MoE-DiffIR: Task-customized Diffusion Priors for Universal Compressed Image Restoration

Yulin Ren[1], Xin Li[1] (✉), Bingchen Li[1], Xingrui Wang[1], Mengxi Guo[2], Shijie Zhao[2], Li Zhang[2], and Zhibo Chen[1] (✉)

[1] University of Science and Technology of China, Hefei, Anhui, China
[2] Bytedance Inc., Beijing, China
{renyulin, lbc31415926, wxrui_18264819595}@mail.ustc.edu.cn
{xin.li, chenzhibo}@ustc.edu.cn, nicolasguo@pku.edu.cn
{zhaoshijie.0526, lizhang.idm}@bytedance.com

Abstract. We present MoE-DiffIR, an innovative universal compressed image restoration (CIR) method with task-customized diffusion priors. This intends to handle two pivotal challenges in the existing CIR methods: (i) lacking adaptability and universality for different image codecs, *e.g.*, JPEG and WebP; (ii) poor texture generation capability, particularly at low bitrates. Specifically, our MoE-DiffIR develops the powerful mixture-of-experts (MoE) prompt module, where some basic prompts cooperate to excavate the task-customized diffusion priors from Stable Diffusion (SD) for each compression task. Moreover, the degradation-aware routing mechanism is proposed to enable the flexible assignment of basic prompts. To activate and reuse the cross-modality generation prior of SD, we design the visual-to-text adapter for MoE-DiffIR, which aims to adapt the embedding of low-quality images from the visual domain to the textual domain as the textual guidance for SD, enabling more consistent and reasonable texture generation. We also construct one comprehensive benchmark dataset for universal CIR, covering 21 types of degradations from 7 popular traditional and learned codecs. Extensive experiments on universal CIR have demonstrated the excellent robustness and texture restoration capability of our proposed MoE-DiffIR. The project can be found at https://renyulin-f.github.io/MoE-DiffIR.github.io/.

Keywords: Compressed Image Restoration · Mixture-of-Experts · Prompt Learning · Stable Diffusion

1 Introduction

Image compression has emerged as a ubiquitous and indispensable technique in human life and industrial applications, aiming to reduce the costs of image transmission and storage. Existing image codecs can be roughly divided into two categories: (i) traditional image codecs [5, 8, 33, 57, 59], which are designed based on elaborate pre-defined transform, and coding modes, *e.g.*, JPEG [59],

(✉) Corresponding authors.

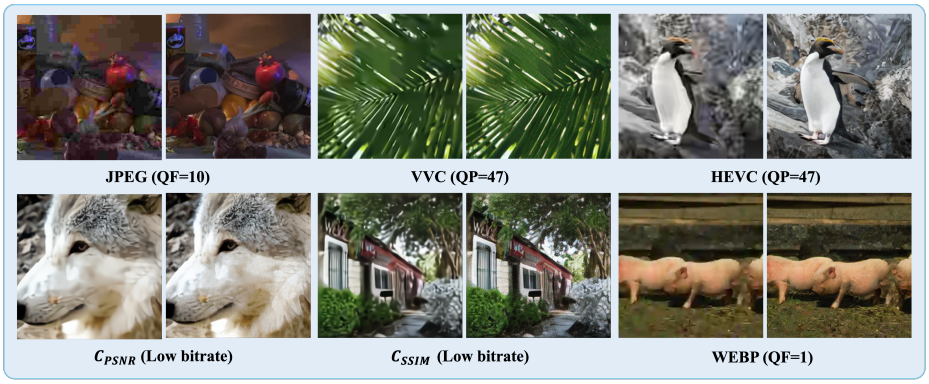


Fig. 1: Visualization of restored compressed images with our MoE-DiffIR on various image codecs and coding modes. Our method can restore diverse compressed images at low bitrates through a single network while possessing high texture generation capability.

BPG [76], and WebP [17] etc; (ii) learned end-to-end image codec [1,19,68], where rate-distortion optimization is achieved with learnable non-linear transform, soft quantization, entropy coding, and other techniques. Despite the substantial success, the compressed images inevitably encounter severe compression artifacts, such as blur, color shift, and block artifacts at low bitrate, which brings an unpleasing visual experience as shown in Fig. 1.

To remove the complicated compression artifacts, the Compressed Image Restoration (CIR) task has been extensively investigated by a series of pioneering studies [15, 49, 72], focusing on the design of the restoration network. Based on advanced Convolution Neural Networks (CNN) [26] and Transformer [58] architecture, some works [6, 11, 13, 22, 28, 31, 71] achieved excellent objective performance (*e.g.*, PSNR, SSIM) on JPEG artifact removal. However, as shown in Fig. 1, these works overlooked two essential challenges in the current CIR task: (i) numerous image codecs and coding modes, leading to diverse compression artifacts. For instance, at low bitrate, the JPEG codec tends to produce blocking artifacts, whereas the learned codecs, *e.g.*, C_{PSNR} [7] is susceptible to more blur artifacts. This raises an urgent requirement for an all-in-one/universal CIR method; (ii) unsatisfied texture recovery due to the lack of generation priors in low-quality images and CIR models.

To address the above challenges, we aim to achieve an all-in-one CIR model by excavating diffusion priors from Stable Diffusion (SD) [3, 10, 23, 38, 61, 74]. Notably, existing works have shown the superior applicability of stable diffusion in image restoration, *e.g.*, StableSR [61], and DiffBIR [38], which reuse the generation priors of diffusion models for a specific task by introducing the modulation module, like ControlNet [80], feature adapter [47, 75, 82]. Nonetheless, the above approaches are inadequate for effectively modulating diffusion models for multiple CIR tasks with shared modulation parameters. Recently, prompt learn-

ing has demonstrated its potential and efficiency for universal image restoration framework [3,18,35,40,43,50]. Inspired by this, we explore how to utilize prompt learning to simultaneously excavate diffusion priors within Stable Diffusion for multiple CIR tasks.

In this work, we present MoE-DiffIR, the first all-in-one diffusion-based framework for universal compressed image restoration with prompt learning. Particularly, the various CIR tasks usually own distinct degradation forms due to different image codecs/modes. This entails the requirement of the task-customized diffusion priors for each CIR task from Stable Diffusion. To this end, we propose the advanced Mixture-of-Experts (MoE) Prompt module, which takes advantage of MoE [44,55,83] to enable dynamic prompt learning for multiple CIR tasks with fixed few prompts. Concretely, we set a series of basic prompts as degradation experts, and design the degradation-aware router to customize the modulation scheme for each task by adaptively selecting top K prompts. In contrast to single prompt or multiple weighted prompts in [3,40,43,50], our MoE-Prompt enables each prompt to perceive different degradations and improve the parameter reuse.

It is noteworthy that Stable Diffusion possesses a rich text-to-image generation prior, which is usually overlooked by existing IR works [38,61]. To activate and reuse these cross-modality priors, we introduce the visual-to-text adapter. Particularly, the CLIP visual encoder is exploited to extract the visual embedding from low-quality images, and the visual-to-text adapter is responsible for transforming the visual embedding into corresponding textual embedding for the guidance of Stable Diffusion. Considering that the low-quality image might damage the extracted visual embedding, we utilize several transform layers as the quality enhancer before the CLIP visual encoder. To validate the effectiveness of our MoE-DiffIR, we construct the first benchmark for the universal CIR task by collecting 7 commonly used image codecs, including 4 traditional codecs and 3 learnable codecs, each with three levels of compression, resulting in 21 types of degradations. Extensive experiments on the universal CIR task have shown the superiority of our MoE-DiffIR in terms of improving perceptual quality and enhancing the robustness for various compression artifacts.

