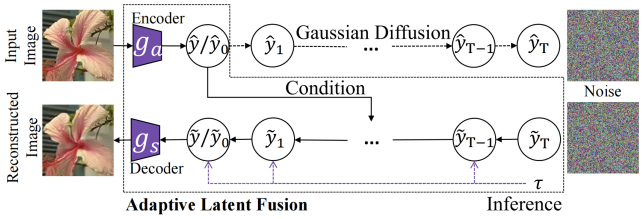

Figure 3: Overview of the latent diffusion process. For simplicity, we omit quantization and entropy coding modules. τ controls the diffusion process to achieve different tradeoffs.

employed as the auxiliary information for the training of our adaptive latent fusion module.

Adaptive Latent Fusion

Our proposed Adaptive Latent Fusion module transforms the decoded $\hat{y}$ into the controllable feature $\tilde{y}$, allowing for a desirable reconstruction that balances distortion and perception. The overview of this procedure is shown in Fig. 3. The compressed latent $\hat{y}$ serves as a condition for the diffusion process and can be fused with the transformed feature during inference.

The architecture of our network, dubbed $\mathcal{D}$, is shown in Fig. 4. We employ the classical latent diffusion architecture, which consists of M units, each containing two time-aware ResNet blocks and an attention block. To utilize the conditional information like the original feature $\hat{y}$ for the diffusion procedure, we generate condition information using the same unit and perform conditioning by concatenation. In addition, given the complexity of predicting noise for features with large channels and the strong condition $\hat{y}$, we directly learn transformed features $\tilde{y}$ instead of the noise, based on the original feature $\hat{y}$ and the pseudo-continuous variable $\frac{t}{T}$. More importantly, we further use the hyper-parameter τ to control the diffusion procedure and produce the features with varying distortion-perception trade-offs.

Training. As shown in Fig. 2 (c), we use the auxiliary encoder to generate auxiliary information $\bar{y}$ as the training target. During training, we add noise to the decoded feature $\hat{y}$. Thus, the diffusion input is given by $\hat{y}_t = \sqrt{\alpha_t}\hat{y} + \sqrt{1-\alpha_t}\epsilon$, where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\beta_t \in (0,1)$ is a monotonically increasing sequence of noise scheduler. t is randomly sampled from $[0, T]$ during training. We use the pseudo-continuous variable $\frac{t}{T}$ to indicate noise intensity to the model. This allows us to use an arbitrary and fewer number of denoising steps during inference.

We disregard noise level t and directly input the result of $\mathcal{D}$ at each step into the decoder g_s. The LPIPS distance between the reconstructed image and the original image is included in the loss function. The training loss function for adaptive latent fusion module is formulated as follows:

$$\min \lambda||\bar{y} - \mathcal{D}(\hat{y}_t, \hat{y}, \frac{t}{T})||_2^2 + \mathcal{L}(x, g_s(\mathcal{D}(\hat{y}_t, \hat{y}, \frac{t}{T}))), \quad (2)$$

where the first term represents the loss for the diffusion procedure and the second term represents the reconstruction

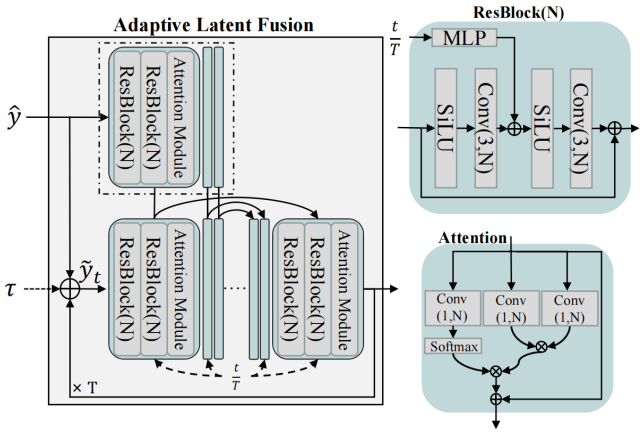

Figure 4: Architecture of our Adaptive Latent Fusion module. ResBlock(N) represents the ResBlock units with N channels. Conv(M,N) is a convolution layer with N channels, with $M \times M$ filters.

loss in image compression. λ is a trade-off parameter for different losses.

