Syntax-Driven Multi-Realism Image Compression With Consistency Guided Diffusion Model

Haowei Kuang, *Graduate Student Member, IEEE*, Wenhan Yang, *Member, IEEE*, Zongming Guo, *Senior Member, IEEE*, and Jiaying Liu, *Fellow, IEEE*

***Abstract*—Given the challenge of balancing high fidelity with perceptual quality, multi-realism image compression is developed to adapt flexibly to varying requirements. It allows images with different levels of realism to be decoded from the same bit stream. Diffusion models are known for generating images with high perceptual quality. However, their inherent process of adding noise and denoising is often difficult to control and will bring more distortion. This limits their direct application in image compression, especially in multi-realism image compression which requires precise control to adapt to different requirements. To address this issue, we propose a *Consistency Guided Diffusion Model* as a post-processing network for multi-realism image compression, aiming to control the addition of detail representations, thereby adjusting the trade-off between subjective quality and fidelity. In detail, our proposed novel method is crafted to introduce an additional consistency guided feature branch into the diffusion model to constrain the deviation caused by randomness in the diffusion process to ensure fidelity. Furthermore, a syntax-driven feature fusion module is constructed to guide the information adaptive fusion of two branches with an input extra ultra-low stream, which contains the context information and trade-off control information. In addition, we design a warm-up based training strategy and adopt a continuous online optimization method to improve coding efficiency and trade-off control precision. Extensive experiments validate the superiority of our method over existing compression techniques, as well as the effectiveness of each component.**

***Index Terms*—Image compression, generative model, denoising diffusion model, neural syntax.**

I. Introduction

IMAGE compression aims to efficiently reduce the storage space and bandwidth needed without compromising the content of the image. In the past few decades, traditional image compression techniques, such as JPEG [1], JPEG2000 [2], and BPG [3], have dominated the field of image processing and have been widely accepted as universal standards. However, these optimizations are limited by their reliance on human-designed systems, which restricts their potential for global optimization. As the number of manually designed strategies increases, the compression model's architecture becomes increasingly complicated, gradually reaching its performance bottlenecks. With the remarkable advancements of deep learning, researchers [4], [5], [6], [7], [8] begin to explore neural image compression to take advantage of the potential of neural networks. Unlike traditional compression frameworks [1], [9], [10] relying on various manually designed coding tools, these methods excel at finding hidden patterns and correlations in image data by training neural networks on large datasets. Therefore, they achieve a remarkable balance between compression efficiency and distortion, outperforming traditional techniques in terms of overall performance. However, despite considerable progress, these methods are starting to hit a performance limit.

Recent studies [11], [12] highlight a critical challenge: the distortion metrics commonly employed in image compression often fail to align with human subjective evaluations of image quality. Some researchers have noticed this limitation and are turning to perceptual image compression techniques [13], [14], [15]. These methods focus on making images look more realistic to the human eye, rather than strictly matching the fidelity measurements. Generative models [16], [17], [18], are at the forefront of these advancements, serving as powerful tools for improving the visual appeal of compressed images. In particular, Generative Adversarial Networks (GANs) [18] have been incorporated into cutting-edge image compression frameworks due to their ability to create realistic details, significantly enhancing perceptual quality.

Nonetheless, these perceptual image compression methods still neglect a critical point: enhancing realism for better visual fidelity of reconstructions will certainly introduce a degree of ambiguity. This ambiguity introduces the possibility that the reconstruction deviates significantly from the original input, leaving it unclear which details are present in the original input and which are artificially introduced for better perceptual quality. Taking this into consideration, multi-realism image compression is introduced. This technology introduces a control mechanism to manage the distortion-perception trade-off, enabling the decoder to produce images with varying levels of realism from the same bit stream. That is, users can choose between an image with higher fidelity but less realism, or one with enhanced realism and better perceptual quality.

Received 17 December 2024; revised 14 December 2025; accepted 30 December 2025. Date of publication 6 January 2026; date of current version 7 May 2026. This work was supported in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology); and in part by the Major Key Project of PCL under Grant PCL2025A03. This article was recommended by Associate Editor L. Zhang. *(Corresponding author: Jiaying Liu.)*

Haowei Kuang, Zongming Guo, and Jiaying Liu are with the Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China (e-mail: kuanghw@stu.pku.edu.cn; guozongming@pku.edu.cn; liujiaying@pku.edu.cn).

Wenhan Yang is with the Pengcheng Laboratory, Shenzhen, Guangdong 518108, China (e-mail: yangwh@pcl.ac.cn).

Digital Object Identifier 10.1109/TCSVT.2026.3651547

Some studies [19], [20] make some efforts in this regard with GAN-based network structure, mitigating the problem to some extent. They are still constrained by the capabilities of existing GAN-based generative models, but as development continues, new opportunities are emerging.

Diffusion-based models [21], [22], [23] have become powerful tools in image generation, showing an impressive ability to create high-quality images. Built on diffusion processes, these models stand out for producing images with exceptional realism. However, despite their impressive achievements in image generation, many studies have pointed out that vanilla diffusion models often create highly realistic images with plenty of visual details, but this comes at the cost of significantly reduced fidelity. This phenomenon is caused by the inherent randomness of their progressive denoising mechanism, which may lead to variations in the reconstructed image. Thus, the direct application of diffusion models to the domain of multi-realism image compression poses a challenge, as it can be difficult to control the trade-off between realism and fidelity. Therefore, the exploration of effective strategies to utilize diffusion models for image compression still requires further research and innovation.

To make use of the power of the diffusion model and enable control over both fidelity and perceptual quality for realism-driven perceptual image compression, we propose constraining the diffusion model with global consistency guidance. Specifically, we propose a new approach called **Consistency Guided Diffusion Model**, which incorporates additional consistency guidance into the network structure of the diffusion model to restrain the deviation in the diffusion process. Besides, we propose the Syntax-Driven Feature Fusion (SFF) Module, which utilizes ultra-low bitrate streams extracted from the encoding phase. These streams carry semantic priors and trade-off control information for guiding the reconstruction process. By utilizing this semantic knowledge, SFF mitigates the uncertainty of post-processing target, resulting in reconstructions that are both precise controlled and faithful to the original image content. Moreover, we incorporate a continuous online optimization paradigm, enabling the model to dynamically adapt and enhance its performance during inference. By combining these efforts, we reduce the randomness in the diffusion process, allowing the model to capture syntax information from the input and helping achieve a controllable balance between how the image looks and how faithful it is to the original. With our consistency guided diffusion model (CGDM) [24] as a strong baseline, we further explore the utilization of diffusion to achieve the realism control mechanism for image reconstructed. We extend CGDM with an additional trade-off control factor in the innovative SFF module to solve the task of multi-realism image compression. With this improvement, our new approach CGDM+ is able to integrate trade-off control information and syntax information into a syntax vector, thereby achieving control over the degree of realism. Additionally, we employ a warm-up training strategy during the training phase to further refine the model's training process. As a result, they lead to precise control between perceptual image quality and fidelity, ultimately resulting in a significant improvement in the overall performance of the proposed approach.

Our contributions are summarized as follows:

- We devise a Consistency Guided Diffusion Model tailored for image compression. By introducing an additional consistency guidance mechanism for regulating the diffusion process, the randomness of diffusion can be constrained, and the trade-off between perception and fidelity can be adjusted through the guidance of consistency features.
- To merge the diffusion feature and consistency feature, we propose a Syntax-Driven Feature Fusion module. Specifically, the module can predict the adaptive convolution kernels with the input of ultra-low bitstream syntax vector, and use them fusing features. With different syntax vectors, our model can achieve different trade-offs between fidelity and perception.
- We employ a warm-up based two-stage training strategy to optimize the multi-realism-oriented network more stably in the training process. We also design a continuous online optimization method for more precise and wider range dynamic control of the distortion-perception trade-off in the inference process.

This paper is an extension of our earlier publication [24]. We make additional contributions in both methodology design and experiments. First, we introduce a novel syntax-driven feature fusion module, extending the task from only perceptual image compression to the broader scope of multi-realism compression. This innovative strategy not only incorporates both subjective quality-driven and fidelity-oriented compression but also enables users to control the balance between these two dimensions. Second, to ensure a smoother and more stable training trajectory, we optimize the training process. These improvements distinguishes it from our previous version, CGDM [24]. Leveraging the latest improvements, CGDM+ significantly expands the scope of its applications. Beyond the methodology contributions, we also enrich the experimental results, providing more quantitative and qualitative comparison results, user studies, and ablation studies. With our new technical contributions, our method on perception-oriented setting achieves BD-rate gains of -3.10 with PIEAPP as metric and -36.20 with PSNR as metric on the Kodak dataset, while performance on fidelity-driven setting also excels with a -43.33 BD-rate gain with PSNR as metric, anchored by CGDM.