The main contributions of this paper are as follows:

- We propose the first all-in-one diffusion-based method for universal compressed image restoration (CIR) by extracting the task-customized diffusion priors from Stable Diffusion for each CIR task.
- Based on the Mixture-of-Experts (MoE), we propose the MoE-Prompt module to enable each prompt expert to perceive the different degradation and cooperate to extract task-customized diffusion priors. Moreover, we active and reuse the cross-modality generation priors with our proposed Visual-to-Text adapter, which further uncovers the potential of stable diffusion.
- We construct the first dataset benchmark for the CIR tasks, consisting of 7 typical traditional and learned image codecs/modes, each with 3 compression levels, resulting in 21 types of degradation tasks.

– Extensive experiments on 21 CIR tasks have validated the effectiveness of our proposed MoE-DiffIR in improving the perceptual quality and the excellent robustness for unseen compression artifacts.

2 Related Work

2.1 Compressed Image Restoration

Compressed image restoration (CIR) aims to restore compressed images generated by different codecs at varying bitrates. Existing CIR methods typically employ CNN-based [11,13,22,34] or Transformer-based approaches [6,27,36,62]. QGAC [13] and FBCNN [22] are typical CNN-based methods that predict quality factors of compressed images to achieve blind restoration of JPEG codecs. The work [62] proposes an unsupervised compression encoding representation learning method specifically for JPEG, improving generalization in the JPEG domain. However, these methods primarily aim to enhance the objective quality of the restored images and have poor perceptual quality at extremely low compression bitrates. Additionally, they only target a specific compression codec like JPEG, lacking generality in practical applications.

2.2 Diffusion-based Image Restoration

The impressive generative capabilities of diffusion models hold potential for various visual domains, including low-level vision tasks [24,25,46,70]. Diffusion-based image restoration (IR) methods can be divided into two categories [32]: supervised IR methods [30, 41, 42, 53, 69] and zero-shot IR methods [9, 25, 52, 65, 66]. Recently, some works [18, 23, 38, 61] have attempted to fine-tune pre-trained SD models to extract diffusion priors for real-world image restoration. The pioneering work in this area is StableSR [61], which fine-tunes a pre-trained Stable Diffusion model with a time-aware encoder for image restoration in real-world scenes. Another method is DiffBIR [38], which combines SwinIR [36] to first perform coarse-level restoration of distorted images and then utilizes Stable Diffusion with ControlNet [80] for details refinement. PASD [74] attempts to employ pre-trained BLIP and ResNet models to extract high-level information from low-quality images to directly guide the Stable Diffusion restoration.

2.3 Prompt Learning in Image restoration

Recently, prompt learning has significantly influenced the fields of language and vision [21,48,84]. Several studies have begun applying prompts to low-level tasks, with PromptIR [50] being a notable example. This work extends Restormer [78], introducing a set of prompts to identify different distortions, and uses soft weights to manage these prompts for all-in-one image restoration. Another pioneering work is ProRes [43], which employs a singular image-like prompt to interact with various distortions. Additionally, PIP [35] suggests a dual-prompt

approach: one type for universal texture perception and another suite for different degradation types, similar to the weighting approach of PromptIR. In diffusion-based methods, DACLIP [40] also incorporates multiple prompts with soft weight combinations at each time step, facilitating multi-task learning.

Unlike previous prompt-based methods, this paper leverages the concept of routers within the Mixture of Experts (MoE) framework, treating different prompts as experts for routing. It schedules combinations of prompts based on different distortion tasks. In this way, basic prompts can cooperate to fully excavate the task-customized diffusion priors for multiple CIR tasks.

3 Proposed Method

We propose a novel framework dubbed MoE-DiffIR for universal compressed image restoration. Firstly, we review the concept of Mixture-of-Experts and Stable Diffusion in Sec. 3.1. In order to fully excavate the task-customized diffusion priors from stable diffusion, we propose a mixture of experts prompt module illustrated in Sec. 3.2. Meanwhile, we design the visual-to-text adapter for MoE-DiffIR in Sec. 3.3 to generate more realistic and consistent texture. Additionally, we introduce the entire framework and fine-tuning process of MoE-DiffIR in Sec. 3.4. Finally, we present our proposed dataset benchmark for compressed image restoration tasks in Sec. 3.5.

3.1 Preliminary

Mixture-of-Experts: The Mixture of Experts (MoE) model is an effective method for increasing the capabilities [39, 44, 55] of the models, and it is frequently employed in various scaling-up tasks. In MoE, routers select and activate different experts based on the input tokens using various routing mechanisms [44, 51, 83]. A particularly typical example is the use of Sparsely Gated MoE [55] where the output y of MoE layer could be described as:

$$y = \sum_{i=1}^{n} G(x)_i E_i(x) \tag{1}$$

Here, G_x and E_i denote the output of router and i-th expert, respectively. In this work, we draw inspiration from the routing concept in MoE framework to combine basic prompts, which enables the prompts to cooperate together and fully excavate the task-customized diffusion priors for universal compression tasks.

Stable Diffusion: Stable diffusion conducts diffusion process in latent space, where a VAE encoder is used to compress input image into latent variable z_0. Then the model predict added noise to noisy latent z_t with a unet network. The optimization function could be written as follows:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \left[\|\epsilon - \epsilon(z_t, t)\|_2^2 \right] \tag{2}$$

Where t denotes the time step and ϵ denotes the noisy map to be estimated.

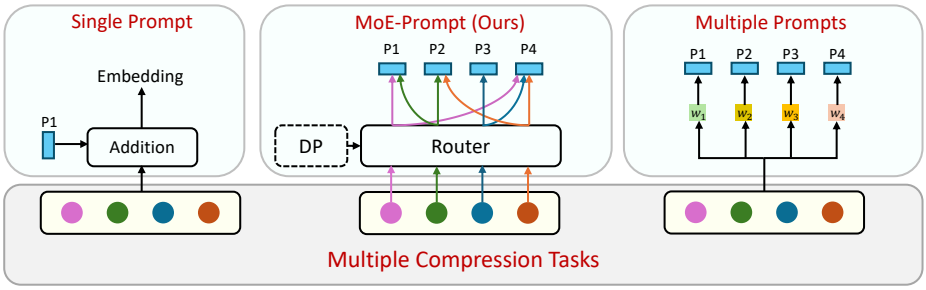


Fig. 2: Comparison of different prompt interaction methods. Here we mainly categorize them into three types: (a) Single Prompt [43], (c) Multiple Prompts [3, 35, 40], (b) MoE-Prompt (Ours). We use Mixture of Experts routing methods to select different combinations of prompts for various compression tasks. In (b), DP stands for Degradation Prior which is obtained from LQ images through pre-trained CLIP encoder of DACLIP.

3.2 Mixture-of-Experts Prompt

As depicted in Fig. 2(b), we propose the mixture-of-experts (MoE) prompt to excavate task-customized diffusion priors. Unlike previous prompt-based IR methods, MoE-Prompt is designed to maximize reusability and the representative capacity of each prompt expert for different tasks. Concretely, there are two commonly used prompt-based IR categories. The first category is the single prompt, as shown in Fig. 2(a), where a single prompt (usually image-like) is used to perceive distortions from different tasks through simple addition. This method struggles to model multiple tasks effectively, particularly as the number of tasks increases. A single prompt makes it difficult to manage complex relationships between different tasks.

The second category involves the use of multiple prompts, as represented in Fig. 2(c), in most works [3, 40, 50]. Specifically, these methods set a prompt pool and generate a set of weights: $w_1, w_2, ..., w_n$, which are used to multiply the predefined prompts and fuse them with soft weighting. However, this method is susceptible to the "mean feature", *i.e.*, these prompts learn similar features, lacking the diversity and reducing the modulation capability of universal tasks (Please see Sec. 1 in the **Supplementary**). The reason is due to the lack of one mechanism to enable these prompts to learn distinct degradation/task knowledge.

