Prompt-Guided Alignment with Information Bottleneck Makes Image Compression Also a Restorer

Xuelin Shen[2]* **Quan Liu**[2,3]* **Jiayin Xu**[2,3] **Wenhan Yang**[1] †

[1]Peng Cheng Laboratory
[2]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3]College of Computer Science and Software Engineering, Shenzhen University
shenxuelin@gml.ac.cn, quanliu@gml.ac.cn
jiayinxu@gml.ac.cn, yangwh@pcl.ac.cn

Abstract

Learned Image Compression (LIC) models face critical challenges in real-world scenarios due to various environmental degradations, such as fog and rain. Due to the distribution mismatch between degraded inputs and clean training data, well-trained LIC models suffer from reduced compression efficiency, while retraining dedicated models for diverse degradation types is costly and impractical. Our method addresses the above issue by leveraging prompt learning under the information bottleneck principle, enabling compact extraction of shared components between degraded and clean images for improved latent alignment and compression efficiency. In detail, we propose an Information Bottleneck-constrained Latent Representation Unifying (IB-LRU) scheme, in which a Probabilistic Prompt Generator (PPG) is deployed to simultaneously capture the distribution of different degradations. Such a design dynamically guides the latent-representation process at the encoder through a gated modulation process. Moreover, to promote the degradation distribution capture process, the probabilistic prompt learning is guided by the Information Bottleneck (IB) principle. That is, IB constrains the information encoded in the prompt to focus solely on degradation characteristics while avoiding the inclusion of redundant image contextual information. We apply our IB-LRU method to a variety of state-of-the-art LIC backbones, and extensive experiments under various degradation scenarios demonstrate the effectiveness of our design. Code is available at `https://github.com/liuquan0521-sys/IB-LRU-compression`.

1 Introduction

In the past decade, Learned Image Compression (LIC) has emerged as a competitive alternative to conventional image coding standards [1, 2, 3] by leveraging deep neural networks to optimize the rate-distortion trade-off in an end-to-end manner. Existing LIC models typically adopt an autoencoder structure, where the encoder extracts a latent representation and the decoder reconstructs the image, characterized by an integrated entropy model that captures spatial dependencies in the latent representation and plays a critical role in compression efficiency. A milestone in entropy modeling is the introduction of the hyperprior [4], which adopts the variational autoencoder (VAE) framework and introduces a Gaussian-based prior to model the distribution of latent representations. This approach enables the compression model to adaptively learn compact latent representations tailored

*These authors contributed equally.
†Correspondence to: Wenhan Yang.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

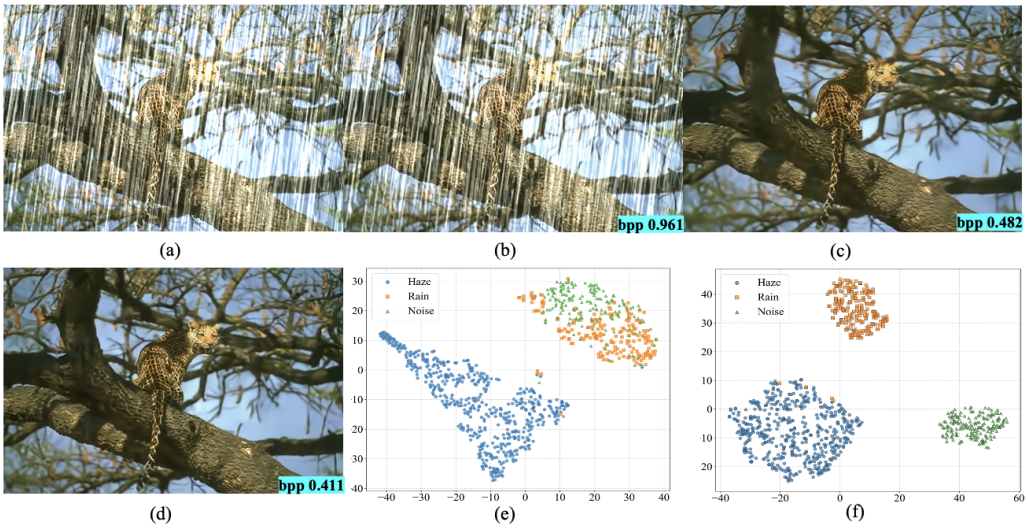


Figure 1: An intuitive illustration of compression performance under degradation scenarios. (a) Input image with rain degradation; (b) Output from a well-trained LIC network [5]; (c) Output from the *restoration-then-compression* paradigm (PromptIR [6] + LIC); (d) Output from our proposed IB-LRU scheme. In addition, we compare the commonly employed factorized prompt [6] (e) and our proposed probabilistic prompt (f) in terms of their capacity to capture discriminative degradation characteristics, visualized via t-SNE plots.

to the statistical properties of natural images, significantly improving rate-distortion performance and forming the foundation for the following LIC models.

Despite their impressive performance under standard conditions, existing LIC models often struggle in real-world settings, where the imaging process typically suffers from various environmental degradations, *e.g.*, noise, fog, rain, or low-light conditions. This limitation arises from their data-driven nature, that are trained on high-quality natural images and relies on learned priors that assume clean image statistics. However, in degraded scenarios, the obtained latent representations deviate significantly from these assumptions. Consequently, as the entropy model is tightly coupled with the prior, it becomes difficult to accurately estimate the distribution of latent features, leading to a substantial drop in compression performance. In other words, compressing degraded images often requires significantly more bits than pristine ones, an intuitive demonstration is provided in Fig. 1 (b). Moreover, the wide range of possible degradation types and intensities in real-world settings makes it impractical to retrain dedicated LIC models for each case to capture their specific distributions.