The remainder of the paper is organized as follows. In Section II, we review existing works within pertinent domains. Subsequently, Section III introduces the motivation and detailed designs or our proposed method. Progressing further, Section IV showcases the effectiveness of our methods through extensive experiments. Finally, Section V concludes the paper while discussing future directions.

II. Related Works

A. Fidelity-Driven Image Compression

Over the past few years, the remarkable progress in deep learning has pushed image compression techniques to new heights, better than traditional methods in achieving an optimal

balance between bit rate efficiency and reconstruction quality. Ballé et al. [25] are the pioneers who utilize the power of neural networks to devise lossy image compression framework, thereby sparking a surge in innovative learning-based compression approaches [26], [27], [28]. Beyond non-linear transformations, extensive research explores the entropy coding of latent representations, leveraging learned probability models such as hyper-priors [4] and context models [29] to further enhance the compression performance.

Based on these, numerous researchers conducted thorough explorations of the components of the end-to-end image compression framework, including enhancing the transform module [8], [30], refining hyper-priors [31], optimizing the context model [6], [30], [32], and improving the entropy model [33], [34]. Each of these methods contributes to enhancing the performance of image compression models in various ways. Furthermore, some researchers worked on methods that use implicit neural representations [35], [36], [37], [38] to compress images by encoding them into the parameters of neural networks, which also show some potential. Nevertheless, as performance continues to rise, the performance of traditional fidelity-driven end-to-end image compression technique is nearing the theoretical ceiling, and enhancing it further has become increasingly challenging.

B. Multi-Realism Image Compression

Recent research [11], [12], [39] points out that the distortion measures often do not match up with how people actually perceive the image quality. Based on this while addressing the trade-off between image fidelity and storage efficiency, perceptual image compression methods have gotten significant attention, aiming to increase the realism of compressed images and enhance their perceptual quality in line with human visual perception. Early perceptual-oriented compression methods primarily rely on either manually designed or deep learning based perceptual image quality metrics to impose constraints, enabling the reconstruction of satisfactory images. After that, Rippel and Bourdev [40] innovatively introduced an adversarial training strategy for image compression, enabling the generation of visually rich details in reconstructed images. This trend was further augmented by many other researchers [13], [14], [41], [42], [43], [44], [45], who improved the designs of generators and discriminators, training approaches, perceptual losses and so on. Furthermore, some methods based on Vector Quantized Variational Autoencoder (VQ-VAE) [46], [47], [48], [49] also exhibit superior performance. These methods employ codebook-based representation learning, where the neural network parameters are utilized to store a diverse distribution of quantized feature vectors. This enables the dynamic selection of contextually appropriate representations tailored to the specific characteristics of individual input samples, and performs exceptionally well even at extremely low bit rates. These methods demonstrate improved ability to reconstruct images with better semantic alignment to the original data, while also producing outputs that more closely match human visual perception.

In addition, there are some VQ-Based methods [46], [47], [48] that store diversified distribution through codebook based on neural network parameters, enabling the selection of suitable representations for different samples. However, due to the limitation of codebook's capacity and optimization strategy, most of these methods are only suitable for ultra-low bitrates, lack the capacity of improving image quality with increased bitrate.

While these methods produce images with good visual quality, they have a crucial drawback. Increasing visual fidelity introduces some uncertainty, which may cause the reconstructed image to differ greatly from the original. It becomes difficult to distinguish between original details and those added to enhance visual quality. Based on GAN model, a notable research [19] emerge that balances the trade-off between fidelity and perceptual quality. This advancement enables users to attain multi-realism results in image reconstruction by adjusting the decoders with factors, which cater to diverse pReferences in image representation. Later, CRDR [20] extend this work to achieve the flexible control of rate, distortion, and realism using consistent model parameters. Though these methods demonstrate impressive performance, their compression capabilities have not been fully explored due to the limitations inherent in GAN models which they used as backbone.

C. Diffusion Model

Diffusion models [21], [22], [23] are a highly powerful class of generative models, defining forward and reverse processes for noise addition and removal, and have become a breakthrough in the field of generation. These models often produce outputs that exhibit superior quality and diversity, and there are many studies exploring the application of diffusion models for generation [50], [51], enhancement [52], [53], [54], [55], [56], [57], understanding [58], [59] and so on.

The great success of diffusion models has also sparked renewed interest in image compression, with several efforts [15], [24], [60], [61], [62], [63], [64], [65], [66] exploring their potential in this domain. Theis et al. [60] took the first step in image compression through diffusion and provided theoretical evidence supporting its feasibility. Following this, Yang and Mandt [15] injected the bitstream as a condition into the diffusion model, groundbreaking achieving the practical image compression. Furthermore, Ma et al. [61] enhanced the performance by correcting the diffusion process with an end-to-end compression model. Additionally, some research [56], [57], [67] explores semantic image compression by integrating diffusion-based large multi-modal models with large language models, and efforts using diffusion as a post-processing module [24], [62], [63] also demonstrates the potential of diffusion in the realm of image compression. However, as we have previously emphasized, the inherent uncertainty in the diffusion process poses challenges to achieving high-fidelity reconstructions, which makes it difficult to achieve multi-realism image compression that takes into account both perceptual quality and fidelity, requiring further exploration. In this paper, we introduce an innovative solution that solves the challenge by integrating supplementary consistency guidance mechanisms and a novel neural syntax-driven approach. By utilizing the power of neural syntax and augmenting it with

consistency guidance, our method advances the frontiers of multi-realism image compression.

III. Consistency Guided Diffusion Model

In this part, we first describe the foundation of diffusion models while underscoring our key motivations in Section III-A, followed by a detailed description of our proposed consistency guided diffusion model which integrates consistency guidance in Section III-B. Expanding our framework further, Section III-C introduces our innovative Syntax-Driven Feature Fusion Module. In Section III-D, we introduce a refined warm-up-based training strategy. Finally, our sample-adaptive continuous online optimization during inference is described in Section III-E.

A. Preliminaries and Motivations

We start with an analysis of the diffusion model. As a kind of generative model, diffusion models have shown their remarkable ability to create images with excellent perceptual quality by leveraging a conditional model that incorporates latent features. Numerous types of research [52], [53], [54], [55] employ diffusion models as a post-processing or enhancement component. These studies typically utilize the degraded image $\tilde{x}$ as a condition, crafting a conditional model to learn the underlying data distribution $p(x|\tilde{x})$ via a fixed multi-step chain with T iterations. The diffusion paradigm is based on a forward process q, which progressively injects Gaussian noise into the image. Mathematically, the progression of this forward process is expressed as:

$$\begin{aligned} q(x_t|x_0) &= \mathcal{N}(x_t|\alpha_t x_0, \sigma_t^2\mathbf{I}), \\ q(x_T) &= \mathcal{N}(x_T|\mathbf{0}, \mathbf{I}), \end{aligned} \tag{1}$$

where α_t and σ_t^2 are hyper-parameter functions of t.

The inference process unfolds as a reverse process from an initial state of pure Gaussian noise $q(x_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ towards the desired target x_0, which can be formally described as:

$$\begin{aligned} p(x_T) &= \mathcal{N}(x_T|\mathbf{0}, \mathbf{I}), \\ p(x_{t-1}|x_t, \tilde{x}) &= \mathcal{N}(x_{t-1}|\mu_\theta(\tilde{x}, x_t, t), \sigma_t^2\mathbf{I}), \end{aligned} \tag{2}$$

where $\mu_\theta(\tilde{x}, x_t, t)$ represents the mean value of the conditional probability distribution $p_\theta(x_{t-1}|x_t, \tilde{x})$, and the diffusion model is trained to acquire an understanding of these conditional distributions through parametric approximation, thereby enabling a precise characterization of the underlying distribution. From the level of neural network structure, numerous existing diffusion-based post-processing frameworks adopt a straightforward approach by directly incorporating the condition $\tilde{x}$, the noise x_t, and the timestamp t as inputs into a U-Net backbone, similar to the vanilla Denoising Diffusion Probabilistic Model (DDPM) and predict the noise $\hat{\epsilon}_t$ at each iterative step.