Therefore, the core principle of our MoE-Prompt method is to treat each prompt as an expert, allowing for the adaptive selection and scheduling of the necessary prompt cooperation for different distortion tasks through a router. This enables prompts to better cooperate and be reused for extracting task-customized diffusion priors. As depicted in Fig. 2(b), it is necessary to provide distortion-related information to the router. Considering that DA-CLIP [40] is trained on large-scale distortion tasks and has demonstrated robustness to out-of-domain data, we use the pre-trained CLIP encoder from DACLIP to extract

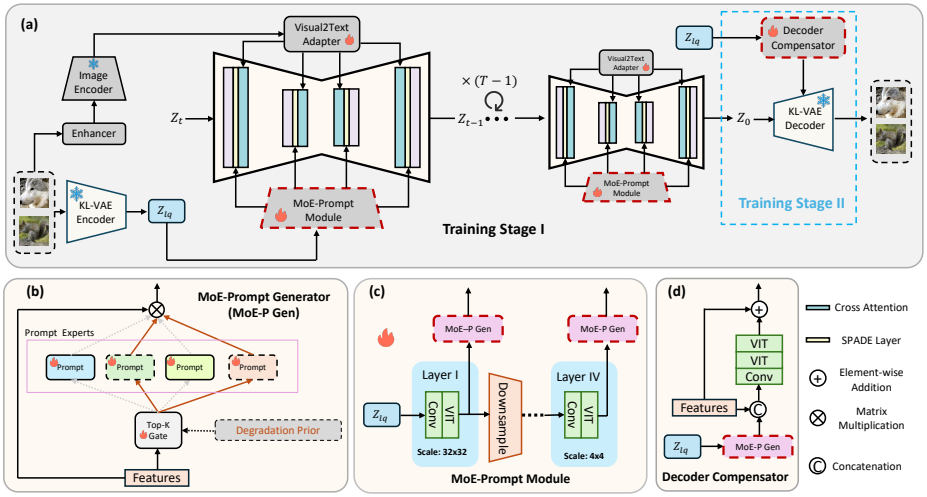


Fig. 3: The framework of the proposed MoE-DiffIR enables dynamic prompt learning for multiple CIR tasks through (b) MoE-Prompt Generator, and introduces a visual-to-text adapter to generate more reasonable texture. In MoE-DiffIR: MoE-Prompt Module (c) aims to extract multi-scale features to interact with (b). Here (a) depicts the process of fine-tuning Stable Diffusion, which consists of two stages. Stage I: only the MoE-Prompt Module is pre-trained to excavate task-customized diffusion priors for each CIR task. Stage II: the (d) Decoder Compensator is fine-tuned for structural correction.

the degradation prior (DP) from low-quality images for various compression tasks. The obtained DP interacts with input features through a cross-attention mechanism and is then fed into the router. A more detailed diagram of this structure could be found in the Sec. 1 of the **Supplementary**. After that, the router adaptively selects a combination of prompts using a noisy Top-K function [55], which is formalized as:

$$G(x) = \text{Top-K}(\text{Softmax}(xW_g + \mathcal{N}(0,1)\text{Softplus}(xW_{\text{noise}}))) \tag{3}$$

where x represents the input features, W_g is the weight matrix for global features, and W_{noise} introduces stochasticity to the selection process, encouraging robustness and diversity in prompt selection. "Softplus" here is the smooth approximation to the ReLU function. Once K prompts have been selected, they interact with the input feature through a form of matrix multiplication.

3.3 Visual2Text Adapter

Stable Diffusion, trained on large-scale datasets [54], stores an abundance of text-to-image priors. However, these priors are often overlooked by some existing SD-based IR works. For instance, StableSR [61] and DiffBIR [38] configure the text condition input for the SD as an empty string. In order to activate

and reutilize textual prior knowledge, we attempt to extract visual information from low-quality(LQ) images and transform it into the text embedding space. Indeed, there are some attempts to leverage pre-trained language models like BLIP for direct textual feature extraction from LQ images, such as PASD [74]. However, in the realm of compressed image restoration (CIR), especially at very low bit rates, the damage to the distorted images is severe. Extracting textual features from these images could potentially degrade the performance of the model. Therefore, as shown in Fig. 3(a), we first enhance the LQ images using several transformer blocks and then employ CLIP's image encoder to directly extract visual features. To better leverage the robust text-to-image capabilities of SD, we employ several MLP layers [16] (referred to the Visual2Text Adapter) to translate visual information into the textual domain of SD. This approach aids in enhancing the reconstruction of textures and details.

3.4 Overall Fine-tuning Procedure

Fig. 3(a) illustrates the entire fine-tuning process of our MoE-DiffIR. Similar to StableSR [61] and AutoDIR [23], we fine-tune the framework in two stages. In the first stage, only the MoE-Prompt Module is trained, while the VAE Codecs and UNet remain fixed. The MoE-Prompt Module modulates the LQ image features onto the multi-scale outputs of Stable Diffusion via the SPADE layer [64]. To achieve this, we employ three downsample layers in the MoE-Prompt Module, and use ViT blocks [12] and convolution layers to extract LQ features at each scale.

In the second stage, all modules are fixed except for the VAE decoder. This fine-tuning process is crucial for ensuring the fidelity of the recovered images, which is also underscored in existing literature [61, 85]. The high compression rate may lead to information loss during the reconstruction stage via the VAE decoder. This occurs because the pre-trained VAE decoder does not align well with varying scenarios, causing the output latent variable z_0 from Stable Diffusion to misalign with the our CIR tasks. Consequently, it is essential to augment the Decoder with some low-quality information, as clearly illustrated in Fig. 3(d). The loss function for second stage fine-tuning is:

$$L_{Decoder} = \mathcal{L}_{lpips}[z_{lq}, z_0, hr] \tag{4}$$

In this phase, we employ the LPIPS perceptual loss function, using high-quality images as the reference. Here z_0 denotes the output of unet denoising network and z_{lq} is latent variable of low quality image.

3.5 CIR dataset benchmark

We introduce the first universal dataset benchmark for compressed image restoration. This benchmark includes four traditional compression methods: (i) JPEG [59], (ii) VVC [5], (iii) HEVC [57], (iv) WebP [17] and three learning-based compression methods: (i) C_{PSNR}, (ii) C_{SSIM}, (iii) HIFIC [45]. Both C_{PSNR} and C_{SSIM}

are adopted from the work [7], optimized by MSE and MS-SSIM loss, respectively. Each codec has three distinct bitrate levels. For JPEG and WebP, we set quality factor (QF) values from [5,10,15]. For VVC and HEVC, we adopt MPEG standard QP values from [37, 42, 47]. For HIFIC, we use three released weights [1] represented for three different bitrates: "low", "med", and "high". We also define cross-degree distortions for unseen test tasks, such as setting QF of JPEG from values [5,25]. Additionally, we create cross-type distortions using AVC codec methods for static images from values [37, 42, 47]. We adopt DF2K [2,37] as our compressed training dataset, containing 3450 images, resulting in 72450 images across 21 compression tasks.

4 Experiments

4.1 Experiment Setup

Implementation Details. We fine-tune Stable Diffusion 2.1-base[2] over two training stages. In the first stage, followed in Sec. 3.4, we fix the decoder of VAE and only train MoE-Prompt module. We use an Adam optimizer($\beta_1 = 0.9$, $\beta_2 = 0.999$) with fixed learning rate of $5e^{-5}$. The total iterations are 0.4M steps, constrained by loss function $\mathcal{L}_{SD}$ as described in the Sec. 3.1. In the second stage, we train only the decoder compensator with other modules fixed. We generate 70,000 latent images using the weights from the first stage and train the decoder with the corresponding LQ images and ground truth images. The learning rate is set to $1e^{-4}$ and total iterations are 0.1M steps. In the whole training process, we resize the input images into 256x256 and employ random flipping and rotation for data augmentation. The batch size is set to 32, and the training is conducted on four NVIDIA RTX 3090 GPUs.