In facing this challenge, a potential path is *restoration-then-compress* paradigm, which involves deploying cutting-edge *multy-in-one* image restoration models [7, 8, 9, 10] (denoting their capacity to handle multiple degradation within a unified process) at encoding end, with the aim of reducing the degradation in a preprocessing stage so that the restored images can be compressed with an acceptable rate cost. However, in our coding practice, this paradigm does not yield optimal performance. Although it may provide satisfactory perceptual quality, the restored images still exhibit distributional divergence from natural images, leading to suboptimal compression results in the context of LIC, as shown in Fig. 1 (c). Moreover, adding a separate restoration model may reduce flexibility in real-world deployments, especially when computational resources at the imaging end are limited.

To tackle the aforementioned challenges, we introduce the Compressor-as-Restoration paradigm, where prompts are employed to adaptively steer the encoder, thereby achieving a unified framework for compression and restoration. We propose a novel Information-Bottleneck-Constrained Latent Representation Unifying (IB-LRU) scheme, a lightweight encoder-side plugin module that constrains latent representations by aligning degraded and clean-image distributions, thereby enhancing compression and reconstruction without modifying the LIC backbone. In particular, to address diverse input degradations within a unified training framework, we adopt a prompt learning strategy enhanced by a VAE-based Probabilistic Prompt Generator (PPG), which models each degradation using distribution parameters. Unlike existing approaches that rely on factorized prompt vectors, our probabilistic design yields more compact and discriminative representations with fewer parameters,

as shown in Fig. 1 (e) and (f). During inference, degradation information is sampled from the PPG's posterior and passed to a Degradation-Adaptive Gating Modulation (DAGM) module, which guides the encoding process. Additionally, we incorporate the Information Bottleneck (IB) principle to constrain the mutual information between prompts and image content, encouraging the prompt to focus on essential degradation characteristics. In experiments, we implement our IB-LRU scheme on multiple LIC backbones, and extensive results demonstrate that the proposed scheme effectively improves compression performance under various degradation settings.

Our key contributions are as follows,

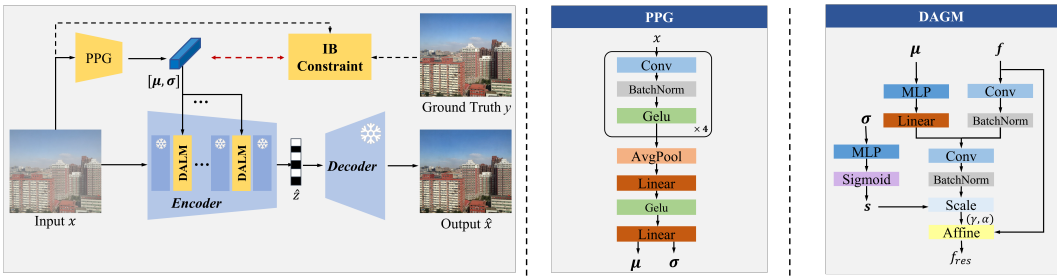
- To the best of our knowledge, the proposed IB-LRU is the first exploration aimed at improving the efficacy of LIC models under various degradation settings through a lightweight plug-and-play design.
- We propose a novel probabilistic prompt learning strategy that characterizes each degradation type using a set of distribution parameters, rather than factorized vectors, resulting in more discriminative degradation representations.
- We introduce an Information Bottleneck (IB)-based optimization criterion for the probabilistic prompt learning process, and derive a variational approximation bound in theory that guides the design of our optimization strategy.

2 Related Work

Learned Image Compression. The past decade has witnessed the rapid advancement of Learned Image Compression (LIC), driven by the progress of deep learning technologies. The first attempt was made by Ballé *et al.* [4], introduced an end-to-end autoencoder pipeline, with a factorized entropy model responsible for constraining the compactness of the latent representation. Furthermore, Ballé *et al.* [11] made a significant advancement by introducing a hyperprior into the entropy coding process, enabling the learning of a hierarchical entropy model in which the distribution of latent features is conditioned on a hyper-latent variable. In the following works, numerous efforts have been made to further enhance the entropy model by leveraging contextual dependencies, such as local context [12], global context information [13], and channel-wise context [14, 15] resulting in incremental improvements. In addition, some studies focus on reducing the computational complexity of the entropy coding process, by leveraging checkerboard context models [16] or sparse sampling methodology [17]. Besides efforts aimed at optimizing the entropy model, other works focus on improving the backbone networks by incorporating architectures such as residual networks [18], invertible neural networks [19, 20], and Swin-Transformers [21, 22, 23, 24]. They enable a more expressive latent representation process, leading to notable improvements in compression performance.

Prompt Learning. The concept of prompt learning originated from the field of natural language processing, involving in inducing pre-trained language models (*e.g.*, BERT [25] or GPT [26]) to generate answers given cloze-style prompts, extracting information useful for downstream tasks. Subsequent research in prompt learning shifted toward automating this process using labeled data, replacing manually designed prompts with learnable ones. For instance, Jiang *et al.*[27] employed text mining and paraphrasing techniques to generate a pool of candidate prompts, from which the optimal ones were selected based on training accuracy. Meanwhile, other studies[28, 29, 30] proposed to factorize prompts into a set of continuous vectors that can be end-to-end optimized with respect to a given objective, a strategy known as continuous prompt learning. In the field of computer vision, CoOp [31] was one of the earliest works to adopt the continuous prompt learning methodology, demonstrating notable improvements in transfer learning performance. Building on this, Zhou *et al.*[32] further applied prompt learning techniques to image classification, achieving significant gains in generalization capability. Similarly, Ju *et al.*[33] explored the use of prompt learning to reformat video-related tasks in alignment with pre-training objectives. Among existing prompt learning works for machine vision, the most relevant to ours is PromptIR [6], which uses a set of factorized vectors as prompts to capture the characteristics of different degradation types for a multi-in-one image restoration network. However, the naive prompt representation and learning strategy yield suboptimal performance in capturing clear and distinct features of various degradation.