However, this paradigm still faces two noteworthy issues, particularly when efforts are made to extend its application to multi-realism image compression:

- The diffusion process often leads to a deviation of the final reconstructed image x_0 from the initial condition $\tilde{x}$, which poses a significant hinder in achieving a high-fidelity reconstruction of the original image.
- Since the condition $\tilde{x}$ represents a degraded image with missing information, it can correspond to multiple possible original images. This makes the task of ensuring that the reconstructed image closely similar to the true original difficult. Additionally, controlling the balance between realism and fidelity during the diffusion process is challenging, making it more difficult to clearly define and optimize the target probability distribution $p(x|\tilde{x})$.

In this paper, we provide an innovative framework to address both of these issues for multi-realism image compression:

- Addressing the first issue, we integrate a supplementary consistency guidance within the network architecture of the diffusion model, called the consistency guided diffusion model, which not only mitigates the deviations but also augments the fidelity of the reconstructed image.
- For the second issue, we devise a novel syntax-driven feature fusion strategy that encodes an additional ultra-low bit stream during the encoding phase. This supplementary bit-stream serves as a carrier for extracted syntax prior information and optimization target, thereby easing the ambiguity of inference target during the post-processing.

B. Overall Structure

Our compression framework is structured into two components: an end-to-end image compression model and a diffusion-based post-processing module called consistency guided diffusion model, with the structure shown in Fig. 1.

1) End-to-End Image Compression Model: We first utilize a state-of-the-art end-to-end image compression framework Transformer-CNN Mixed architecture [30], which is renowned for its exceptional performance in distortion-oriented compression tasks. This framework enables us to apply lossy compression on the original image x, resulting in a compressed representation $\tilde{x}$ that effectively captures the main content of the input image despite some inevitable loss of details.

2) Diffusion-Based Post-Processing Model: Our diffusion-based post-processing framework follows an encoder-decoder paradigm augmented with skip connections like U-Net architecture [68]. Differently, our structure comprises two encoders and a single decoder informed by prior work in diffusion models [69].

The upper encoder branch, as shown in Fig. 2, transforms the noisy input image into a set of N multi-resolution diffusion feature maps d_i, each corresponding to a different scale, where N represents the depth of the U-Net backbone and $i \in \{0, \ldots, N\}$. In parallel, the lower encoder branch extracts feature maps e_i from the degraded image $\tilde{x}$, maintaining consistency in scale with the d_i maps.

The decoder part of our model integrates a set of syntax-driven feature fusion modules. This module, detailed in the subsequent section, utilizes an ultra-low bitrate bitstream s to guide the fusion process. Starting with the fusion of d_N and e_N to yield u_N, the SFF module iteratively fuses the corresponding features u_{i+1} and e_i at each decoder layer, then fuses the output of SFF with the corresponding skip connection d_i, producing u_i. This hierarchical fusion and up-sampling process is iterated continuously to achieve the purpose of predicting the output noise.

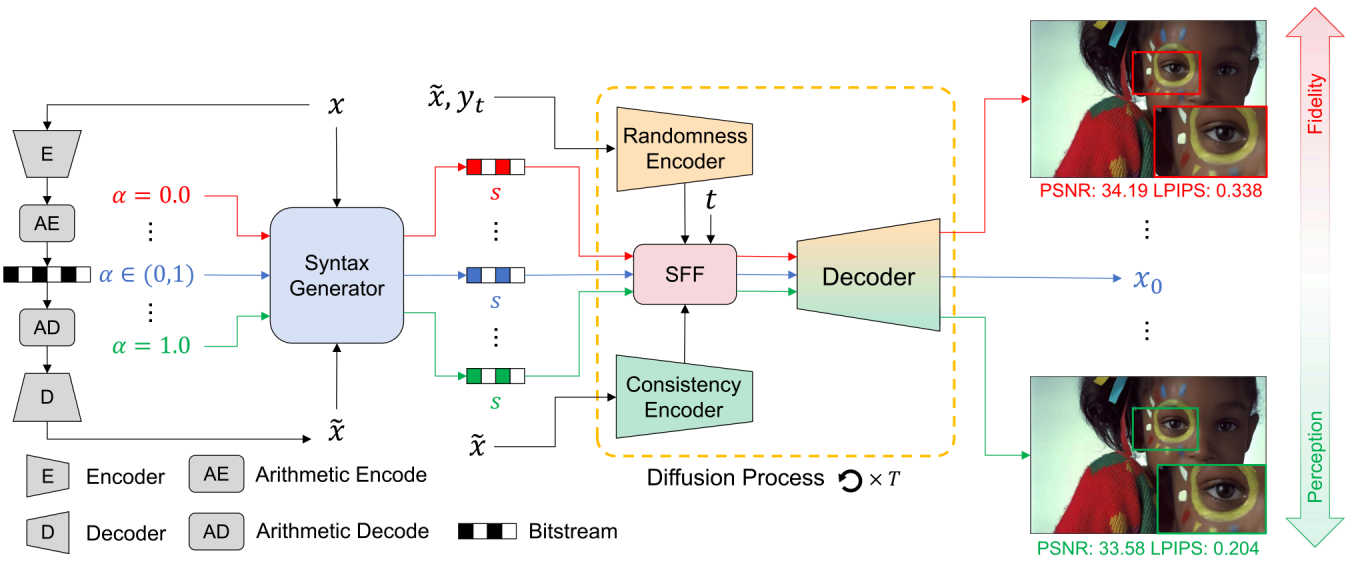


Fig. 1. The overall structure of our proposed method. The blue path represents a general sample of how our approach is utilized in image compression, while the red and green marked paths demonstrate two extremes sample, with the red path emphasizing high fidelity and the green path prioritizing perception. Starting with an input image x for encoding, we first obtain $\tilde{x}$ with a fidelity-driven lossy compression method. Then, we extract a syntax vector s with a realism control factor α for controlling the trade-off between fidelity and perception. After a complete T steps diffusion process with noise y_t at each step, we obtain a higher-quality reconstructed image x_0. x_0 can display different levels of realism with different injected the control factor α, achieving multi-realism image compression.

By taking this methodology, we steadily inject a consistency guidance feature extracted from the degraded image into the denoising diffusion process of our post-processing model. This feature guides the model's inference to stay close to the conditional distribution of the degraded image, and we achieve a final output that enhances perceptual quality while maintaining fidelity to the original image, striking a good balance between both of them.

C. Syntax Driven Feature Fusion

As shown in Section III-A, the target probability distribution $p(x|\tilde{x})$ optimized by the post-processing module is ambiguous. To address this ambiguity and ensure that the reconstructed image x_0 is in line with our expectations for the fidelity-perceptual quality trade-off, we propose transmitting a compact syntax vector utilizing an ultra-low bitstream for imparting the syntactic information and the optimization target. Based on this idea and drawing from the inspiration of [8], we encode the syntax prior of the original image into a compact, one-dimensional vector using a syntax generator. This syntax vector acts as a dynamic convolution kernel in our syntax driven feature fusion module. By decoding this syntax vector into kernels and applying them to the features for fusion, we naturally integrate syntax information into the neural processing flow. The structures of the syntax generator and syntax driven feature fusion module are shown in Fig. 2.

1) Syntax Generator: The architecture of our syntax generator follows the design principles outlined in [8], incorporating a multi-scale framework grounded in the hyper-priors entropy model [4], [31]. For capturing both the rich syntax context of the image and the precise optimization objective into the syntax vector, we incorporate both pre- and post- lossy compression image representations into the syntax generator's input, and inject a control factor α to provide guidance to the optimization target. In implementation, the control factor α is injected by expanding to match the dimensions of the input image and cascading with them. Then, this network strategically uses global average pooling at each scale to compress the features into a compact, one-dimensional vector. This strategic approach not only makes use of the richness of multi-scale information but also ensures the global semantic consistency, thereby enhancing the overall efficiency and robustness of the syntax generation process.

2) Syntax Driven Feature Fusion: The syntax-driven feature fusion module simply takes in features d_i, e_i, the syntax vector s, and the current timestamp t. By concatenating s and t, we use a fully connected network to generate two dynamic convolutional kernels, W_e^i and W_d^i. These kernels then independently convolve with e_i and d_i, achieving adaptive fusion of features at each layer:

$$u_i = W_e^i * e_i + W_d^i * d_i, \tag{3}$$

where * denotes convolution. Since the generation of these kernels relies on the original image through the input of syntax vector, the fusion process is able to capture the image's unique characteristics, allowing for a more tailored and effective combination of the two features during reconstruction, resulting in an improved final image.