Compared Methods. To validate the effectiveness of MoE-DiffIR, we compare it with several state-of-the-art (SOTA) methods. These methods include two for all-in-one IR: PromptIR [50] and Airnet [29], one method for compression artifact removal: HAT [6], one GAN-based method: RealESRGAN [63], and four diffusion-based methods: StableSR [61], DiffBIR [38], SUPIR [77] and PASD [74]. Here, we present only a subset of the quantitative results. A more comprehensive set of quantitative results will be detailed in the Sec. 2 of the **Supplementary**. For training settings, we adhere to the configurations provided in the official code repositories of these methods. We set batch size to 32 for all methods.

4.2 Comparisons with State-of-the-arts

We validate MoE-DiffIR on five commonly used compressed test sets: LIVE1 [56], Classic5 [79], BSDS500 [4], DIV2K Testset [2], and ICB [14]. We employ PSNR, SSIM as distortions metrics, and LPIPS [81], FID [20] as perceptual metrics. In Sec. 2 of the **Supplementary**, we show more results using some non-reference metrics like ClipIQA [60], ManIQA [73] to further validate the perceptual quality.

[1] `https://github.com/Justin-Tan/high-fidelity-generative-compression`
[2] `https://huggingface.co/stabilityai/stable-diffusion-2-1-base`

Table 1: Quantitative comparison for compressed image restoration on seven codecs (average on three distortions). Results are tested on with different compression qualities in terms of PSNR↑, SSIM↑, LPIPS↓, FID↓. Red and blue colors represent the best and second best performance, respectively. **All compared methods are retrained on our constructed CIR datasets except for SUPIR [77].**