Information Bottleneck The Information Bottleneck (IB) principle, originally proposed by Tishby *et al.* [34], offers a theoretical framework for extracting the most relevant information from input



(a) (b) (c)

Figure 2: (a) Compression pipeline of the proposed IB-LRU scheme. (b) Details of the proposed PPG. (c) Details of the proposed DAGM module.

data with respect to the task target by compressing the input representation while preserving task-relevant features. In the context of deep learning, the IB framework has been adapted to encourage models to learn compact and generalizable representations by constraining the mutual information between the input and intermediate features[35]. In particular, this work introduced the variational information bottleneck (VIB), which formulates the IB objective in a tractable form using variational approximations, enabling its integration into neural network training pipelines. More recently, the IB principle has been employed in computer vision tasks to enhance robustness and generalization [7, 36, 37, 38]. These works inspire our use of IB to guide prompt learning, aiming to extract clear degradation-relevant information while suppressing redundancy from irrelevant image context.

3 Methodology

3.1 Motivation and Overview

Existing LIC models typically follow an autoencoder-like pipeline. An encoder $G_a(\cdot)$ maps the input image x to a latent representation $z = G_a(x)$, which is then fed to a uniform quantizer that converts z to its discrete form $\hat{z}$. Subsequently, an entropy encoder $R(\cdot)$ losslessly encodes $\hat{z}$ into a binary bitstream, transmits it to the decoder side, and provides its rate estimation. At the decoder side, the decoder $G_s(\cdot)$ reconstructs the image $\hat{x} = G_s(\hat{z})$. The entire framework is optimized under the *rate-distortion* criterion,

$$\mathcal{L}_{rd} = R(\hat{z}) + \lambda D(x, \hat{x}), \tag{1}$$

where $R(\hat{z})$ is the estimated *rate* measured in bits per pixel (bpp), $D(\cdot, \cdot)$ denotes the *distortion* metric (*e.g.*, MSE: $||x - \hat{x}||^2$), and λ is a Lagrange parameter, being responsible for obtaining codecs at different compression levels. However, while LIC models perform well on clean natural images, their effectiveness degrades in real-world conditions involving noise, fog, blur, *et al*. This is because:

- The latent distributions of degraded images deviate significantly from those of clean images;
- The entropy model relies heavily on prior assumptions, making accurate distribution estimation challenging under degradation.

To address the generalization issues of LIC models, we propose an Information-Bottleneck-Constrained Latent Representation Unifying (IB-LRU) scheme, whose central idea is to perform prompt optimization with information bottleneck constraint to guide the encoding of degraded images so that the distribution of their latent representations aligns with that of pristine images, thereby improving compactness and ensuring high-quality outputs at the decoder side. In general, the proposed IB-LRU scheme is designed with three key objectives: flexibility through a lightweight plug-in design; compatibility with existing LIC backbones without requiring parameter modification; and generalizability to support diverse degradation types within a single training process.

The overall framework is illustrated in Fig. 2. The proposed IB-LRU scheme includes two main components: 1) **Probabilistic Prompt Generator** (**PPG**): A VAE-based encoder that learns degradation-specific prompts using a set of distribution parameters drawn from a multivariate Gaussian. At inference time, degradation representations are sampled and used to guide the encoder; 2) **Degradation-Adaptive Gating Modulation** (**DAGM**): A modulation module that adjusts the encoding process using the sampled prompt, effectively aligning the latent distribution across degradation types. Moreover, to promote probabilistic prompt learning to capture clear and distinct representations of different

degradations, we investigate: 3) **Information Bottleneck (IB) Principle**: a variational approximation of IB in our context is explored to constrain the mutual information between the learned prompt and the input image while preserving the task-relevant information, thus suppressing redundant information of the learned prompt from the broader image context. The detailed implementations of the proposed PPG and DAGM are presented in Subsection 3.2, while the derivation of the variational approximation of IB is provided in Subsection 3.3.

3.2 Prompt-Guide Latent Representation Alignment

Probabilistic Prompt Generator: As shown in Fig. 2 (b), the PPG involves a a VAE-style encoder, denoted as $\mathbf{p} = G(x; \boldsymbol{\theta})$,where $\boldsymbol{\theta}$ represents the model parameters and $\mathbf{p}$ is the probabilistic prompt characterized by its distribution parameters $[\mu_{\mathbf{p}}, \sigma_{\mathbf{p}}]$. In particular, during the inference stage, they are sampled from the approximate posterior $q(\mathbf{p}|x; \boldsymbol{\theta}) = \mathcal{N}(\mathbf{p}; \mu_{\mathbf{p}}, \text{diag}(\sigma^2_{\mathbf{p}})$, where $\mu_{\mathbf{p}} \in \mathbb{R}^d$ is the mean vector, and the diagonal covariance matrix is given by $\text{diag}(\sigma^2_{\mathbf{p}})$, with $\sigma^2_{\mathbf{p}} \in \mathbb{R}^d$ representing the per-dimension variances. Herein, d denotes the dimensionality of the prompt.