Inference. During inference, we follow DDIM (Song, Meng, and Ermon 2021), which replaces the original Markov process with a deterministic generative process to improve sampling speed. Our method aims to generate a decodable feature directly and calculate the predicted noise. As shown in Fig. 3, to differentiate from the training process, we refer to $\tilde{y}_t$ as the noisy feature during inference. At timestep t given the predicted result $\mathcal{D}(\tilde{y}_t, \hat{y}, \frac{t}{T})$, we can derive the equivalence with $\epsilon(\tilde{y}_t, \hat{y}, \frac{t}{T}) = \frac{\tilde{y}_t - \sqrt{\alpha_t}\mathcal{D}(\tilde{y}_t, \hat{y}, \frac{t}{T})}{\sqrt{1-\alpha_t}}$, which is the predicted noise. The sampling process can be formulated as:

$$\tilde{y}_{t-1} = \sqrt{\alpha_{t-1}}\mathcal{D}(\tilde{y}_t, \hat{y}, \frac{t}{T}) + \sqrt{1-\alpha_{t-1}}\epsilon(\tilde{y}_t, \hat{y}, \frac{t}{T}), \quad (3)$$

We can use the above recurrence equation to sample $\tilde{y}_{t-1}$ from $\tilde{y}_t$ as t gradually decreases from T to 0, while $\tilde{y}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. It should be noted that we use different values for T in the training and inference process, which significantly reduces the number of diffusion steps required for inference.

In practice, different scenarios have different requirements for distortion and perception. Therefore, we aim to achieve a controllable trade-off between them during the inference phase. We use the following equation to guide the sampling process and apply τ for weighted interpolation between the predicted output $\tilde{y}_{t-1}$ and the original latent $\hat{y}$.

$$\begin{aligned}\tilde{y}_{t-1} = \sqrt{\alpha_{t-1}}[(1-\tau^2) \times \mathcal{D}(\tilde{y}_t, \hat{y}, \frac{t}{T}) + \tau^2 \times \hat{y}] \\ + (1-\tau^2) \times \sqrt{1-\alpha_{t-1}}\epsilon(\tilde{y}_t, \hat{y}, \frac{t}{T}), \end{aligned} \quad (4)$$

We use τ^2 instead of τ to ensure that the two latent representations are weighted appropriately, achieving a more linear control effect. The parameter τ ranges from [0, 1], where $\tau = 0$ results in outputs composed entirely of perception-oriented latents. Conversely, When $\tau = 1$, $\tilde{y}_0 = \hat{y}$, which