D. Warm-Up Based Training Strategy

In the task of multi-realism compression, the model's optimization objective dynamically shifts in response to the injected control factor α, which poses a challenge to the

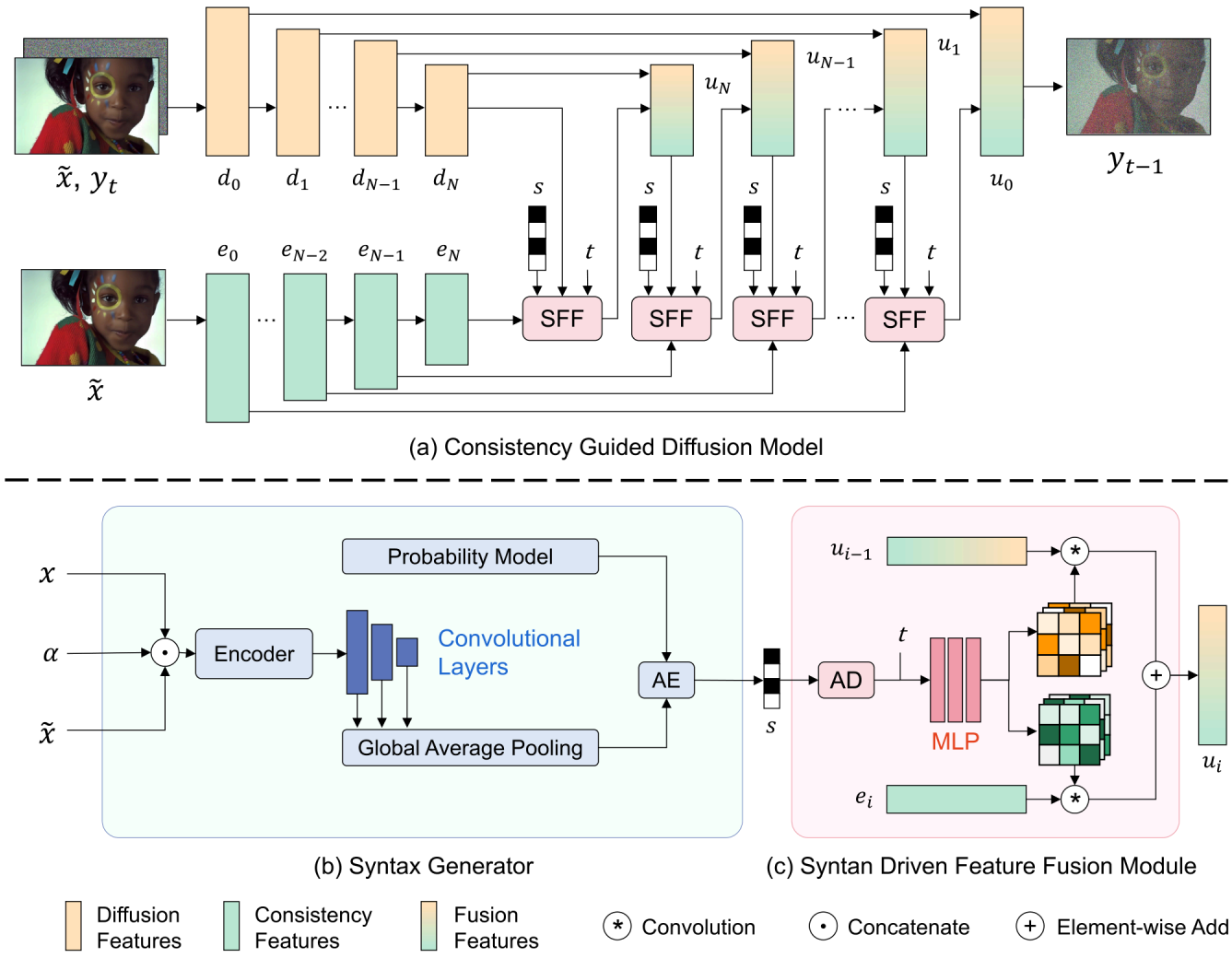


Fig. 2. Detailed structures of the consistency guided diffusion model (top), Syntax generator (bottom left), and syntax driven feature fusion module (bottom right). In the consistency guided diffusion model, features d and e are fused and got u utilizing the SFF Module guided by syntax s. The syntax generator module serves as the extractor of the global syntax vector s from the image x and $\tilde{x}$ with a controlling factor α during encoding, whereas the SFF module plays the role of adaptively integrating consistency and diffusion features within the diffusion framework.

training process. To address this challenge, we devise a two-stage warm-up based training strategy. The first-stage warm-up training stage aims for a warm start for the diffusion model by constraining the noise under a unified optimization objective, and the following training stage further optimizes the diffusion model with a variable optimization objective by imposing constraints on the sampled clean images.

1) Warm-Up Training Phase: Our warm-up training phase is similar to the conventional training strategy for other diffusion models [21], focusing on the estimation of noise. In difference, we introduce a bit-rate control term, which acts as a constraint on the bit cost associated with syntax vector, and enables more accurate control over the trade-off between compression efficiency and reconstruction quality. The resulting loss function for the warm-up training phase is formulated as follows:

$$\mathcal{L} = \|\epsilon_t - \mu_\theta(x_t, \tilde{x}, s, t, \alpha)\|_2^2 + \lambda_c R(s), \tag{4}$$

where μ_θ represents our noise prediction network CGDM+, ϵ_t signifies the noise added at step t during the forward diffusion process, $R(s)$ denotes the bit rate cost following the calculation method in [30], and hyper-parameter λ_c is used to balance the rate-distortion trade-off. In this stage, the control factor α is set to a fixed value of 0.5.

2) Official Training Phase: After undergoing a warm-up training process, our diffusion network acquire the capability to reconstruct images. In the following official training stage, we train it to learn how to adjust the perception-fidelity trade-off with the trade-off control factor α. In conventional diffusion model training strategy, its optimization modeling derivation is based on the KL divergence and constraints are imposed on the predicted noise to optimize the diffusion model, which limit the direct distortion-perception trade-off optimization on the reconstructed image. For solving this problem, we adopt a novel rate-distortion-perception trade-off optimization strategy for diffusion models, inspired by [70]. Specifically, we fix the diffusion sampling strategy to a 4-step DDIM sampling

approach. During each step of the training process, we perform complete sampling to obtain the final sampled image denoted as $\hat{x}_0$. In this phase, the injected control factor α at each training iteration will be randomly valued from 0 to 1. Then, we apply constraints to the final sampled image $\hat{x}_0$, and the loss function can be formulated as:

$$\begin{aligned}\mathcal{L} &= \lambda_d \mathcal{L}_d(\hat{x}_0, x) + \alpha \cdot \lambda_p \mathcal{L}_p(\hat{x}_0, x) + \lambda_f R(s),\\ \mathcal{L}_d &= |\hat{x}_0 - x|,\\ \mathcal{L}_p &= \frac{1}{3}(DISTS(\hat{x}_0, x) + LPIPS_{Alex}(\hat{x}_0, x)\\ &\quad + LPIPS_{VGG}(\hat{x}_0, x)),\end{aligned} \tag{5}$$

where $\mathcal{L}_d$ represents the distortion loss, $\mathcal{L}_p$ denotes the perception loss, λ_d, λ_p and λ_f are the hyper-parameters, α has the same value with the injected control factor. Subsequently, the gradients computed by the loss function are back-propagated through the complete 4-step DDIM sampling process, thereby optimizing the entire diffusion process and enabling image-level constraints to refine the diffusion model. By doing this, the control factor α is used to regulate the trade-off between realism and fidelity in the loss function and the network is able to learn the correlation between the control factor and the optimization goal.

E. Sample-Adaptive Continuous Online Optimization

We contend that the sample adaptation of our method underscores the potential for a more flexible design during the inference phase to utilize its full performance capabilities and attain even better results. Drawing inspiration from rate-distortion optimization via the online mode decision from conventional image compression techniques, we introduce a sample-adaptive continuous online optimization mechanism, tailored to adapt the diverse resolutions and styles of images, and achieve more accurate perception-fidelity trade-off control. Through online optimization, the extracted syntax vector can be more suitable for its expected fidelity, and the performance can be further enhance.