Degradation-Adaptive Gating Modulation: These posterior statistics, $\mu_{\mathbf{p}}$ and $\log \sigma^2_{\mathbf{p}}$, are then fed into a gated modulation process, as shown in Fig. 2 (c). Given an intermediate feature f from a frozen backbone LIC encoder, a modulation network $\text{M}_{\text{mod}}(\cdot)$ tasks as input the $\mu_{\mathbf{p}}$ and f to produce basis modulation maps $[\gamma_{\text{b}}, \alpha_{\text{b}}]$:

$$[\gamma_{\text{b}}, \alpha_{\text{b}}] = \text{M}_{\text{mod}}(\mu_{\mathbf{p}}, f). \tag{2}$$

Concurrently, an MLP layer computes a gating signal s from $\log \sigma^2_{\mathbf{p}}$, which provide information related to the spread of the approximate posterior. This signal s derives final modulation maps $[\gamma, \alpha]$:

$$\begin{aligned} s &= \text{sigmoid}(\text{MLP}(\log \sigma^2_{\mathbf{p}})), \\ \gamma &= 1 + s \odot (\gamma_{\text{b}} - 1), \\ \alpha &= s \odot \alpha_{\text{b}}. \end{aligned} \tag{3}$$

The restored feature f_{res} is then obtained via an affine transformation:

$$f_{\text{res}} = \gamma \odot f + \alpha. \tag{4}$$

3.3 Information Bottleneck-based Prompt Learning Constraint

As for the training process of the proposed PPG, we posit that an effective prompt $\mathbf{p}$ should act as a minimal sufficient statistic of the degradation signal relative to the content already captured in the intermediate feature f of the frozen backbone LIC encoder. Directly learning $\mathbf{p}$ might lead it to capture redundant information, *e.g.*, image content information, which would inevitably prevent the learned prompt from capturing clean and distinct characteristics of different degradation types. Therefore, we investigate the Information Bottleneck (IB) principle [34] to provide reliable and effective guidance for the prompt learning process. In the context of IB, we aim to find a prompt $\mathbf{p}$ that minimizes the mutual information it retains about the input x while maximizing the mutual information relevant to the target task y, which denotes the pristine ground truth in this case. We formulate the IB-based optimization objective for $\mathbf{P}$ [3] as:

$$\mathcal{L}_{\text{IB\_prompt}} = \mathcal{I}(\mathbf{P}; X) - \beta \mathcal{I}(\mathbf{P}; Y|F), \tag{5}$$

where the left term represents the information (potentially redundant) about the input X that $\mathbf{P}$ contains, and the right term quantifies the additional information $\mathbf{P}$ offers about the target Y beyond that already captured by the backbone features F, β denotes the trade-off parameter.

As directly optimizing Eq. (5) is intractable, we follow Variational Information Bottleneck (VIB) [35] to derive tractable variational upper and lower bounds for the left and right terms, respectively.

Variational Lower Bound for $\mathcal{I}(\mathbf{P}; Y|F)$: By definition, $\mathcal{I}(\mathbf{P}; Y|F) = \mathcal{H}(Y|F) - \mathcal{H}(Y|\mathbf{P}, F)$. Since F is a deterministic function of X (frozen), $\mathcal{H}(Y|F)$ can be treated as constant with respect to $\mathbf{P}$. Thus, maximizing $I(\mathbf{P}; Y|F)$ is equivalent to minimizing the conditional entropy $\mathcal{H}(Y|\mathbf{P}, F)$:

$$\mathcal{H}(Y|\mathbf{P}, F) = -\mathbb{E}_{(\mathbf{p},f)\sim p(\mathbf{p},f)} \mathbb{E}_{y\sim p(y|\mathbf{p},f)}[\log p(y|\mathbf{p}, f)]. \tag{6}$$

[3] In the following derivations, X, Y, $\mathbf{P}$, F represent random variables, while x, y, $\mathbf{p}$ and f are scalar or single instances of random variables

We apply a variational approximation $q(y|\mathbf{p}, f)$ for $p(y|\mathbf{p}, f)$, corresponding to our decoder. Using the non-negativity of KL divergence, $D_{\mathrm{KL}}\left[p(y|\mathbf{p}, f) \parallel q(y|\mathbf{p}, f)\right] \geq 0$, we have for any given $\mathbf{p}$, f:

$$\mathbb{E}_{p(y|\mathbf{p},f)}[\log p(y|\mathbf{p}, f)] \geq \mathbb{E}_{p(y|\mathbf{p},f)}[\log q(y|\mathbf{p}, f)]. \tag{7}$$

Substituting this inequality (after taking expectation over $p(\mathbf{p}, f)$) into the entropy expression yields an upper bound for the conditional entropy:

$$\mathcal{H}(Y|\mathbf{P}, F) \leq -\mathbb{E}_{p(y,\mathbf{p},f)}[\log q(y|\mathbf{p}, f)]. \tag{8}$$

Therefore, a variational lower bound for the mutual information is:

$$\mathcal{I}(\mathbf{P}; Y|F) \geq \mathbb{E}_{p(y,\mathbf{p},f)}[\log q(y|\mathbf{p}, f)]. \tag{9}$$

Given the above derivation, we found that maximizing this lower bound is equivalent to maximizing the expected log-likelihood $\mathbb{E}_{p(y,\mathbf{p},f)}[\log q(y|\mathbf{p}, f)]$. Thus, in actual implementation, the *distortion* term in Eq. (1) is employed and acting as a proxy for $-\log q(y|\mathbf{p}, f)$.

Variational Upper Bound for $\mathcal{I}(\mathbf{P}; X)$: We aim to derive a tractable upper bound for the mutual information $\mathcal{I}(\mathbf{P}; X)$. By definition, we obtain:

$$\begin{aligned}\mathcal{I}(\mathbf{P}; X) &= \mathbb{E}_{(x,\mathbf{p})\sim p(x,\mathbf{p})}\left[\log \frac{p(\mathbf{p}|x)}{p(\mathbf{p})}\right] \\ &= \mathbb{E}_{(x,\mathbf{p})\sim p(x,\mathbf{p})}[\log p(\mathbf{p}|x)] - \mathbb{E}_{\mathbf{p}\sim p(\mathbf{p})}[\log p(\mathbf{p})].\end{aligned} \tag{10}$$