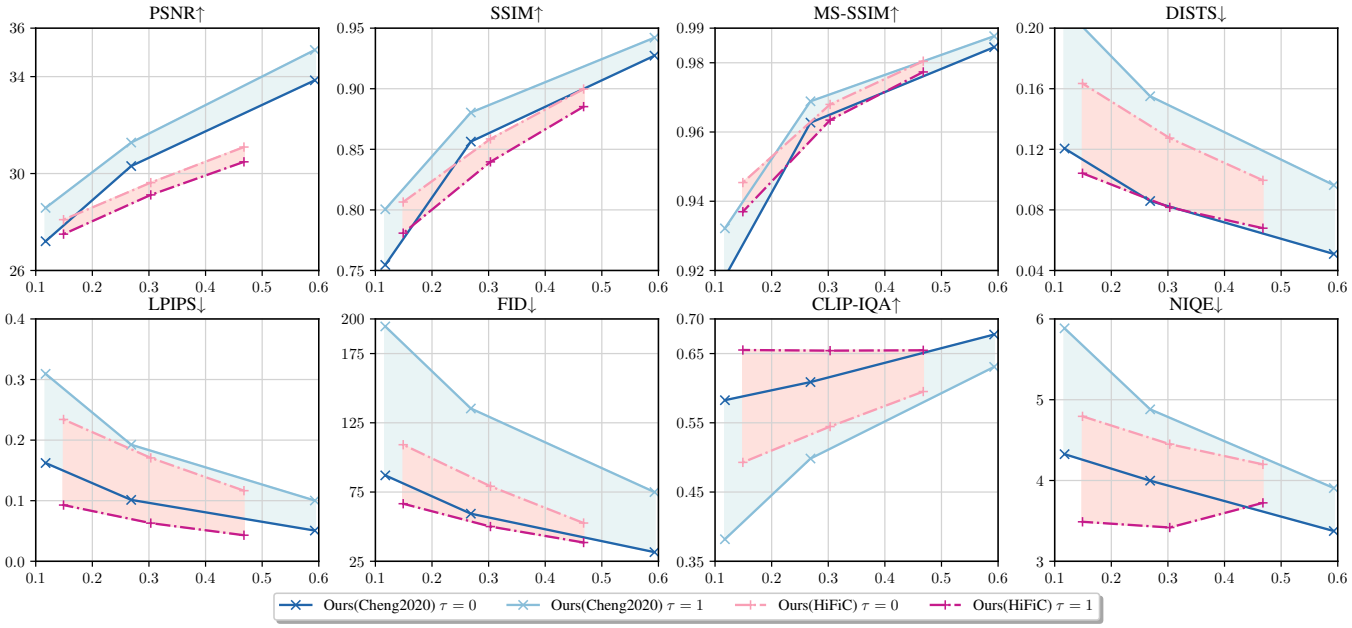

Figure 5: Trade-offs between bitrate and different metrics for various base codecs tested on Kodak dataset. Arrows in the plot titles indicate whether high(↑) or low(↓) values indices a better score.

corresponds to the base codec. The final decoded image $\tilde{x}$ is obtained by inputting the last $\tilde{y}_0$ into the decoder g_s.

Experiments

Experimental Settings

Training Details. For training, we use a high-quality Flickr2W dataset (Liu et al. 2020), and randomly crop images to a resolution of 256×256. To train the auxiliary encoder and the adaptive latent fusion module, we use the AdamW optimizer with a batch size of 32. The learning rate is maintained at a fixed value of 5×10^{-5}.

Evaluation. We evaluate our method using both distortion metrics and perceptual quality metrics. All evaluations are performed on full-resolution images. We select eight widely used metrics for image quality evaluation. To measure distortion and perception, we use PSNR, SSIM, DISTS, LPIPS (Zhang et al. 2018), FID (Heusel et al. 2017), CLIP-IQA (Wang, Chan, and Loy 2023), and NIQE (Mittal, Soundararajan, and Bovik 2013). Among these metrics, FID, CLIP-IQA, NIQE are non-reference metrics, while others are full-reference metrics. When calculating FID, we follow the procedure of previous compression works (Mentzer et al. 2020) by segmenting images into non-overlapping patches of 256×256 resolution. The evaluations are conducted on two common image compression benchmark datasets: the CLIC2020 test set and the Kodak dataset.

State-of-the-art Methods. We compare our method with several representative neural compression approaches. MSHyper (Ballé et al. 2018) introduces the hyperprior for enhanced compression. Cheng2020 (Cheng et al. 2020) employs an attention mechanism and outperforms the traditional VVC codec. HiFiC (Mentzer et al. 2020), a GAN-based codec trained for specific rate-perception trade-offs, exemplifies a leading perceptual codec. CDC (Yang and Mandt 2023) uses a conditional DDIM decoder to generate reconstructions from latent representations, providing distortion-oriented (CDC) and perception-oriented (CDC-lpips) models. MRIC (Agustsson et al. 2023) introduces β to achieve various distortion-perception tradeoffs. We limit comparisons to studies with publicly available codes and models for consistent testing and evaluation.