Similar with prior studies such as [5] and [8], our approach utilizes a generator to extract global syntax information from images. This allows for online optimization during encoding at inference time. This is similar to the mode decision process in traditional hybrid coding frameworks, where a best-fit mode is picked from a predefined set. However, by incorporating iterative optimization, we extend this to a continuous selection from a virtually limitless range of options, thereby enhancing the flexibility and potential of our online optimization strategy. Specifically, during the inference phase for each individual image, we iteratively refine the encoder parameters of the syntax generator on randomly sampled patches. This iterative optimization strategy enhances the generator's ability to produce syntax vectors that are increasingly aligned with the image's semantics and the expected realism. During the iterative optimization in inference, the optimization objective remains consistent with the formula 5 described previously.

TABLE I
DETAILED STRUCTURE OF OUR MODEL

Model Size	124.63M
Inner Channel	64
Channel Multiplier	[1,2,2,3]
Depth	2
Dropout	0.2
Attention Resolution	None

IV. EXPERIMENTS

A. Implementation

1) Network Implementation: We implement our diffusion model based on the architecture of [69] while reducing the parameters. To further minimize video memory consumption, we strategically delete the self-attention module. The detailed structure of our proposed Consistency Guided Diffusion Model is demonstrated in Table I.

2) Training Details: We utilize the DIV2K dataset [71] as our training dataset, which consisting of 800 high-quality natural images with an average resolution of 2K. To enhance our model's robustness and adaptation to diverse image resolutions, we incorporate an augmentation strategy by down-sampling each image to half of its original resolution. During the training phase, we adopt a random approach, randomly extracting 256×256 patches from each image, thereby exposing our model to various spatial contexts and further enhancing its capabilities.

Our training process employ the Adam optimizer [72] with an initial learning rate tuned to 1×10^{-4}. Also, any gradients with norm values exceeding 0.5 are clipped to 0.5 or -0.5 to avoid exploding gradients. Besides, we use the exponential moving average strategy for more stable training, with decay parameters set to 0.999. The entire training process is conducted on a single NVIDIA GeForce RTX 4090 GPU, using a batch size of 8 during the warm-up training stage and a batch size of 2 during the following official training stage. We train 4 distinct models, each tailored to a unique compression rate. During warm-up training phase, we strategically set the hyper-parameter λ_c to 1×10^{-5}. And in the following official training stage, we train the model by assigning specific values of hyper-parameter λ_d, λ_p and λ_f to 65.0, 34.56 and 1×10^{-5}, respectively. Each model is training with 20k iterations on the warm-up training stage and 50k iterations on the following official training stage.

3) Inference Details: During model training, we utilize patches of a uniform size. However, during inference, feeding images with different resolutions directly into the diffusion model can impair performance due to distribution mismatches. In order to adapt images with any resolution, we employ a tiling method. To minimize the block artifacts from this tiling approach, we adopt a straightforward strategies that overlaps the patches slightly, allowing pixels at the edges to be predicted by multiple patches. This helps smooth out the transitions at the edges, reducing the visibility of tiling artifacts. During the inference phase, we employ a patch size

TABLE II
AVERAGE BD-RATE FOR DIFFERENT METHODS ANCHORED ON BPG. BOLD INDICATES THE BEST PERFORMANCE, WHILE UNDERLINED INDICATING THE SECOND

Category	Methods	LPIPS	PIEAPP	FID	PSNR	VIF	MS-SSIM	Average
Fidelity-Driven	TCM [30]	-24.44	-25.47	+2.637	**-33.29**	-14.57	-17.10	-18.70
	ELIC [6]	-24.68	-28.20	-1.92	-31.54	-13.87	**-19.42**	-19.94
	NLIC [73]	-23.73	-25.07	-1.17	-32.55	-13.52	-16.26	-18.72
	NIF [37]	+42.33	+70.61	+40.88	+81.54	+67.05	+88.58	+65.17
Perceptual-Driven	ILLM [14]	-75.05	**-73.83**	-86.19	+32.71	-3.81	-3.245	-34.90
	HiFiC [13]	-63.19	-47.35	-81.73	+53.74	+5.57	+3.77	-21.53
	HFD [62]	-63.71	-67.46	-81.80	+42.38	+22.52	+40.41	-17.94
	CGDM [24]	**-82.05**	-70.52	**-89.11**	+17.96	-3.39	-8.27	-39.23
Multi-Realism	$\text{CRDR}_{0.00}$ [20]	-40.98	-37.18	-22.01	-21.82	-7.97	-16.26	-24.37
	$\text{CRDR}_{5.12}$ [20]	-70.88	-72.84	-80.95	+3.03	-3.35	+3.62	-36.90
	$\text{MRIC}_{0.00}$ [19]	-57.07	-32.32	-53.48	-9.12	-16.66	-8.239	-29.48
	$\text{MRIC}_{2.56}$ [19]	-68.61	-67.03	-81.98	+20.46	-12.92	+9.94	-33.36
	$\text{CGDM+}_{0.00}$	-29.06	-28.49	-6.14	-33.28	**-15.16**	-17.46	-21.60
	$\text{CGDM+}_{1.00}$	-69.41	-66.16	-75.84	-24.78	-12.31	-7.40	**-42.65**

of 256×256 pixels, with an overlap of 8 pixels at the edges. For each image, starting from the pre-trained network, we fine-tune the encoder using the Adam optimizer with a learning rate of 5×10^{-5} for 100 iterations. The goal of this fine-tuning is to minimize the same loss function as during training, to further improve the encoder's performance specifically for that image.

4) Evaluation Protocol: We assess our method on these three datasets: the Kodak image dataset [74], the professional subset of the CLIC validation dataset [75], and DIV2K validation dataset [71]. The Kodak image dataset consists of 24 images, each with a resolution of 768×512. The CLIC professional validation dataset comprises 41 images and the DIV2K contains 100 images. The images contained within these two datasets exhibit higher resolutions of about 2K. Evaluation on these datasets shows the performance of our approach on images that have higher resolutions.

To underscore the superiority of our approach in terms of both distortion reduction and perceptual quality enhancement, we utilize a comprehensive suite of evaluation metrics. Specifically, for assessing distortion, we adopt PSNR, VIF [76], and MS-SSIM [11], where PSNR quantifying the mean square error (MSE) between the original and compressed images, multi-scale structural similarity (MS-SSIM) [11] comparing patch-level similarities and visual information fidelity (VIF) offer insights into the fidelity of the reconstructed images in four scales. To evaluate perceptual quality, we utilize VGG-based [77] LPIPS [78], PIEAPP [79], and FID [80]. In them, LPIPS [78] calculates the sum of MSE between deep feature pyramids, and offers flexibility in choosing different backbone networks. PIEAPP evaluates image quality using a deep learning model trained on many visual perception datasets. In contrast to image-level metrics, FID [80] captures human perception of image quality beyond only pixel-level accuracy, and stands apart as a widely adopted metric for evaluating the divergence between distributions.

TABLE III
CODE LINKS OF THE COMPARED METHODS

Method	Code Link
BPG [3]	https://bellard.org/bpg/
TCM [30]	https://github.com/jmliu206/LIC_TCM
ELIC [6]	https://github.com/VincentChandelier/ELiC-ReImplemetation
NIF [37]	https://github.com/aegroto/nif
ILLM [14]	https://github.com/facebookresearch/NeuralCompression
HiFiC [13]	https://github.com/Justin-Tan/high-fidelity-generative-compression
CDC [15]	https://github.com/buggyyang/CDC_compression
CRDR [20]	https://github.com/iwa-shi/CRDR
MRIC [19]	https://github.com/Nikolai10/MRIC

We present R-D curves and BD-rate [81] analyses across these various evaluation metrics to compare different methods.