As the true prior $p(\mathbf{p}) = \mathbb{E}_{x\sim p(x)}[p(\mathbf{p}|x)]$ is intractable, we introduce $r(\mathbf{p})$, a standard Gaussian distribution, as a approximation to the true prior $p(\mathbf{p})$: Using the non-negativity of the KL divergence between the true prior and our variational approximation, $D_{\mathrm{KL}}\left[p(\mathbf{p}) \parallel r(\mathbf{p})\right] \geq 0$, we have:

$$-\mathbb{E}_{\mathbf{p}\sim p(\mathbf{p})}[\log p(\mathbf{p})] \leq -\mathbb{E}_{\mathbf{p}\sim p(\mathbf{p})}[\log r(\mathbf{p})]. \tag{11}$$

Substituting this inequality back into Eq. (10):

$$\mathcal{I}(\mathbf{P}; X) \leq \mathbb{E}_{(x,\mathbf{p})\sim p(x,\mathbf{p})}[\log p(\mathbf{p}|x)] - \mathbb{E}_{\mathbf{p}\sim p(\mathbf{p})}[\log r(\mathbf{p})]. \tag{12}$$

This bound still involves the true posterior $p(\mathbf{p}|x)$ inside the expectations. Now, we further approximate the expectation involving the true posterior by replacing $p(\mathbf{p}|x)$ with our PPG $q(\mathbf{p}|x; \boldsymbol{\theta})$. This step makes the bound computable using samples from the encoder. Recognizing the structure resembles the KL divergence, we arrive at the commonly used VIB upper bound for mutual information:

$$\begin{aligned}\mathcal{I}(\mathbf{P}; X) &\leq \mathbb{E}_{x\sim p(x)}\, \mathbb{E}_{\mathbf{p}\sim p(\mathbf{p}|x)}\left[\log \frac{p(\mathbf{p}|x)}{r(\mathbf{p})}\right] \\ &\approx \mathbb{E}_{x\sim p(x)} D_{\mathrm{KL}}[q(\mathbf{p}|x) \parallel r(\mathbf{p})].\end{aligned} \tag{13}$$

Things have to be mentioned that, although replacing the true posterior $p(\mathbf{p}|x)$ with the approximate posterior $q(\mathbf{p}|x; \boldsymbol{\theta})$ inside the expectation introduces an additional approximation beyond the rigorous bound derived using $D_{\mathrm{KL}}(p \parallel r) \geq 0$, Eq. (13) provides a practical objective for IB implementations.

With the trackable IB-based optimization creation for prompt learning, the final loss function for our IB-LRU scheme is formulated as,

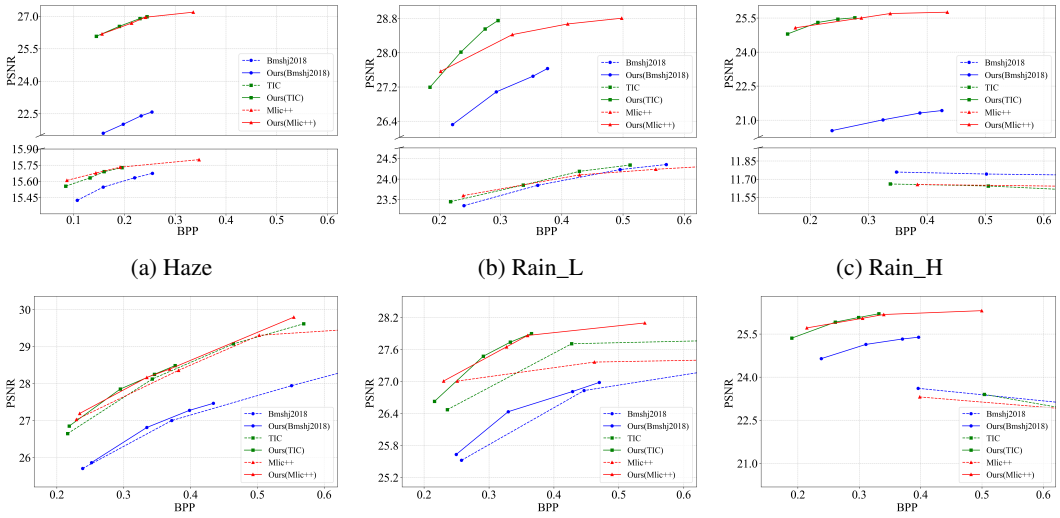
$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{rd}} + \beta\, \mathbb{E}_{x\sim p(x)} D_{\mathrm{KL}}[q(\mathbf{p}|x) \parallel r(\mathbf{p})]. \tag{14}$$

4 Experiments

In the experimental stage, we first implement the proposed IB-LRU scheme on multiple backbone LIC networks and evaluate its effectiveness across a variety of degradation settings. Subsequently, we compare its performance with the *restoration-then-compression* paradigm, which is built upon cutting-edge *multy-in-one* restoration models. Moreover, comprehensive ablation studies are conducted to validate the effectiveness of our design.

Table 1: Comparison between the proposed scheme and the *restoration-then-compression* paradigm regrading parameters and inference speed.

Model	**AirNet+TIC**	**Restormer+TIC**	**PromptIR+TIC**	**MOCE-IR+TIC**	**Ours+TIC**
Extra_Param.	8.93M	26.12M	35.59M	25.35M	4.6M
Inference Speed (ms)	483.3	363.9	416.5	349.4	109.5



(a) Haze (b) Rain_L (c) Rain_H

(d) Noise_15 (e) Noise_25 (f) Noise_50

Figure 3: Performance comparison between the LIC backbone networks with and without the integration of the proposed IB-LRU scheme.