Main Results

Distortion-Perception Trade-off Ability. As a plug-and-play approach, our approach can be easily integrated with the existing image codecs. We select two representative methods, Cheng et al. (Cheng et al. 2020)(denoted as Ours(Cheng2020)) and HiFiC (Mentzer et al. 2020)(denoted as Ours(HiFiC)) as our base codecs and evaluate the versatility and effectiveness of the proposed approach. More test results for other base models can be found in the Appendix.

In Fig. 5, the shaded area represents the adjustable rate-distortion-perception range achievable by a single model. Our model is presented in two configurations: $\tau \in \{0, 1\}$. When $\tau = 1$, our codec is equivalent to the base codec. When $\tau = 0$, the optimization direction of our loss is opposite to the original codec. It is noted that our proposed approach can achieve a wide range of trade-offs between distortion and perception in different bitrates. In addition, our pipeline preserves the original codec and this enables us to maintain state-of-the-art (SOTA) performance in areas where the original codec excels.

For the distortion-optimized base codec, our approach, Ours (Cheng2020, $\tau = 0$), achieves a significant improve-

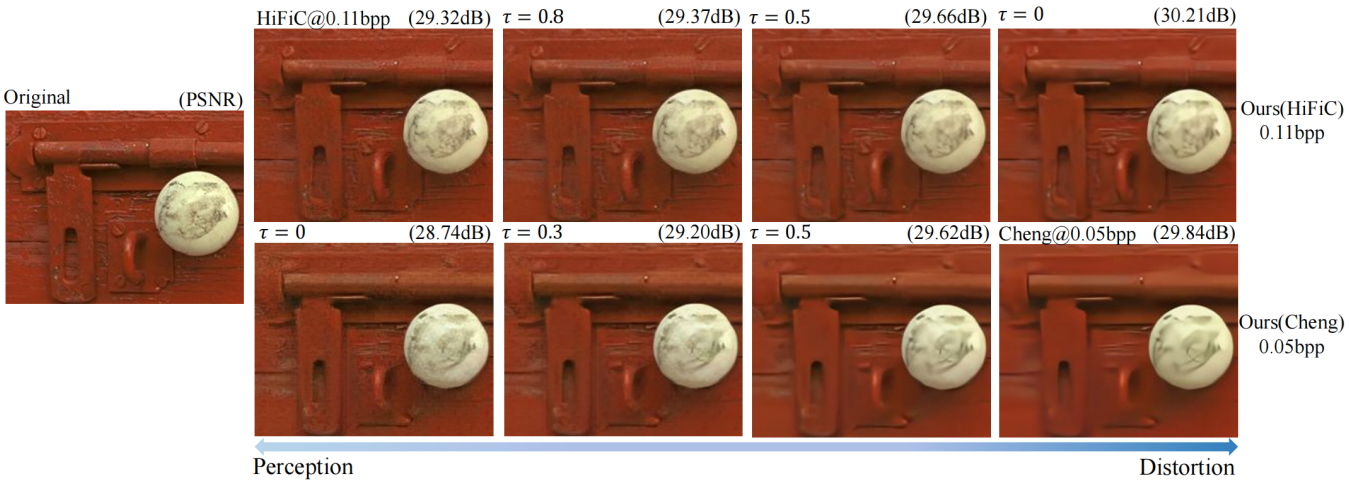

Figure 6: Kodak reconstructions of our method for different rate-distortion-perception. Shown scores are for full image.