B. Quantitative Comparison

We evaluate our method against a diverse set of compression techniques, encompassing both traditional methods such as BPG [3], and learning-based approaches tailored for Mean Squared Error (MSE) optimization like ELIC [6], TCM [30] and NLIC [73], Implicit Neural Representation (INR)-based image compression methods like NIF [37], alongside perceptual quality-oriented methods including HiFiC [13], ILLM [14], HFD [62] and CDC [15]. Additionally, we also incorporate a comparison with multi-realism compression methods exemplified by MLIC [19] and CRDR [20]. In comparision, for methods with open-source code, we use official implementations, and for MRIC and ELIC, community-maintained versions are employed. For HFD and NLIC, we directly utilize the authors' provided compressed images and bpp values for comparison. The code links for the comparison methods are listed in Table III. Fig. 3 comprehensively presents the Rate-Distortion (R-D) curves of our method alongside many comparative approaches, evaluated on the benchmark datasets

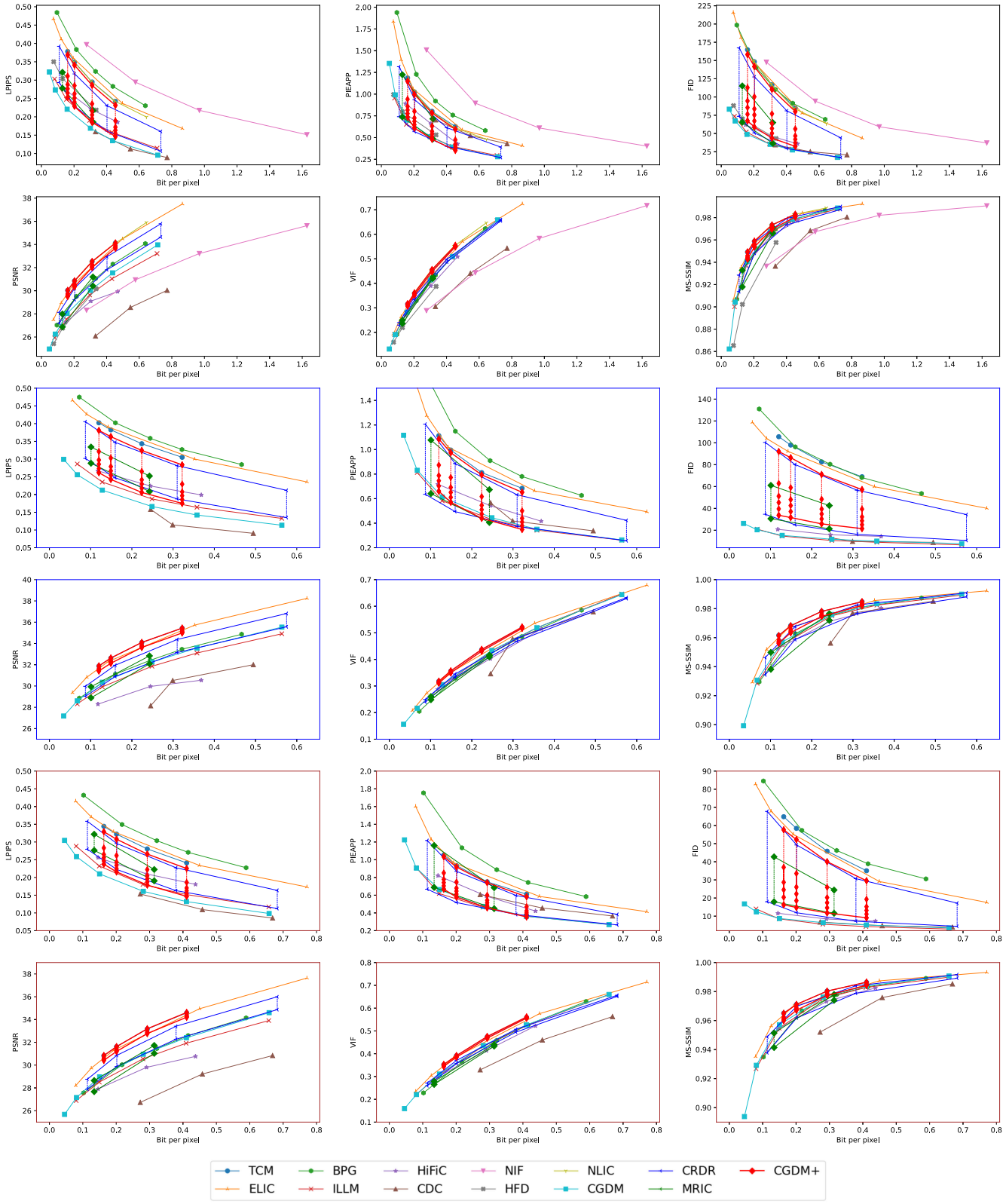


Fig. 3. Trade-offs between bitrate (x-axes, in bpp) and different metrics (y-axes) for various models tested on Kodak, CLIC and DIV2K. We consider both perceptual (LPIPS, PIEAPP, FID) and distortion metrics (PSNR, VIF, MS-SSIM). The upper 2 rows (black frame) are the performance on Kodak datasets, the middle 2 rows (blue frame) are on CLIC professional dataset, and the bottom 2 rows (brown frame) are on DIV2K test dataset.

of Kodak [74], CLIC [75], and DIV2K [71], across various evaluation metrics. Under the control of different syntax vectors taking ultra-low bitstream, our method demonstrates the remarkable capability to reconstruct images across a wide range of fidelity-perceptual trade-offs utilizing a consistent main bitstream. When prioritizing perceptual quality, our

approach demonstrates exceptional performance that is comparable with state-of-the-art perceptual-optimized methods [13], [14], [15], [24], [62], while achieving more desirable fidelity. In contrast, when shifting towards distortion, our method not only aligns with the SOTA fidelity-optimized methods [6], [30], [73] but also excels in perceptual evaluations, showcasing a unique advantage. Furthermore, in comparison to multi-realism methodologies [19], [20], our technique stands out by enabling an even broader range of trade-offs and better performance, especially showing significant advantages than other methods on fidelity.

To show a more intuitive evaluation of our method's performance with other benchmark approaches across all evaluation metrics, we compute the BD-rate [81] for each metric. Leveraging BPG as an anchor, Tab. II showcases the average BD-rate attained by each method on both distortion and perceptual metrics on Kodak datasets. For most methods, we follow the standards [81] to calculate the BD-rate with at least 4 rate points. For some comparison methods that lack enough rate points for calculating, we utilized their shared rate points to fit the RD-curve by quadratic interpolation, enabling comparison with BD-rate. For CRDR, MRIC, and CGDM+ in Tab. II, the smaller the subscript, the more fidelity is preferred, and the larger the subscript, the more perceptual quality is preferred. This comprehensive comparison underscores the superiority of our method in terms of overall performance.

C. Qualitative Comparison

To better demonstrate the perceptual quality achieved by our approach, we showcase several cases in Fig. 4. Notably, our CGDM+, MRIC [19], and CRDR [20] use configuration that mostly favors perceptual quality in comparison. As shown in Fig. 4, conventional compression methods like BPG [3] and mean squared error (MSE)-optimized techniques like NLIC [73] and TCM [30] result in overly smoothed images like the grass in the bottom two rows of Fig. 4, accompanied by a significant loss of details.

Conversely, perception-driven methods, including ILLM [14], HiFiC [13], CDC [15] and so on, tend to compromise the fidelity of the details (such as the tattoo in the middle two rows of Fig. 4) or structure (such as the shape of aircraft tire in the bottom two rows) of the reconstructed image, resulting in significant distortion. In stark contrast, images reconstructed using our method have richer visual detail and fewer artifacts when using fewer or comparable bits. This observation underscores the superior perceptual quality of our compression technique.

D. Time Complexity Analysis

We also present a comprehensive complexity analysis of our proposed method in Table IV, conducted on a single NVIDIA GeForce RTX 4090 GPU. In terms of encoding time, online optimization strategies inherently entail higher computational overhead. However, our method achieves superior encoding efficiency compared to alternative approaches [49], [61] that employ online optimization techniques, while also outperforming the widely adopted image compression standard VVC [10].

TABLE IV
TIME COMPLEXITY FOR A 768 × 512 IMAGE

Method	Enc. Time	Dec. Time
VVC [10]	31.53s	0.06s
IEEE Std 1857.11-2024 [82]	216.19s	244.58s
CorrDiff [61]	10.95s	2.16s
CGOO [10]	36.34s	0.25s
TCM [30]	0.25s	0.24s
Ours	7.77s	0.71s

TABLE V
RESULTS OF THE USER STUDY. BOLD INDICATES THE BEST

Methods	Votes	Percentage
BPG [3]	8	1.33%
ELIC [6]	77	12.83%
TCM [30]	125	20.83%
HiFiC [13]	43	7.17%
ILLM [14]	58	9.67%
CDC [15]	98	16.33%
CGDM [24]	20	3.33%
CGDM+	**171**	**28.50%**

Regarding decoding time, diffusion-based methods also exhibit higher computational complexity. Nevertheless, our approach demonstrates faster decoding performance compared to other diffusion-based methods [61]. Additionally, we conduct a comparative analysis with the application-oriented end-to-end image compression standard IEEE Std 1857.11-2024 [82]. Our method consistently achieves better encoding and decoding times than the standard, highlighting its potential for practical applications.