Codecs	Methods	LIVE1 [56]				Classic5 [79]				BSDS500 [4]				DIV2K [2]				ICB [14]			
		PSNR	SSIM	LPIPS	FID	PSNR	SSIM	LPIPS	FID	PSNR	SSIM	LPIPS	FID	PSNR	SSIM	LPIPS	FID	PSNR	SSIM	LPIPS	FID
JPEG [59]	AirNet [29]	30.06	0.858	0.2241	113.59	35.84	0.955	0.1213	62.78	30.99	0.864	0.1975	96.56	29.91	0.872	0.2037	107.22	31.25	0.878	0.2565	179.67
	HAT [6]	30.18	0.860	0.1952	115.53	33.69	0.911	0.1185	57.80	31.07	0.865	0.1809	86.32	29.58	0.869	0.1980	107.22	30.98	0.881	0.2308	173.19
	PromptIR [50]	31.42	0.885	0.2131	111.15	35.66	0.949	0.1124	49.84	31.95	0.881	0.1791	87.50	31.14	0.896	0.1879	98.83	31.67	0.896	0.2113	167.68
	RealESRGAN [63]	30.26	0.860	0.1423	76.28	33.21	0.915	0.1106	56.49	30.24	0.848	0.1485	81.43	29.39	0.866	0.1357	74.46	30.22	0.876	0.1443	130.66
	DiffBIR [38]	28.42	0.812	0.0995	67.23	30.38	0.86	0.1026	56.95	28.67	0.810	0.1013	73.43	27.99	0.805	0.0898	65.84	29.72	0.837	0.1086	113.05
	PASD [74]	28.39	0.806	0.1055	70.46	29.97	0.839	0.0950	54.98	27.70	0.758	0.0934	70.93	28.04	0.784	0.0906	67.61	28.75	0.668	0.1736	115.67
	StableSR [61]	30.19	0.855	0.1069	68.19	31.08	0.875	0.1227	46.97	30.60	0.847	0.1127	67.94	29.44	0.863	0.1067	62.55	31.00	0.875	0.1213	125.56
	SUPIR [77]	27.41	0.747	0.1254	71.04	29.05	0.834	0.1110	60.60	27.89	0.713	0.1076	69.63	27.56	0.780	0.1362	70.23	28.78	0.673	0.1319	121.60
	Ours	30.50	0.857	0.0964	62.72	32.30	0.890	0.0902	40.83	30.95	0.852	0.1006	68.00	29.86	0.868	0.0906	55.75	31.48	0.881	0.1059	100.93
VVC [5]	AirNet [29]	29.08	0.803	0.3055	185.2	33.74	0.914	0.1715	165.37	29.71	0.796	0.3147	183.47	28.15	0.808	0.2863	155.61	28.17	0.797	0.3252	208.07
	HAT [6]	29.19	0.806	0.2801	182.68	32.27	0.874	0.1629	156.23	29.75	0.797	0.2874	170.61	28.51	0.812	0.2832	154.78	29.04	0.817	0.3069	213.37
	PromptIR [50]	30.00	0.825	0.3053	168.36	34.06	0.916	0.1662	148.57	30.32	0.811	0.2818	162.25	29.41	0.832	0.2628	145.96	29.35	0.816	0.2798	205.30
	RealESRGAN [63]	28.64	0.786	0.2082	131.83	32.56	0.884	0.1196	87.79	28.24	0.765	0.2185	130.61	27.77	0.792	0.1978	114.20	27.89	0.790	0.2118	188.82
	DiffBIR [38]	27.54	0.771	0.1687	95.49	29.64	0.811	0.1407	85.55	27.97	0.773	0.1775	111.42	27.06	0.763	0.1468	93.68	28.03	0.775	0.1746	154.91
	PASD [74]	27.48	0.766	0.1746	101.67	28.78	0.779	0.1381	86.43	26.84	0.749	0.1717	109.04	26.91	0.738	0.1355	96.20	26.80	0.708	0.1857	154.37
	StableSR [61]	28.49	0.771	0.1679	98.45	30.23	0.822	0.1318	84.59	28.75	0.756	0.1831	100.38	27.85	0.789	0.1473	92.08	28.46	0.789	0.1824	165.84
	SUPIR [77]	27.46	0.714	0.1468	99.73	28.45	0.774	0.1415	87.62	27.16	0.784	0.1759	105.71	26.49	0.733	0.1833	99.80	26.84	0.715	0.1990	159.98
	Ours	28.76	0.777	0.1444	88.83	31.01	0.845	0.1121	78.28	28.94	0.755	0.1577	84.92	28.05	0.786	0.1316	80.96	28.63	0.781	0.1540	144.51
HEVC [57]	AirNet [29]	28.70	0.792	0.3159	167.81	33.59	0.906	0.1790	172.80	29.31	0.784	0.3244	174.49	27.78	0.794	0.3010	158.35	27.35	0.827	0.3041	209.11
	HAT [6]	28.82	0.795	0.2957	167.72	31.80	0.862	0.1717	165.78	29.42	0.786	0.3002	176.55	28.19	0.800	0.2994	153.22	29.45	0.831	0.3075	232.96
	PromptIR [50]	29.54	0.814	0.3138	172.58	33.87	0.909	0.1721	157.71	29.88	0.799	0.2901	156.25	29.04	0.822	0.2677	146.35	29.85	0.834	0.2847	223.55
	RealESRGAN [63]	28.31	0.776	0.2269	139.67	32.43	0.873	0.1266	104.17	28.02	0.754	0.2355	137.33	27.53	0.781	0.2148	117.01	28.33	0.807	0.2192	200.91
	DiffBIR [38]	27.53	0.762	0.1790	104.09	29.54	0.809	0.1317	94.85	27.79	0.773	0.1784	111.20	26.96	0.751	0.1595	96.19	28.35	0.780	0.1666	176.30
	PASD [74]	27.50	0.76	0.1801	108.89	29.02	0.78	0.1300	91.91	26.07	0.747	0.1853	106.77	26.15	0.704	0.1461	100.83	27.02	0.706	0.1801	177.79
	StableSR [61]	28.19	0.759	0.1845	106.00	30.39	0.821	0.1323	85.98	28.64	0.751	0.1902	108.52	27.67	0.780	0.1656	95.47	28.84	0.803	0.1895	184.51
	SUPIR [77]	26.80	0.679	0.1605	108.69	28.39	0.777	0.1339	97.12	26.40	0.743	0.1895	107.19	26.47	0.700	0.1721	103.54	27.05	0.710	0.1938	182.37
	Ours	28.50	0.768	0.1622	98.66	31.08	0.839	0.1157	82.82	28.71	0.745	0.1749	96.23	27.83	0.777	0.1470	84.63	28.87	0.791	0.1776	171.01
WebP [17]	AirNet [29]	29.40	0.822	0.2537	154.54	34.59	0.930	0.1453	106.65	29.55	0.803	0.2701	161.94	28.31	0.831	0.2253	145.74	27.02	0.831	0.2804	221.70
	HAT [6]	29.51	0.825	0.2384	153.36	32.31	0.882	0.1404	99.30	29.68	0.804	0.2582	152.92	28.84	0.832	0.2252	139.85	30.33	0.857	0.2358	211.06
	PromptIR [50]	30.49	0.856	0.2501	149.40	35.1	0.936	0.1383	94.66	30.35	0.832	0.2533	145.45	30.04	0.866	0.2137	135.30	29.61	0.866	0.2352	207.33
	RealESRGAN [63]	28.95	0.816	0.1697	100.49	33.66	0.913	0.0928	89.85	28.39	0.784	0.1890	117.83	28.16	0.824	0.1568	94.41	29.23	0.844	0.1735	173.90
	DiffBIR [38]	27.96	0.779	0.1472	91.40	30.01	0.875	0.1021	66.42	28.30	0.777	0.1705	103.80	27.73	0.778	0.1318	91.27	29.05	0.807	0.1407	136.89
	PASD [74]	27.88	0.77	0.1511	96.90	29.71	0.844	0.0940	63.11	27.68	0.810	0.1638	101.99	26.22	0.723	0.1218	97.58	27.70	0.707	0.1379	138.46
	StableSR [61]	28.88	0.805	0.1206	75.13	32.11	0.872	0.1074	69.10	29.00	0.780	0.1528	97.23	28.34	0.820	0.1235	84.54	29.86	0.839	0.1487	142.77
	SUPIR [77]	27.36	0.671	0.1418	95.55	30.14	0.837	0.1073	69.50	26.17	0.757	0.1791	100.12	26.84	0.716	0.1530	101.02	27.72	0.711	0.1389	141.48
	Ours	29.28	0.815	0.1098	70.64	32.82	0.895	0.0781	61.70	29.13	0.783	0.1300	80.32	28.57	0.825	0.1019	68.87	30.27	0.845	0.1255	136.27
C_{PSNR} [7]	AirNet [29]	30.33	0.847	0.2386	151.30	35.15	0.940	0.1502	165.52	30.35	0.825	0.2678	156.72	29.92	0.859	0.2110	145.42	28.83	0.863	0.2421	188.50
	HAT [6]	30.40	0.850	0.2070	153.25	33.33	0.898	0.1520	167.65	30.43	0.827	0.2460	155.76	29.79	0.861	0.2095	135.84	31.38	0.880	0.2384	178.90
	PromptIR [50]	30.84	0.865	0.2341	141.38	35.87	0.945	0.1461	153.46	30.91	0.838	0.2478	145.03	30.54	0.879	0.1857	129.72	31.49	0.880	0.2067	175.59
	RealESRGAN [63]	30.07	0.840	0.1467	85.38	33.76	0.917	0.1096	80.31	29.62	0.813	0.2235	142.41	29.12	0.845	0.1443	81.47	29.79	0.859	0.1580	151.03
	DiffBIR [38]	28.42	0.812	0.1064	85.38	30.13	0.841	0.1307	79.13	28.06	0.795	0.1534	103.09	28.06	0.804	0.1030	80.08	29.75	0.835	0.1096	109.86
	PASD [74]	28.34	0.81	0.1101	89.69	29.24	0.810	0.1269	66.64	26.78	0.782	0.1549	99.44	27.43	0.786	0.0933	82.53	28.56	0.785	0.1802	117.99
	StableSR [61]	29.85	0.833	0.1082	73.86	31.29	0.847	0.1199	70.77	29.64	0.806	0.1595	103.12	28.76	0.835	0.1003	68.97	30.40	0.853	0.1280	114.55
	SUPIR [77]	27.48	0.721	0.1074	90.06	29.71	0.803	0.1410	71.20	27.95	0.803	0.1719	108.30	27.00	0.779	0.1046	84.73	28.58	0.796	0.1252	120.47
	Ours	30.18	0.837	0.0996	72.23	31.75	0.866	0.1029	64.18	30.12	0.821	0.1623	101.35	29.46	0.848	0.0865	58.97	31.25	0.864	0.1041	109.05
C_{SSIM} [7]	AirNet [29]	27.54	0.816	0.3325	171.53	33.99	0.947	0.1490	131.64	28.32	0.806	0.3270	173.52	27.09	0.822	0.3183	165.87	26.64	0.824	0.3268	197.18
	HAT [6]	27.63	0.819	0.2780	167.27	32.79	0.911	0.1428	120.24	28.35	0.809	0.2864	161.61	26.88	0.813	0.3180	161.40	28.88	0.854	0.2651	193.93
	PromptIR [50]	27.95	0.826	0.3284	160.14	34.26	0.948	0.138	123.52	28.49	0.817	0.2939	158.59	27.23	0.822	0.2904	156.75	28.97	0.843	0.2543	185.23
	RealESRGAN [63]	27.08	0.802	0.1986	108.08	32.72	0.92	0.1009	61.25	27.47	0.791	0.2765	150.52	26.20	0.786	0.2131	116.11	27.38	0.815	0.1996	169.79
	DiffBIR [38]	26.50	0.766	0.1364	82.54	29.62	0.841	0.1354	69.74	26.67	0.757	0.1517	95.84	26.17	0.748	0.1351	89.88	28.44	0.806	0.1352	128.31
	PASD [74]	26.45	0.762	0.1377	86.59	28.64	0.819	0.1273	68.14	26.72	0.744	0.1483	95.09	25.56	0.725	0.1372	87.22	27.48	0.706	0.1758	127.86
	StableSR [61]	27.00	0.789	0.1535	90.74	30.53	0.844	0.1271	71.17	27.81	0.785	0.1897	109.78	26.31	0.786	0.1506	96.62	28.35	0.824	0.1509	129.93
	SUPIR [77]	25.78	0.661	0.1358	87.00	28.57	0.813	0.1414	70.98	25.54	0.695	0.1587	92.55	25.00	0.719	0.1277	90.01	27.50	0.709	0.1842	133.68
	Ours	26.85	0.781	0.1352	80.23	31.03	0.87	0.1154	66.02	28.09	0.798	0.1573	101.87	25.97	0.773	0.1294	84.58	28.20	0.819	0.1322	122.25
HIFIC [45]	AirNet [29]	29.22	0.853	0.1258	78.78	32.35	0.897	0.1061	60.69	27.63	0.841	0.1339	77.67	28.93	0.864	0.1227	82.71	29.04	0.880	0.1956	135.95
	HAT [6]	29.32	0.857	0.1254	79.54	32.48	0.909	0.1009	62.95	27.70	0.841	0.1368	75.22	28.10	0.857	0.1220	76.08	30.35	0.888	0.1723	126.98
	PromptIR [50]	29.89	0.876	0.1206	82.81	32.69	0.918	0.0958	64.17	28.12	0.858	0.1325	79.23	29.06	0.879	0.1153	79.58	30.14	0.895	0.1577	122.82
	RealESRGAN [63]	28.86	0.849	0.0766	53.47	30.86	0.918	0.084	47.83	26.76	0.818	0.1040	70.85	27.62	0.839	0.0982	66.55	29.40	0.875	0.1014	104.11
	DiffBIR [38]	28.03	0.805	0.0595	45.14	28.72	0.849	0.0777	41.64	26.46	0.795	0.0920	69.37	27.13	0.796	0.0781	59.07	29.15	0.838	0.0776	79.42
	PASD [74]	28.00	0.801	0.0652	50.70	28.40	0.818	0.0774	40.57	27.52	0.787	0.0846	65.70	27.59	0.734	0.0690	58.63	28.90	0.762	0.1297	78.86
	StableSR [61]	28.87	0.841	0.0706	46.69	29.78	0.877	0.0792	39.27	26.96	0.807	0.0905	62.49	27.49	0.837	0.0705	53.48	29.46	0.867	0.0876	85.59
	SUPIR [77]	27.81	0.748	0.0735	48.40	28.43	0.812	0.0799	44.23	27.37	0.756	0.1021	63.67	27.17	0.733	0.0834	62.03	28.93	0.769	0.0936	83.47
	Ours	29.18	0.847	0.0593	43.58	30.24	0.88	0.0767	35.89	27.54	0.824	0.0753	53.70	28.08	0.848	0.0628	48.57	30.15	0.875	0.0685	74.12