4.1 Experimental Setting

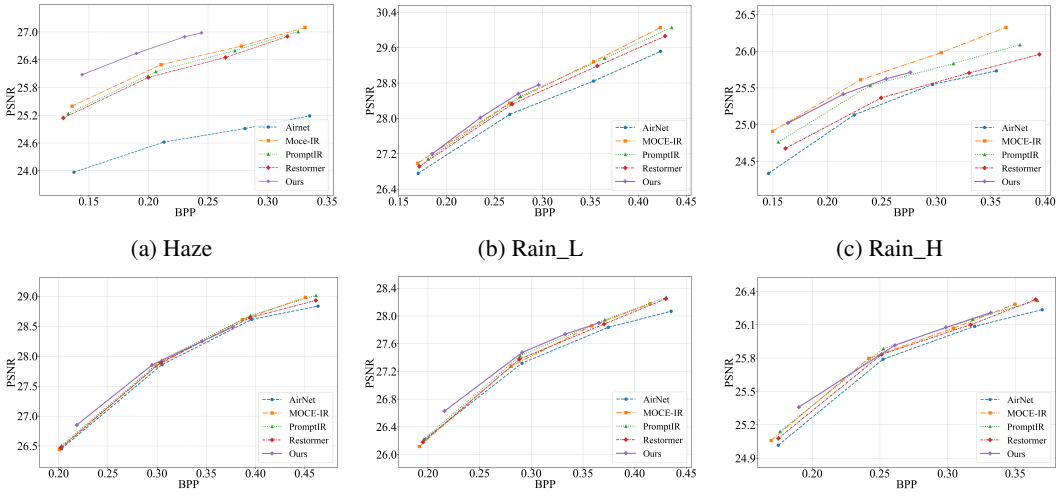
Benchmark: The performance is assessed across three degradation types, including haze, rain, and noise, with multiple degradation intensities employed for each type: 1) **Rain:** The Rain100 dataset [39] consists of 2,000 degraded-clean pairs for each of the *heavy* and *light* rain scenarios. In our experiments, 1,800 pairs are used for training and 200 pairs for testing in each scenario. 2) **Noise:** For the noise scenarios, the training set was constructed using 400 images from BSD400 [40] and 4,744 images from WED [41], with degraded versions generated by adding Additive White Gaussian Noise (AWGN) with $\sigma \in \{15, 25, 50\}$. The testing set is formed by combining BSD68 [42] and Urban100 [43], using the same noise settings. 3) **Haze:** The SOTS dataset [44] is employed, with 72,135 / 500 images for training /testing. The training sets of all above datasets are combined to support the training of our IB-LRU scheme.

LIC backbone: Three widely adopted LIC networks are employed, including the milestone Bmshj2018 [11], as well as two cutting-edge models: the Swin Transformer-based TIC [5] and MLIC++ [45]. These three LIC backbone networks were pre-trained exclusively on the clean COCO 2017 dataset. For each model, four checkpoints corresponding to different compression levels are employed, and their parameters are kept frozen during the training of our IB-LRU scheme.

Anchors: Four cutting-edge *multy-in-one* image restoration networks are employed for the *restoration-then-compression* paradigm, including AirNet[8], Restormer[9], and PromptIR[6], and MOCE-IR[46]. All models are trained from scratch using the same training set as ours to ensure fair comparisons. The restored output images are then fed into well-trained LIC backbones for evaluation.

Evaluation Criteria: We introduce *rate-perception* criterion for this unique task. In this context, *perception* refers to the distance between decoded images and corresponding pristine ground truths, measured in terms of PSNR (dB), while the bit-per-pixel (bpp) value is used to quantify the *rate*.

Implementation Details: We train our model using the AdamW optimizer with settings $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 1×10^{-4}. This rate is maintained for the first 80 epochs (80% of the total training) and then decayed to 1×10^{-5} for the remaining 20 epochs. The



(a) Haze (b) Rain_L (c) Rain_H

(d) Noise_15 (e) Noise_25 (f) Noise_50

Figure 4: Performance comparison between the *restoration-then-compression* paradigm and our IB-LRU scheme, both of which are based on the TIC backbone.

model is trained for a total of 100 epochs. Training is performed on randomly cropped 256×256 patches using a setup of 8 NVIDIA RTX 6000 GPUs.

4.2 Experimental Results

Effectiveness Evaluation: Performance is compared between the LIC backbones with and without our IB-LRU implementation to examine the effectiveness of the proposed scheme. It is worth noting that although the LIC backbones are not designed for image restoration, we focus on bitrate savings in this part, while still reporting their *perception* indices to ensure consistent and clear representation.

The corresponding results, presented in Fig. 3, show encouraging improvements. In particular, under the *light rain*, *heavy rain*, *noise_15*, *noise_25*, and *noise_50* scenarios, our proposed method brings notable improvements in *rate* performance, resulting in average bpp reductions of 27.1%, 51.6%, 18.2%, 34.6%, and 64.9%, respectively, across the three LIC backbones. Moreover, significant improvements in perceptual quality can also be observed, demonstrating the contribution of latent representation unification to the image reconstruction process. It is worth noting that in the *haze* scenario, although the IB-LRU scheme remarkably improves perceptual quality, the bpp values are even increased. According to our analysis and a comprehensive examination of the corresponding dataset, this is primarily because haze tends to act as a smoothing effect on image content rather than introducing additional degradation signals. As a result, hazy images may still be well represented by the prior of the LIC model, as they resemble smooth images.

Comparison with *Restoration-then-Compression* Paradigm: Comparisons results between our propose IB-LRU and the *restoration-then-compression* methods regarding *rate-perception* performances are provided in Fig. 4. As shown, our scheme is capable of achieving competitive perceptual quality compared to these cascaded approaches, while offering overall advantages in bpp reduction—especially in high bitrate regions. In particular, an interesting observation arises from the fact that both our IB-LRU and the *restoration-then-compression* approaches are built upon the same frozen LIC checkpoint. While both are capable of achieving satisfactory perceptual quality, they tend to result in a significant increase in bpp. As previously discussed, the underlying reason is that although the restored images may align well with perceptual preferences, their distributions still diverge considerably from those of natural images due to the generative nature of the restoration models. Consequently, during entropy modeling, these distributions cannot be accurately estimated, leading to increased bitrate costs. Meanwhile, our method focuses on unifying the latent distribution, achieving a better trade-off between *rate* and *perception*. To provide intuitive insight into the performance, a set of visual examples is presented in Fig. 5.