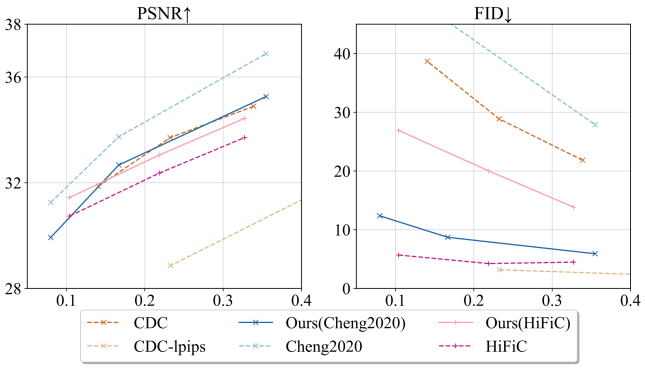

Figure 7: Trade-offs between bitrate and different metrics for various models tested on CLIC2020 test set. Ours are shown with $\tau = 0$.

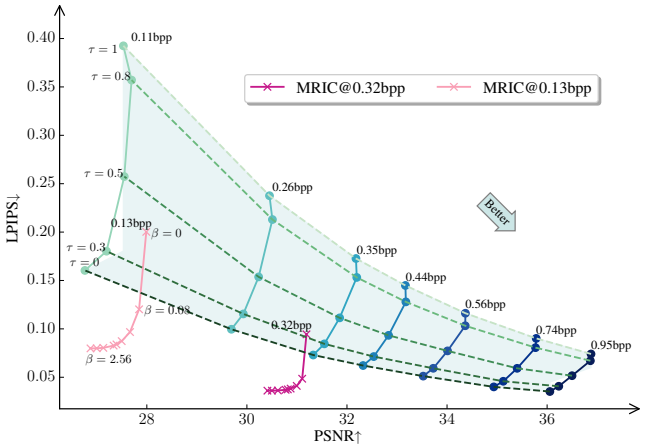

Figure 8: Distortion (PSNR) vs. perception(LPIPS) on Kodak for different rate-distortion-perception tradeoffs.

ment over the original Cheng's method, enhancing LPIPS-BDRate by 158.75%, which corresponds to an average LPIPS gain of 0.096, with only a modest PSNR degradation of 1.08 dB on average. On the high-perception side ($\tau = 0$), our method matches or surpasses the state-of-the-art generative approach HiFiC in DISTS, while also delivering a substantial average PSNR improvement of 1.48 dB.

For the perception-oriented base codec, compared with the original HiFiC approach, our approach Ours(HiFiC, $\tau = 0$) saves 22.59% bitrate in terms of PSNR performance on Kodak dataset. On the low-distortion side ($\tau = 0$), we match or outperform the state-of-the-art distortion method Cheng2020 in SSIM and MS-SSIM, while also significantly outperforming it in perceptual quality.

Comparison with SOTA Methods. Fig. 7 provides more results when comparing our approach ($\tau = 0$) with existing image codecs in the CLIC2020 dataset.

When compared with CDC method, which also employs a diffusion model, Ours(HiFiC, $\tau = 0$) achieves a comparable distortion performance, while significantly enhancing perceptual quality. Ours(Cheng2020, $\tau = 0$) significantly outperforms CDC-lpips in terms of PSNR while achieving a similar LPIPS score. Additionally, it surpasses CDC in LPIPS and FID while maintaining a similar PSNR. Despite not achieving the most SOTA performance in perceptual metrics due to the fixed decoder and the inherent limitations of encoder-extracted features, we observe substantial enhancement in perceptual quality and achieve comparable results with perception-oriented models, which is the main purpose of our design. In fact, given such a low LPIPS, further reductions may not yield perceptually significant improvements but could potentially introduce more high-frequency noise at low bit rates.