Quantitative Analysis. Table 1 shows comprehensive performance of our MoE-DiffIR compared with SOTA methods across 7 compression codecs. Here, for each codec, we average the metrics of its three distortion levels. Since the primary objective of this work is to enhance the perceptual quality of images at low bitrates, our comparisons primarily focus on the perceptual quality against generative models. From the table, we can see that our method almost surpasses all other methods in terms of perceptual metrics like LPIPS and FID. Moreover, our method is also competitive in distortion metrics such as PSNR compared to transformer-based methods, thanks to the fine-tuning stage of the VAE decoder. Specifically, on the LPIPS metric, we achieved a 10.9% reduction compared to

SUPIR, a decrease of 5.4 on the FID metric, and also an average increase of 0.41dB over StableSR on the PSNR metric.

Qualitative Analysis. Additionally, we also present some perceptual visual results in Fig. 4, covering different quality factors from various codecs (More visual comparsions with SUPIR and PASD are shown in Sec. 3 of the **Supplementary**). It is observable that, in scenarios with lower compression rates, the Transformer-based all-in-one model PromptIR tends to restore images too smoothly, whereas DiffBIR is prone to generating some erroneous texture details, as shown in the textual information in the third row of Fig. 4. Thanks to the compensation in the second stage of the VAE decoder, MoE-DiffIR is capable of generating more accurate textures in terms of fidelity. Moreover, our MoE-Prompt enables MoE-DiffIR to effectively handle different compression distortions, demonstrating excellent perceptual restoration capabilities, including color correction and texture detail generation.

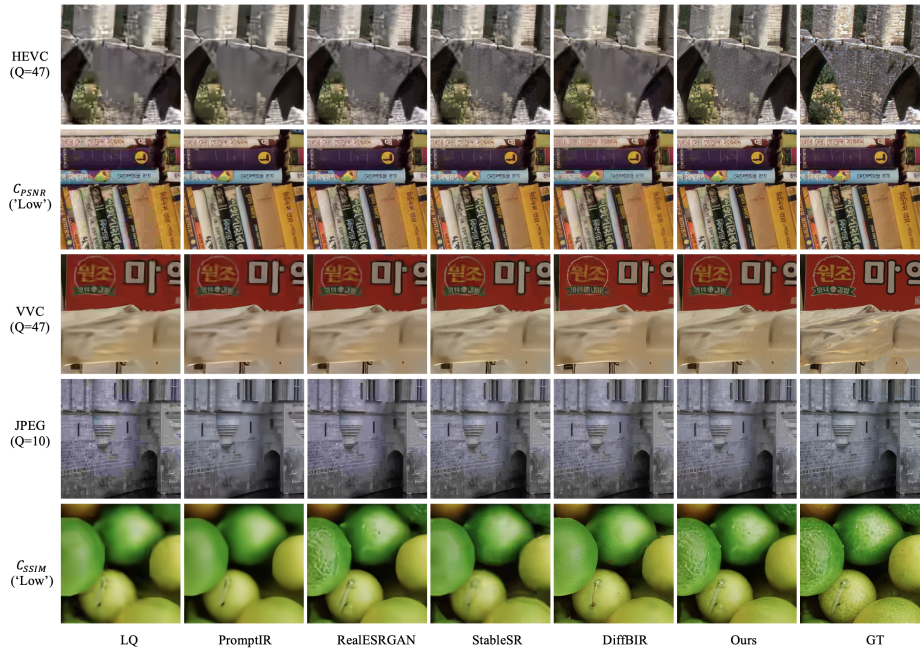


Fig. 4: Visual comparisons between our methods and other state of the arts methods. This figure demonstrate 5 different compression tasks: JPEG (QF=10), VVC (QP=47), HEVC (QP=47), C_{SSIM}("Low" bitrates), C_{PSNR}("Low" bitrates). More visual results can be found in Sec. 3 of the **Supplementary**.

4.3 Ablation Studies

The effects of MoE-Prompt. We conduct experiments with different prompt designs mentioned in Sec. 3. The results are presented in Table 2. We compare

four prompt designs: No Prompt, Single Prompt, Multiple Prompts, and our MoE-Prompt. In addition to the 21 tasks in our CIR dataset, we also test the performance of these models on some unseen tasks to assess their generalizability. Specifically, we employ two types of unseen tasks to validate the generalization performance. The first type is "Cross Degrees", which involves selecting one of the seven codecs, but with unseen quality factors. In this experiment, we choose VVC with QP from values [32,52] as the distortion types. The second type is "Cross Type", where we select codec AVC [67] with QP from values [47, 42, 37]. Specifically, without prompts, the model has a reduced ability to distinguish between various distortions, leading to a notably lower average performance across tasks than prompt-based models, particularly on unseen tasks. Furthermore, using a single prompt results in lower average performance across the 21 tasks compared to using multiple prompts or MoE-Prompt, indicating that a single prompt lacks the capability for multi-task modeling. In contrast, our MoE-based method exceeds multiple prompt design by an average of 0.05dB on 21 tasks and improves perceptual quality, reducing LPIPS by 5% and FID by about 4. This proves that MoE-Prompt can more effectively utilize and share prompts across various distortion tasks, uncovering task-customized diffusion priors than other prompt interaction methods.

Table 2: Impacts of different prompt designs. Results are reported on LIVE1, DIV2K and ICB. For seen tasks, the value is the average result of 21 compression tasks on our CIR dataset. For unseen tasks, the value is the average result of "Cross Degrees" (VVC: QP form values [32,52]) and "Cross Types" (AVC: QP from values [47,42,37]). Best performances are in **bold**.