To further demonstrate the *flexibility* of our proposed IB-LRU scheme, we compare the number of additional parameters and inference speed with those of the *restoration-then-compression* paradigm.

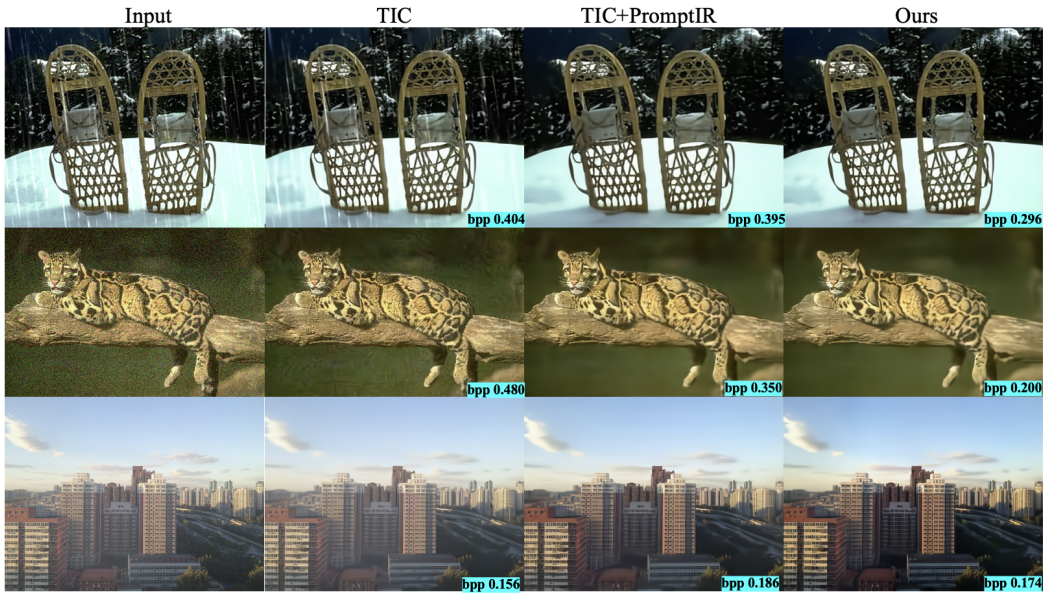
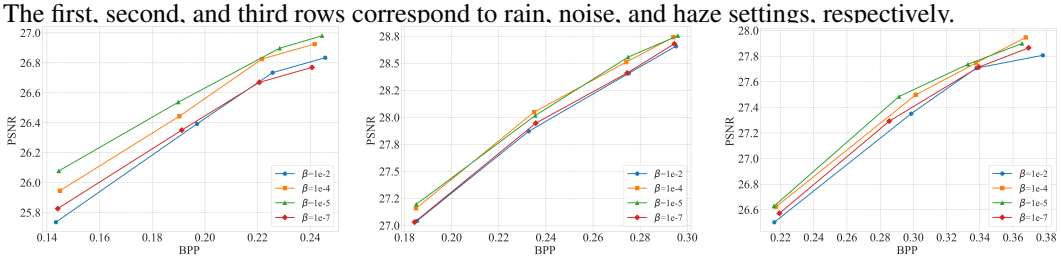


Figure 5: An intuitive illustration of compression performance under different degradation scenarios. The first, second, and third rows correspond to rain, noise, and haze settings, respectively.



(a) Haze (b) Rain_L (c) Noise_25

Figure 6: Ablation results on the IB constraint by varying its weight parameter β.

The corresponding results, presented in Table 1, demonstrate our overwhelming advantages and highlight the strong potential of IB-LRU for real-world implementation.

4.3 Ablation Studies

This part specifically examines the effectiveness of the IB constraint and its influence on the compression pipeline. In particular, the IB principle is employed to regularize the probabilistic prompt, guiding it to capture minimal yet sufficient degradation information. In this part, we retrain the entire scheme with varying values of β in Eq. (14) based on the TIC backbone, specifically $\{1 \times 10^{-2}, 1 \times 10^{-4}, 1 \times 10^{-5}, 1 \times 10^{-7}\}$ to examine the influence of the IB constraint. Fig. 6 illustrates the corresponding results, where we observe that the optimal *rate-perception* performance is generally achieved when β is set to 1×10^{-4} or 1×10^{-5}, whereas values that are too large or too small lead to suboptimal results.

To gain deeper insight,

- When β is too small (*e.g.*, 1×10^{-7}), the IB constraint becomes weak, allowing the prompt to capture excessive redundant image context rather than distinct and clear degradation characteristics. This, in turn, compromises latent representation unification, as well as compression efficiency and restoration quality.

- Increasing β to 1×10^{-5} and 1×10^{-4} significantly improves *rate-perception* performance, as a moderate IB constraint effectively regularizes the prompt, guiding it to discard irrelevant information while retaining what is essential for adaptive restoration.

- When β becomes too large (*e.g.,* 1×10^{-2}), the *rate-perception* performance begins to degrade. This suggests that overly strong IB constraints may excessively compress the prompt, resulting in loss of useful information necessary for capturing degradation characteristics.