E. User Studies

In addition to the comprehensive quantitative and qualitative analyses detailed in Sections IV-B and IV-C, we conduct a user study to measure the perceptual quality of images reconstructed by various techniques. This assessment requires gathering participants' preferences for the top-3 most favorably reconstructed images among a pool of eight methods: our perception-driven CGDM+ approach and seven benchmark methods (BPG [3], ELIC [6], TCM [30], HiFiC [13], ILLM [14], CGDM [24] and CDC [15]). The evaluation criteria contain two vital aspects: (1) satisfactory of structural fidelity to the original image — consistency in key contents, and (2) pleasing overall visual quality — no artifacts, color bias, etc. A total of 20 volunteers participate in this study and 600 votes are collected during the evaluation process. As summarized in Tab. V, our proposed CGDM+ outperforms other SOTA image compression methods, demonstrating our approach's ability to create visually appealing images.

Fig. 4. Visual comparisons with state-of-the-art methods. We provide further analysis that focuses on subjective results in the main text. [*Zoom in for best view*].

F. Ablation Studies

In this section, we conduct extensive ablation studies for our proposed network architecture on the Kodak dataset to show the rationality and necessity of the components.

1) Online Optimization: Our initial focus is on the Online Optimization strategy in our approach. By eliminating the online optimization process and only relying on the pre-trained model parameters while inference, the online optimization strategy is removed. The results without the Online

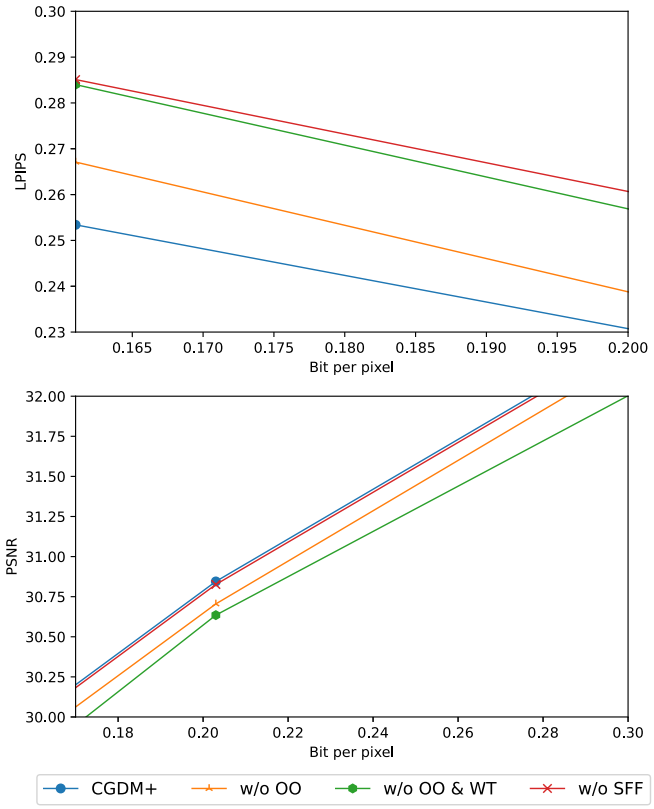


Fig. 5. The ablation study results tested on Kodak dataset. The upper chart demonstrates the trade-offs between bitrate (x-axes, in bpp) and LPIPS (y-axes) with models use the configuration that most favors realism. The bottom chart demonstrates the trade-offs between bitrate (x-axes, in bpp) and PSNR (y-axes) with models use the configuration that most favors fidelity.

Optimization strategy, denoted as w/o OO in Fig. 5, show a narrower range of achievable trade-offs of perception and fidelity in the reconstructed images compared to our full method, demonstrating the role of Online Optimization.

2) Warm-Up Training Strategy: We further study the Warm-up Training Strategy. By removing the warm-up training stage and adding the iterations of the official training stage to 70k iterations, the Warm-up Training Strategy is moved. The results without Warm-up Training Strategy and Online Optimization are denoted as w/o OO & WT in Fig. 5, demonstrating a further reduction in its performance, illustrating the role of Warm-up Training Strategy.

3) Syntax Driven Feature Fusion: Finally, we show the role of Syntax driven Feature Fusion module. By replacing the SFF module with direct element-wise addition, the model fuses information directly without syntax guided feature fusion. Without the SFF module, the model will not be able to adjust the direction of the trade-offs. Therefore, we train two models for fidelity and perception-oriented, marked as w/o SFF in Fig. 5. The experimental results demonstrate that while training with a single optimization objective could enhance performance, the fidelity and perceptual quality of the reconstructed images falls short of CGDM+, attributed to the absence of syntax vector guidance.

V. Conclusion

In this work, a novel consistency guided diffusion model tailored for multi-realism image compression is introduced, carefully balancing the trade-off between subjective visual quality and fidelity. The consistency guidance architecture, neural syntax driven strategy, warm-up-based training strategy, and online optimization technique enable the diffusion model to reconstruct images precisely and meet the required fidelity-perception trade-off. Qualitative and quantitative results demonstrate the superiority of our proposed method.

References

[1] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, Apr. 1991.
[2] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Proc. DCC. Data Compress. Conf.*, 2000, pp. 523–541.
[3] F. Bellard. (2017). *BPG Image Format*. [Online]. Available: http://bellard.org/bpg/
[4] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–23.
[5] G. Lu et al., "Content adaptive and error propagation aware deep video compression," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 456–472.
[6] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5708–5717.
[7] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "MLIC: Multi-reference entropy model for learned image compression," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7618–7627.
[8] D. Wang, W. Yang, Y. Hu, and J. Liu, "Neural data-dependent transform for learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17358–17367.
[9] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[10] B. Bross et al., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.
[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, pp. 1398–1402.
[12] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 675–685.
[13] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 11913–11924.
[14] M. J. Muckley, A. El-Nouby, K. Ullrich, H. Jégou, and J. Verbeek, "Improving statistical fidelity for neural image compression with implicit local likelihood models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 25426–25443.
[15] R. Yang and S. Mandt, "Lossy image compression with conditional diffusion models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 64971–64995.
[16] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.
[17] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1945–1954.
[18] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[19] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, "Multi-realism image compression with a conditional generator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22324–22333.
[20] S. Iwai, T. Miyazaki, and S. Omachi, "Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 2888–2897.

[21] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 33, 2024, pp. 6840–6851.

[22] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–22.

[23] C. Lü, Y. Zhou, F. Bao, J. F. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2022, pp. 5775–5787.

[24] H. Kuang, Y. Ma, W. Yang, Z. Guo, and J. Liu, "Consistency guided diffusion model with neural syntax for perceptual image compression," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 1622–1631. [Online]. Available: https://openreview.net/forum?id=nSUMQhITdd

[25] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[26] D. Minnen et al., "Spatially adaptive image compression using a tiled deep network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2796–2800.

[27] D. Minnen, G. Toderici, S. Singh, S. J. Hwang, and M. Covell, "Image-dependent local entropy models for learned image compression," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2018, pp. 430–434.

[28] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–27.

[29] J. Lee, S.-H. Cho, and S. Beack, "Context-adaptive entropy model for end-to-end optimized image compression," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–20.

[30] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-CNN architectures," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14388–14397.

[31] Y. Hu, W. Yang, and J. Liu, "Coarse-to-fine hyper-prior modeling for learned image compression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 11013–11020.

[32] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Trans. Image Process.*, vol. 30, pp. 3179–3191, 2021.

[33] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7936–7945.

[34] C. Li, S. Yin, C. Jia, F. Meng, Y. Tian, and Y. Liang, "Multirate progressive entropy model for learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7725–7741, Aug. 2024.

[35] Y. Strümpler, J. Postels, R. Yang, L. V. Gool, and F. Tombari, "Implicit neural representations for image compression," in *Proc. Eur. Conf. Comput. Vis.*, 2021, pp. 74–91.

[36] E. Dupont, H. Loya, M. Alizadeh, A. Goliński, Y. W. Teh, and A. Doucet, "COIN++: Neural compression across modalities," *IEEE Trans. Mach. Learn. Res.*, vol. 11, pp. 1–26, Nov. 2022.

[37] L. Catania and D. Allegra, "NIF: A fast implicit image compression with bottleneck layers and modulated sinusoidal activations," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 9022–9031.

[38] T. Ladune, P. Philippe, F. Henry, G. Clare, and T. Leguay, "COOL-CHIC: Coordinate-based low complexity hierarchical image codec," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13469–13476.