Rate-Distortion-Perception Trade-off. Our method enables the transformation of distortion and perception-oriented latent representations, allowing for a flexible trade-off between these dimensions during the inference phase. By combining this approach with a variable bitrate scheme,

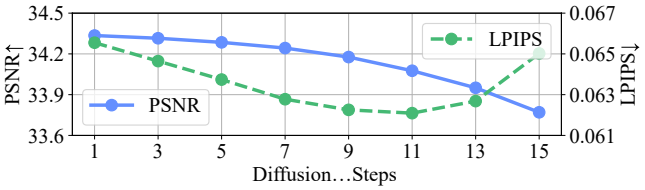

Figure 9: Reconstruction quality in the terms of PSNR and LPIPS versus the number of sampling timesteps.

we can achieve a three-dimensional exploration of the rate-distortion-perception tradeoff. Here we apply the variable bitrate method proposed by (Cui et al. 2021) to the baseline method (Ballé et al. 2018), denoted as Ours(MSHyper).

As illustrated in Fig. 8, we evaluate the effect of different τ values ($\tau \in \{0, 0.3, 0.5, 0.8, 1\}$) at various bit rates. Our results Ours(MSHyper) demonstrate that the proposed approach performs a controllable and smooth traversal between low distortion (high PSNR) and high perceptual quality (low LPIPS). MRIC (Agustsson et al. 2023), on the other hand, is not a variable bitrate scheme and involves training the entire model from scratch, resulting in a non-linear and less controllable tradeoff between distortion and perception. Their method shows little or no change from $\beta = 0.64$ to $\beta = 2.56$, making their methods less controllable. Our approach allows for a linear transition between distortion and perception. Our method can achieve a much larger range of conversion, significantly improving perceptual quality, as evidenced by a 170% improvement in LPIPS-BDRate.

Quantitative Results. The quantitative results are presented in Fig. 6. When $\tau = 1$, our codec is equivalent to the base codec. Ours(Cheng, $\tau = 0$) generates reconstructed image with comparable perceptual quality to HiFiC, achieving much higher PSNR even at a significantly lower BPP. Decreasing τ results in a drastic change in perceptual quality. Ours(HiFiC, $\tau = 0$) increases PSNR while maintaining more texture details, avoiding the smoothness and blurriness that Cheng exhibits. Decreasing τ slightly reduces reconstruction texture while boosting PSNR.

Ablation Study

Effectiveness of Latent Diffusion Module. As shown in Table 1, we evaluate two variants. Negative values indicate bitrate savings compared to our method. Variant-1 excludes the diffusion process, omitting the auxiliary encoder and its associated loss. Our method outperforms Variant-1 in both PSNR and LPIPS, demonstrating the effectiveness of the latent diffusion process and auxiliary encoder. As shown in Fig. 10, Variant-1's features reveal a critical issue: PSNR fails to correlate positively with τ when fused with untreated features. While this design can decode the original features and produce modified features, it fails to achieve effective fusion of these two through simple weighted interpolation. To obtain Variant-2, the decoder is unfrozen and trained jointly. While Variant-2 achieves results similar to ours in terms of LPIPS and better PSNR, the decoder cannot handle unprocessed latents from the encoder as it has been modified. Consequently, Variant-2 does not offer the same flexibility in distortion-perception trade-offs as our method.

Methods	Diffusion Process	Decoder	BDRate PSNR	BDRate LPIPS
Ours	✓	✗	0	0
Variant-1	✗	✗	+6.91%	+4.69%
Variant-2	✗	✓	-2.36%	+0.63%

Table 1: Ablation study on different variants of our model on Kodak dataset.

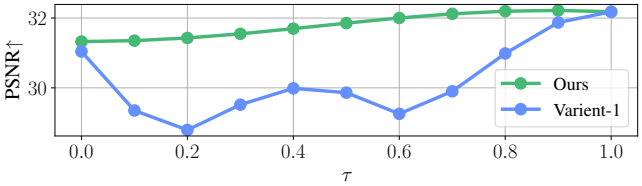

Figure 10: Comparison of PSNR values with varying hyperparameters τ across different methods.