Methods	Seen tasks						Unseen Tasks			
	LIVE1		BSDS500		DIV2K		LIVE1 (Cross Degrees)		ICB (Cross Types)	
	PSNR/SSIM	LPIPS/FID	PSNR/SSIM	LPIPS/FID	PSNR/SSIM	LPIPS/FID	PSNR/SSIM	LPIPS/FID	PSNR/SSIM	LPIPS/FID
No Prompt	28.73/0.806	0.1343/85.87	28.56/0.770	0.1591/96.45	27.86/0.813	0.1295/79.5	31.79/0.893	0.064/37.33	28.61/0.787	0.1933/210.84
Single Prompt	28.86/0.806	0.1272/79.45	28.78/0.791	0.1530/89.62	28.02/0.816	0.1143/71.26	33.25/0.910	0.0457/28.60	28.88/0.793	0.179/187.29
Multiple Prompt	28.98/0.810	0.1212/77.09	28.93/0.794	0.1482/89.34	28.22/0.817	0.1124/71.65	33.32/0.913	0.0432/28.23	28.89/0.792	0.1756/187.89
MoE-Prompt (Ours)	**29.02/0.811**	**0.1179/75.86**	**28.97/0.794**	**0.1430/88.14**	**28.29/0.821**	**0.1071/68.91**	**33.45/0.916**	**0.0411/25.65**	**29.02/0.800**	**0.1690/176.87**

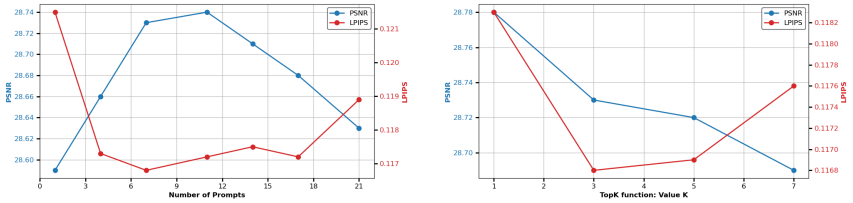
Table 3: Impacts of Visual2Text(V2T) adapter and Degradation Prior (DP). Results are reported on LIVE1, BSDS500, ICB. Here the value is the average result of 21 compression tasks. Best performances are in **bold**.

Methods	Datasets					
	LIVE1		BSDS500		ICB	
	PSNR/SSIM	LPIPS/FID	PSNR/SSIM	LPIPS/FID	PSNR/SSIM	LPIPS/FID
MoE-Prompt	29.02/0.810	0.1179/75.86	28.97/0.794	0.1430/88.14	29.83/0.839	0.1277/122.59
MoE-Prompt+V2T Adapter	29.03/0.812	0.1145/74.13	28.94/0.796	0.1367/**86.77**	29.83/0.840	0.1239/119.78
MoE-Prompt+DP	29.07/0.814	0.1154/76.60	**29.06**/0.795	0.1405/88.00	29.87/0.841	0.1269/122.32
MoE-Prompt+V2T Adapter+DP	**29.10/0.814**	**0.1136/73.60**	29.02/**0.797**	**0.1356**/86.81	**29.88/0.841**	**0.1235/119.29**

The effects of Visual2Text adapter and Degradation Prior. In Sec. 3.3, we describe using a cross-modal adapter to convert visual information into text embeddings. Additionally, in Sec. 3.2, we employ the pre-trained DACLIP [40] to provide degradation priors (DP), enhancing the router's adaptive selection of optimal prompts. Ablation studies validate these methodologies by integrating

the Visual2Text adapter or DP into the MoE-Prompt backbone. Table 3 shows that adding V2T adapter could reduce LPIPS by 3-5% and improves FID by 1-3 points on average, indicating better perceptual quality. The use of degradation prior (DP) mainly contributes to distortion metrics, with an average PSNR increase of 0.07dB across 21 tasks, indicating that adding visual information could enhance perceptual quality while adding DP may improve fidelity. Visual comparisons in Fig. 5 also show an interesting phenomenon: at extremely low bitrates, Stable Diffusion may convert severe distortions into noise spots, which could be smoothed with the use of the V2T adapter or by adding degradation prior (DP), thereby enhancing model performance.

Fig. 5: Visual ablation results: different prompt interaction designs, use of V2T adapter and use of degradation prior (DP).



(a) Effects of the number of prompts **(b)** Effects of K value

Fig. 6: The effect of the number of prompts and the value K of the Top-K function.

The effects of number of total prompts and selected prompts. In our proposed MoE-Prompt Module, the router uses a Top-K function to select K prompts from N predefined prompts. We conduct ablation experiments to evaluate the effects of varying N and K. Since our CIR dataset consits of total 21 compression tasks, we set N to a series of values (1, 4, 7, 11, 14, 17, 21) and use

the LIVE1 dataset for testing. As shown in Fig. 6(a), changing N significantly impacts both the distortion metric PSNR and the perceptual metric LPIPS. When N is small, both PSNR and LPIPS are poor because a single prompt (N=1) cannot model different distortion tasks effectively. As N increases, performance improves but then declines past a certain point due to the difficulty in learning task-relevant distortion features from too many prompts, leading to underutilization. The data suggests that performance is optimal at N=7, similar to N=11, indicating that around N=7 is sufficient for parameter economy. We then fix N at 7 and vary K with values (1, 3, 5, 7). Fig. 6(b) shows that while K=1 yields higher PSNR, it results in poor perceptual quality. We can conclude that multiple prompts could cooperate together for perceiving different tasks with better perceptual quality. The best perceptual performance is seen when K is between 3 and 5. Thus, we select N=7 and K=3 for the final settings.

5 Conclusion

In this work, we propose the MoE-Prompt to excavate task-customized diffusion priors for universal compressed image restoration, dubbed MoE-DiffIR. Our method maximizes the utilization of different prompts, enabling them to collaboratively perceive different distortions. By utilizing a Visual2Text adapter, we integrate visual information into the text inputs of the Stable Diffusion model, thereby improving the perceptual restoration capabilities of the model at low bitrates. We also construct a comprehensive dataset benchmark for CIR tasks. Our extensive experiments have demonstrated that MoE-DiffIR not only improves perceptual performance at low bitrates but also facilitates rapid transferability across various compression tasks. In the future, we intend to design novel approaches within our CIR benchmark to further improve the performance of the model.

Limitation

In this work, we propose a novel universal compressed image restoration (CIR) framework using the MoE-Prompt Modules. Although our model outperforms other methods in terms of perceptual quality, there remains a notable gap between the restored images and the ground truth at extremely low bitrates, as shown in Fig. 4. In future work, we aim to focus on this area and strive for improvements.

Acknowledgements

This work was supported in part by NSFC under Grant 623B2098, 62371434 and 62021001.

References

1. Agustsson, E., Minnen, D., Toderici, G., Mentzer, F.: Multi-realism image compression with a conditional generator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22324–22333 (2023)
2. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
3. Ai, Y., Huang, H., Zhou, X., Wang, J., He, R.: Multimodal prompt perceiver: Empower adaptiveness, generalizability and fidelity for all-in-one image restoration. arXiv preprint arXiv:2312.02918 (2023)
4. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE transactions on pattern analysis and machine intelligence **33**(5), 898–916 (2010)
5. Bross, B., Wang, Y.K., Ye, Y., Liu, S., Chen, J., Sullivan, G.J., Ohm, J.R.: Overview of the versatile video coding (vvc) standard and its applications. IEEE Transactions on Circuits and Systems for Video Technology **31**(10), 3736–3764 (2021)
6. Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., Dong, C.: Hat: Hybrid attention transformer for image restoration. arXiv preprint arXiv:2309.05239 (2023)
7. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7939–7948 (2020)
8. Christopoulos, C., Skodras, A., Ebrahimi, T.: The jpeg2000 still image coding system: an overview. IEEE transactions on consumer electronics **46**(4), 1103–1127 (2000)
9. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687 (2022)
10. Chung, H., Ye, J.C., Milanfar, P., Delbracio, M.: Prompt-tuning latent diffusion models for inverse problems. arXiv preprint arXiv:2310.01110 (2023)
11. Dong, C., Deng, Y., Loy, C.C., Tang, X.: Compression artifacts reduction by a deep convolutional network. In: Proceedings of the IEEE international conference on computer vision. pp. 576–584 (2015)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Ehrlich, M., Davis, L., Lim, S.N., Shrivastava, A.: Quantization guided jpeg artifact correction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 293–309. Springer (2020)
14. Ehrlich, M., Davis, L., Lim, S.N., Shrivastava, A.: Quantization guided jpeg artifact correction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. pp. 293–309. Springer (2020)
15. Foi, A., Katkovnik, V., Egiazarian, K.: Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. IEEE transactions on image processing **16**(5), 1395–1411 (2007)
16. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)