[39] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 11517–11529.

[40] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2922–2930.

[41] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Van Gool, "Generative adversarial networks for extreme learned image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 221–231.

[42] S. Dash, G. Kumaravelu, V. Naganoor, S. K. Raman, A. Ramesh, and H. Lee, "CompressNet: Generative compression at extremely low bitrates," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2314–2322.

[43] S. Iwai, T. Miyazaki, Y. Sugaya, and S. Omachi, "Fidelity-controllable extreme image compression with generative adversarial networks," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8235–8242.

[44] D. He et al., "PO-ELIC: Perception-oriented efficient learned image coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1763–1768.

[45] A. El-Nouby, M. J. Muckley, K. Ullrich, I. Laptev, J. Verbeek, and H. Jégou, "Image compression with product quantized masked image modeling," 2022, *arXiv:2212.07372*.

[46] W. Jiang, W. Wang, and Y. Chen, "Neural image compression using masked sparse visual representation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4177–4185.

[47] Q. Mao et al., "Extreme image compression using fine-tuned VQGANs," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2024, pp. 203–212.

[48] N. Xue, Q. Mao, Z. Wang, Y. Zhang, and S. Ma, "Unifying generation and compression: Ultra-low bitrate image coding via multi-stage transformer," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2024, pp. 1–6.

[49] H. Kuang, W. Yang, Z. Guo, and J. Liu, "Cross-granularity online optimization with masked compensated information for learned image compression," *Proc. Int. Conf. Comput. Vis.*, 2025, pp. 16514–16523.

[50] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 8780–8794.

[51] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3813–3824.

[52] Y. Ma, H. Yang, W. Yang, J. Fu, and J. Liu, "Solving diffusion ODEs with optimal boundary conditions for better image super-resolution," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–19.

[53] S. Panagiotou and A. S. Bosman, "Denoising diffusion post-processing for low-light image enhancement," 2023, *arXiv:2303.09627*.

[54] B. Fei et al., "Generative diffusion prior for unified image restoration and enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9935–9946.

[55] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12268–12277.

[56] M. Careil, M. J. Muckley, J. Verbeek, and S. Lathuilière, "Towards image compression with perfect realism at ultra-low bitrates," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–21.

[57] C. Li et al., "CMC-bench: Towards a new paradigm of visual signal compression," 2024, *arXiv:2406.09356*.

[58] Z. Li, Q. Zhou, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Open-vocabulary object segmentation with diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1–21.

[59] J. Wang et al., "Diffusion model is secretly a training-free open vocabulary semantic segmenter," 2023, *arXiv:2309.02773*.

[60] L. Theis, T. Salimans, M. D. Hoffman, and F. Mentzer, "Lossy compression with Gaussian diffusion," 2022, *arXiv:2206.08889*.

[61] Y. Ma, W. Yang, and J. Liu, "Correcting diffusion-based perceptual image compression with privileged end-to-end decoder," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 1–19.

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

[63] N. F. K. M. Ghouse, J. Petersen, A. J. Wiggers, T. Xu, and G. Sautiere. (2022). *Neural Image Compression With a Diffusion-Based Decoder*. [Online]. Available: https://openreview.net/forum?id=4Jq0XWCZQel

[64] Z. Li, Y. Zhou, H. Wei, C. Ge, and A. Mian, "RDEIC: Accelerating diffusion-based extreme image compression with relay residual diffusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 11, pp. 11540–11552, Nov. 2025.

[65] Z. Li, Y. Zhou, H. Wei, C. Ge, and J. Jiang, "Toward extreme image compression with latent feature guidance and diffusion prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 1, pp. 888–899, Jan. 2025.

[66] M. Cao et al., "Generative probabilistic entropy modeling with conditional diffusion for learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 9, pp. 9443–9459, Sep. 2025.

[67] C. Li et al., "MISC: Ultra-low bitrate image semantic compression driven by large multimodal model," 2024, *arXiv:2402.16749*.

[68] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent*, 2015, pp. 234–241.

[69] S. Gao et al., "Implicit diffusion models for continuous super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10021–10030.

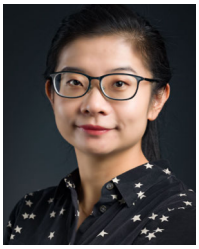
[70] L. Wang, Q. Yang, C. Wang, W. Wang, J. Pan, and Z. Su, "Learning a coarse-to-fine diffusion transformer for image restoration," 2023, *arXiv:2308.08730*.

[71] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1122–1131.
[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[73] Z. Ge, S. Ma, W. Gao, J. Pan, and C. Jia, "NLIC: Non-uniform quantization based learned image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 9647–9663, Oct. 2024.
[74] E. Kodak. (2024). *Kodak Lossless True Color Image Suite*. [Online]. Available: https://r0k.us/graphics/kodak/
[75] G. Toderici et al., "Workshop and challenge on learned image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2020. [Online]. Available: https://archive.compression.cc/challenge/
[76] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[78] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[79] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.
[80] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6626–6637.
[81] B. Gisle, *Calculation of Average PSNR Differences Between RD Curves*, document VCEG-M33, 2001.
[82] *IEEE Standard for Neural Network-Based Image Coding*, Standard 1857.11-2024, 2024, pp. 1–159.

Haowei Kuang (Graduate Student Member, IEEE) received the B.S. degree in computer science from the Southern University of Science and Technology, Guangdong, China, in 2023. He is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include image compression and image enhancement.

Wenhan Yang (Member, IEEE) received the B.S. and Ph.D. (Hons.) degrees in computer science from Peking University, Beijing, China, in 2012 and 2018, respectively. He is currently an Associate Researcher with the Peng Cheng Laboratory, Shenzhen, Guangdong, China. He has authored over 50 technical articles in refereed journals and proceedings and holds nine granted patents. His current research interests include image/video processing/restoration, bad weather restoration, and human–machine collaborative coding. He received the 2023 IEEE Multimedia Rising Star Runner-Up Award, the IEEE ICME-2020 Best Paper Award, the IFTC 2017 Best Paper Award, the IEEE CVPR-2018 UG2 Challenge First Runner-Up Award, and the MSA-TC Best Paper Award of ISCAS 2022. He was a Candidate of the CSIG Best Doctoral Dissertation Award in 2019. He served as the Area Chair for IEEE ICME-2021/2022/2023/2024, the Session Chair for IEEE ICME-2021, and the Organizer for IEEE CVPR-2019/2020/2021 UG2+ Challenge and Workshop.

Zongming Guo (Senior Member, IEEE) received the B.S. degree in mathematics and the M.S. and Ph.D. degrees in computer science from Peking University, Beijing, China, in 1987, 1990, and 1994, respectively. He is currently a Professor with the Wangxuan Institute of Computer Technology, Peking University. His current research interests include video coding, processing, and communication. He is an Executive Member of China-Society of Motion Picture and Television Engineers. He was a recipient of the First Prize of the State Administration of Radio Film and Television Award in 2004, the First Prize of the Ministry of Education Science and Technology Progress Award in 2006, the Second Prize of the National Science and Technology Award in 2007, the Wang Xuan News Technology Award and the Chia Tai Teaching Award in 2008, the Government Allowance granted by the State Council in 2009, and the Distinguished Doctoral Dissertation Advisor Award of Peking University in 2012 and 2013.



Jiaying Liu (Fellow, IEEE) received the Ph.D. degree (Hons.) in computer science from Peking University, Beijing, China, 2010. She is currently a Professor with the Wangxuan Institute of Computer Technology, Peking University, China. She has authored more than 100 technical articles in refereed journals and proceedings and holds 70 granted patents. Her current research interests include multimedia signal processing, compression, and computer vision. She was a member of the Multimedia Systems and Applications Technical Committee (MSA TC) and the Visual Signal Processing and Communications Technical Committee (VSPC TC) in the IEEE Circuits and Systems Society. She was also an ACM ICMR Steering Committee Member and the CAS Representative at the ICME Steering Committee. She received the IEEE ICME 2020 Best Paper Award and the IEEE MMSP 2015 Top10% Paper Award. She was the Technical Program Chair of ACM MM Asia in 2023 and 2025, IEEE ICME in 2021, ACM ICMR in 2021, and IEEE VCIP in 2019. She served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS SYSTEMS FOR VIDEO TECHNOLOGY, and *Journal of Visual Communication and Image Representation*. She was an APSIPA Distinguished Lecturer from 2016 to 2017.