Generation Speed. We investigate the impact of sampling timesteps on the reconstruction quality. The trade-off between generation speed and quality is illustrated in Fig. 9. Our method achieves competitive performance with just 10 timesteps, adopted as the default setting. However, exceeding 10 timesteps degrades LPIPS scores, as the diffusion module introduces excessive high-frequency components, causing a drift from the original content. In comparison, CDC (Yang and Mandt 2023) employs DDIM in the pixel domain with 17 timesteps for direct image output.

Conclusion

In this paper, we present a plug-and-play method for neural image codecs, which allows for various distortion-perception reconstruction tradeoffs from a single latent representation. An adaptive latent fusion module can either transform the feature to a high-perception or a low-distortion one. We keep the image codecs unchanged, but allow the trade-off between realism and distortion to happen on the receiver side, with no change in the bit stream. By integrating with the variable bitrate codecs, users can select the desired rate and switch between reconstructions which are as close to the original as possible and those with a better level of detail. Experiments demonstrate that our method works effectively with both distortion-driven and perception-driven models, achieving remarkable performance improvements in areas where the original codec is less effective, while minimally sacrificing their respective advantages.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant 2024YFF0509700, National Natural Science Foundation of China(62471290,62331014) and the Fundamental Research Funds for the Central Universities.

References

Agustsson, E.; Minnen, D.; Toderici, G.; and Mentzer, F. 2023. Multi-Realism Image Compression with a Conditional Generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 22324–22333.

Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations(ICLR)*.

Bellard, F. 2018. Bpg image format. https://bellard.org/bpg/.

Blau, Y.; and Michaeli, T. 2019. Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff. In *Proceedings of the 36th International Conference on Machine Learning(ICML)*, 675–685. PMLR.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021. Overview of the Versatile Video Coding (VVC) Standard and its Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 3736–3764.

Careil, M.; Muckley, M. J.; Verbeek, J.; and Lathuilière, S. 2023. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations(ICLR)*.

Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020. Learned Image Compression With Discretized Gaussian Mixture Likelihoods and Attention Modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*.

Cui, Z.; Wang, J.; Gao, S.; Guo, T.; Feng, Y.; and Bai, B. 2021. Asymmetric Gained Deep Image Compression With Continuous Rate Adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 10532–10541.

Ghouse, N. F.; Petersen, J.; Wiggers, A.; Xu, T.; and Sautiere, G. 2023. A residual diffusion model for high perceptual quality codec augmentation. *arXiv preprint arXiv:2301.05489*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, D.; Zheng, Y.; Sun, B.; Wang, Y.; and Qin, H. 2021. Checkerboard Context Model for Efficient Learned Image Compression. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 14771–14780. Computer Vision Foundation / IEEE.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems(NeurIPS)*, 6626–6637.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems(NeurIPS) 33*.

Hoogeboom, E.; Agustsson, E.; Mentzer, F.; Versari, L.; Toderici, G.; and Theis, L. 2023. High-fidelity image compression with score-based generative models. *arXiv preprint arXiv:2305.18231*.

Hu, Z.; Lu, G.; and Xu, D. 2021. FVC: A New Framework Towards Deep Video Compression in Feature Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 1502–1511.

Iwai, S.; Miyazaki, T.; and Omachi, S. 2024. 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*, 2900–2909.

Li, J.; Li, B.; and Lu, Y. 2023. Neural Video Compression with Diverse Contexts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 22616–22626.

Liu, J.; Lu, G.; Hu, Z.; and Xu, D. 2020. A Unified End-to-End Framework for Efficient Deep Image Compression. *arXiv preprint arXiv:2002.03370*.

Liu, L.; Zhao, M.; Yuan, S.; Lyu, W.; Zhou, W.; Li, H.; Wang, Y.; and Tian, Q. 2023. Exploring Effective Mask Sampling Modeling for Neural Image Compression. *arXiv preprint arXiv: 2306.05704*.