17. Ginesu, G., Pintus, M., Giusto, D.D.: Objective assessment of the webp image coding algorithm. Signal processing: image communication **27**(8), 867–874 (2012)
18. Gou, Y., Zhao, H., Li, B., Xiao, X., Peng, X.: Exploiting diffusion priors for all-in-one image restoration. arXiv preprint arXiv:2312.02197 (2023)
19. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5718–5727 (2022)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
21. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
22. Jiang, J., Zhang, K., Timofte, R.: Towards flexible blind jpeg artifacts removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4997–5006 (2021)
23. Jiang, Y., Zhang, Z., Xue, T., Gu, J.: Autodir: Automatic all-in-one image restoration with latent diffusion. arXiv preprint arXiv:2310.10123 (2023)
24. Jin, Y., Ye, W., Yang, W., Yuan, Y., Tan, R.T.: Des3: Adaptive attention-driven self and soft shadow removal using vit similarity. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2634–2642 (2024)
25. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. Advances in Neural Information Processing Systems **35**, 23593–23606 (2022)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
27. Li, B., Li, X., Lu, Y., Feng, R., Guo, M., Zhao, S., Zhang, L., Chen, Z.: Promptcir: Blind compressed image restoration with prompt learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2024)
28. Li, B., Li, X., Zhu, H., Jin, Y., Feng, R., Zhang, Z., Chen, Z.: Sed: Semantic-aware discriminator for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25784–25795 (2024)
29. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17452–17462 (2022)
30. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
31. Li, X., Jin, X., Lin, J., Liu, S., Wu, Y., Yu, T., Zhou, W., Chen, Z.: Learning disentangled feature representation for hybrid-distorted image restoration. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16. pp. 313–329. Springer (2020)
32. Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X., Chen, Z.: Diffusion models for image restoration and enhancement–a comprehensive survey. arXiv preprint arXiv:2308.09388 (2023)
33. Li, X., Shi, J., Chen, Z.: Task-driven semantic coding via reinforcement learning. IEEE Transactions on Image Processing **30**, 6307–6320 (2021)

34. Li, X., Sun, S., Zhang, Z., Chen, Z.: Multi-scale grouped dense network for vvc intra coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 158–159 (2020)
35. Li, Z., Lei, Y., Ma, C., Zhang, J., Shan, H.: Prompt-in-prompt learning for universal image restoration. arXiv preprint arXiv:2312.05038 (2023)
36. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
37. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
38. Lin, X., He, J., Chen, Z., Lyu, Z., Fei, B., Dai, B., Ouyang, W., Qiao, Y., Dong, C.: Diffbir: Towards blind image restoration with generative diffusion prior. arXiv preprint arXiv:2308.15070 (2023)
39. Luo, F., Xiang, J., Zhang, J., Han, X., Yang, W.: Image super-resolution via latent diffusion: A sampling-space mixture of experts and frequency-augmented decoder approach. arXiv preprint arXiv:2310.12004 (2023)
40. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. arXiv preprint arXiv:2310.01018 (2023)
41. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. arXiv preprint arXiv:2301.11699 (2023)
42. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1680–1691 (2023)
43. Ma, J., Cheng, T., Wang, G., Zhang, Q., Wang, X., Zhang, L.: Prores: Exploring degradation-aware visual prompt for universal image restoration. arXiv preprint arXiv:2306.13653 (2023)
44. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review **42**, 275–293 (2014)
45. Mentzer, F., Toderici, G.D., Tschannen, M., Agustsson, E.: High-fidelity generative image compression. Advances in Neural Information Processing Systems **33**, 11913–11924 (2020)
46. Moser, B.B., Shanbhag, A.S., Raue, F., Frolov, S., Palacio, S., Dengel, A.: Diffusion models, image super-resolution and everything: A survey. arXiv preprint arXiv:2401.00736 (2024)
47. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
48. Nie, X., Ni, B., Chang, J., Meng, G., Huo, C., Xiang, S., Tian, Q.: Pro-tuning: Unified prompt tuning for vision tasks. IEEE Transactions on Circuits and Systems for Video Technology (2023)
49. Nosratinia, A.: Embedded post-processing for enhancement of compressed images. In: Proceedings DCC'99 Data Compression Conference (Cat. No. PR00096). pp. 62–71. IEEE (1999)
50. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. arXiv preprint arXiv:2306.13090 (2023)
51. Puigcerver, J., Riquelme, C., Mustafa, B., Houlsby, N.: From sparse to soft mixtures of experts. arXiv preprint arXiv:2308.00951 (2023)

52. Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., Shakkottai, S.: Solving linear inverse problems provably via posterior sampling with latent diffusion models. Advances in Neural Information Processing Systems **36** (2024)
53. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
54. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
55. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
56. Sheikh, H.: Live image quality assessment database release 2. http://live. ece. utexas. edu/research/quality (2005)
57. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Transactions on circuits and systems for video technology **22**(12), 1649–1668 (2012)
58. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
59. Wallace, G.K.: The jpeg still picture compression standard. Communications of the ACM **34**(4), 30–44 (1991)
60. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2555–2563 (2023)
61. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. arXiv preprint arXiv:2305.07015 (2023)
62. Wang, X., Fu, X., Zhu, Y., Zha, Z.J.: Jpeg artifacts removal via contrastive representation learning. In: European Conference on Computer Vision. pp. 615–631. Springer (2022)
63. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1905–1914 (2021)
64. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 606–615 (2018)
65. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022)
66. Wang, Z., Zhang, Z., Zhang, X., Zheng, H., Zhou, M., Zhang, Y., Wang, Y.: Dr2: Diffusion-based robust degradation remover for blind face restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1704–1713 (2023)
67. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the h. 264/avc video coding standard. IEEE Transactions on circuits and systems for video technology **13**(7), 560–576 (2003)
68. Wu, Y., Li, X., Zhang, Z., Jin, X., Chen, Z.: Learned block-based hybrid image compression. IEEE Transactions on Circuits and Systems for Video Technology **32**(6), 3978–3990 (2021)

69. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. arXiv preprint arXiv:2303.09472 (2023)
70. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)
71. Yang, R., Timofte, R., Li, B., Li, X., Guo, M., Zhao, S., Zhang, L., Chen, Z., Zhang, D., Arora, Y., et al.: Ntire 2024 challenge on blind enhancement of compressed image: Methods and results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6524–6535 (2024)
72. Yang, R., Timofte, R., Li, X., Zhang, Q., Zhang, L., Liu, F., He, D., Li, F., Zheng, H., Yuan, W., et al.: Aim 2022 challenge on super-resolution of compressed image and video: Dataset, methods and results. In: European Conference on Computer Vision. pp. 174–202. Springer (2022)
73. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1191–1200 (2022)
74. Yang, T., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. arXiv preprint arXiv:2308.14469 (2023)
75. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
76. Yee, D., Soltaninejad, S., Hazarika, D., Mbuyi, G., Barnwal, R., Basu, A.: Medical image compression based on region of interest using better portable graphics (bpg). In: 2017 IEEE international conference on systems, man, and cybernetics (SMC). pp. 216–221. IEEE (2017)
77. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. arXiv preprint arXiv:2401.13627 (2024)
78. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
79. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7. pp. 711–730. Springer (2012)
80. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
81. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
82. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Unicontrolnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems **36** (2024)
83. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)

84. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15659–15669 (2023)
85. Zhu, Z., Feng, X., Chen, D., Bao, J., Wang, L., Chen, Y., Yuan, L., Hua, G.: Designing a better asymmetric vqgan for stablediffusion. arXiv preprint arXiv:2306.04632 (2023)