Lu, G.; Ge, X.; Zhong, T.; Hu, Q.; and Geng, J. 2024a. Preprocessing Enhanced Image Compression for Machine Vision. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lu, G.; Ouyang, W.; Xu, D.; Zhang, X.; Cai, C.; and Gao, Z. 2019. DVC: An End-To-End Deep Video Compression Framework. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 11006–11015.

Lu, G.; Zhang, X.; Ouyang, W.; Chen, L.; Gao, Z.; and Xu, D. 2021. An End-to-End Learning Framework for Video Compression. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10): 3292–3308.

Lu, H.; Yang, G.; Fei, N.; Huo, Y.; Lu, Z.; Luo, P.; and Ding, M. 2024b. VDT: General-purpose Video Diffusion Transformers via Mask Modeling. In *The Twelfth International Conference on Learning Representations(ICLR)*.

Luo, X.; Xie, Y.; Qu, Y.; and Fu, Y. 2024. SkipDiff: Adaptive Skip Diffusion Model for High-Fidelity Perceptual Image Super-resolution. In *Thirty-Eighth AAAI Conference on Artificial Intelligence(AAAI)*, 4017–4025.

Mentzer, F.; Toderici, G.; Tschannen, M.; and Agustsson, E. 2020. High-Fidelity Generative Image Compression. In *Advances in Neural Information Processing Systems(NeurIPS) 33*.

Minnen, D.; Ballé, J.; and Toderici, G. 2018. Joint Autoregressive and Hierarchical Priors for Learned Image Compression. In *Advances in Neural Information Processing Systems(NeurIPS) 31*, 10794–10803.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.*, 20(3): 209–212.

Muckley, M. J.; El-Nouby, A.; Ullrich, K.; Jégou, H.; and Verbeek, J. 2023. Improving Statistical Fidelity for Neural Image Compression with Implicit Local Likelihood Models. In *International Conference on Machine Learning(ICML)*, volume 202, 25426–25443. PMLR.

Pan, Z.; Zhou, X.; and Tian, H. 2022. Extreme generative image compression by learning text embedding from diffusion models. *arXiv preprint arXiv:2211.07793*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations(ICLR)*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 10674–10685.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *9th International Conference on Learning Representations(ICLR)*.

Sullivan, G. J.; Ohm, J.; Han, W.; and Wiegand, T. 2012. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12): 1649–1668.

Wallace, G. K. 1991. The JPEG Still Picture Compression Standard. *Communication ACM*, 34(4): 30–44.

Wang, J.; Chan, K. C. K.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence(AAAI)*, 2555–2563.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*.

Yan, Z.; Wen, F.; and Liu, P. 2022. Optimally Controllable Perceptual Lossy Compression. In *International Conference on Machine Learning(ICML)*, volume 162 of *Proceedings of Machine Learning Research*, 24911–24928. PMLR.

Yang, R.; and Mandt, S. 2023. Lossy Image Compression with Conditional Diffusion Models. In *Advances in Neural Information Processing Systems(NeurIPS) 36*.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling Up to Excellence: Practicing Model Scaling for Photo-Realistic Image Restoration In the Wild. *arXiv preprint arXiv:2401.13627*.

Zhang, G.; Qian, J.; Chen, J.; and Khisti, A. 2021. Universal Rate-Distortion-Perception Representations for Lossy Compression. In *Advances in Neural Information Processing Systems(NeurIPS) 34*, 11517–11529.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 586–595.

Zhao, S.; Song, J.; and Ermon, S. 2017. Towards Deeper Understanding of Variational Autoencoding Models. *arXiv preprint arXiv:1702.08658*.

Zheng, H.; and Gao, W. 2024. End-to-End RGB-D Image Compression via Exploiting Channel-Modality Redundancy. In *Thirty-Eighth AAAI Conference on Artificial Intelligence(AAAI)*, 7562–7570.

Zhu, Y.; Yang, Y.; and Cohen, T. 2022. Transformer-based Transform Coding. In *The Tenth International Conference on Learning Representations(ICLR)*.