5 Conclusion

We propose IB-LRU, a lightweight and modular scheme that enhances the robustness of Learned Image Compression (LIC) models under various real-world degradations. By leveraging a probabilistic prompt guided by the Information Bottleneck principle, our method constrains latent representations to retain task-relevant degradation information while suppressing redundancy. IB-LRU operates without altering the original LIC parameters and is compatible with multiple pre-trained backbones. Empirical results demonstrate consistent gains in compression and restoration performances across diverse degradation scenarios, with minimal additional computational cost. Our findings suggest that latent distribution alignment is a promising direction for improving the generalization of LIC models in practical settings.

Acknowledgements

This work was in part by the Interdisciplinary Frontier Research Project of PCL under Grant 2025QYB013, in part by the Major Key Project of PCL (PCL2025A03), in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010454, in part by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No. GML-KF-24-27, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515011667.

References

[1] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12):1649–1668, 2012.

[2] Siwei Ma, Tiejun Huang, Cliff Reader, and Wen Gao. AVS2? making video coding smarter [standards in a nutshell]. *IEEE Signal Processing Magazine*, 32(2):172–183, 2015.

[3] B. Bross, Y. Wang, Y. Ye, S. Liu, J. Chen, G. Sullivan, and J. Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. In *Proceedings of the International Conference on Learning Representations*, 2017.

[5] Ming Lu, Peiyao Guo, Huiqing Shi, Chuntong Cao, and Zhan Ma. Transformer-based image compression. In *Proceedings of the Data Compression Conference*, pages 469–469, 2022.

[6] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. PromptIR: Prompting for all-in-one image restoration. In *Advances in Neural Information Processing Systems*, volume 36, pages 71275–71293, 2023.

[7] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *Proceedings of the European Conference on Computer Vision*, pages 200–216, 2020.

[8] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17452–17462, 2022.

[9] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[10] Jun Cheng, Dong Liang, and Shan Tan. Transfer CLIP for generalizable image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25974–25984, 2024.

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

[12] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 31, 2018.

[13] Zongyu Guo, Zhizheng Zhang, Runsen Feng, and Zhibo Chen. Causal contextual prediction for learned image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):2329–2341, 2021.

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

[15] Wei Jiang, Jiayu Yang, Yongqi Zhai, Peirong Ning, Feng Gao, and Ronggang Wang. MLIC: Multi-reference entropy model for learned image compression. In *Proceedings of the ACM International Conference on Multimedia*, pages 7618–7627, 2023.

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

[17] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Beyond learned metadata-based raw image reconstruction. *International Journal of Computer Vision*, 132(12):5514–5533, 2024.

[18] F Chen, Y Xu, and L Wang. Two-stage octave residual network for end-to-end image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3922–3929, 2022.

[19] Y Xie, K Cheng, and Q Chen. Enhanced invertible encoding for learned image compression. In *Proceedings of the ACM International Conference on Multimedia*, pages 162–170, 2021.

[20] Shilv Cai, Liqun Chen, Zhijun Zhang, Xiangyun Zhao, Jiahuan Zhou, Yuxin Peng, Luxin Yan, Sheng Zhong, and Xu Zou. I2C: Invertible continuous codec for high-fidelity variable-rate image compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6): 4262–4279, 2024.

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

[22] Yinhao Zhu, Yang Yang, and Taco Cohen. Transformer-based transform coding. In *International Conference on Learning Representations*, 2022.

[23] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17492–17501, 2022.

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

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[26] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[27] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[28] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[30] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[32] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[33] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 105–124, 2022.

[34] N Tishby. The information bottleneck method. *Computing Research Repository*, 2000.

[35] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *Proceedings of the International Conference on Learning Representations*, 2017.

[36] Rana Ali Amjad and Bernhard C Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2225–2239, 2019.

[37] Jing Wang, Yuanjie Zheng, Jingqi Song, and Sujuan Hou. Cross-view representation learning for multi-view logo classification with information bottleneck. In *Proceedings of the ACM International Conference on Multimedia*, pages 4680–4688, 2021.

[38] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6778–6787, 2019.

[39] Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1377–1393, 2019.

[40] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2010.

[41] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016.

[42] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 416–423, 2001.

[43] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2015.

[44] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28 (1):492–505, 2018.

[45] Wei Jiang, Jiayu Yang, Yongqi Zhai, Feng Gao, and Ronggang Wang. MLIC++: Linear complexity multi-reference entropy modeling for learned image compression. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(5):1–25, 2025.

[46] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12753–12763, 2025.

[47] Benoit Brummer and Christophe De Vleeschouwer. On the importance of denoising when learning to compress images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we delineate the scope of our research, analyze the current limitations of Learned Image Compression (LIC), and highlight our proposed methodology and key contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: The paper has limitations, but those are not discussed in the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full set of assumptions and a complete (and correct) proof for each theoretical result are provided in Section 3.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide a comprehensive description of our experimental setup in Section 4, which covers datasets, baselines, evaluation metrics, and pertinent experimental details. To further support reproducibility, our code will be released as open source.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   We have open-sourced our codes with corresponding training and testing scripts).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide our experimental settings, such as datasets and optimizers in Section 4. We will subsequently organize and open-source our code, making detailed specifics/information available.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Owing to constraints on time and page length, the current presentation of our experimental results does not include an analysis of error or other statistical significance measures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In section 4.1, we specified in our implementation details that 8 NVIDIA RTX 6000 GPUs with 48 GB of memory were used for training.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that the research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 5, we discuss the potential positive societal impacts of our work. Furthermore, we do not anticipate any adverse societal impacts from this work.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of all assets used in the paper, including code, data, and models, are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have released our source code, accompanied by comprehensive documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve any crowd sourcing or experiments with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve experiments with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Supplementary Experiments and Analyses

A.1 Effectiveness of the Probabilistic Prompt Design

This section provides a straightforward comparison between our probabilistic prompt and the existing factorized prompt design. Specifically, we replace our probabilistic prompt with the factorized prompt from PromptIR, keep the feature-modulation module unchanged, and retrain the plug-in on the TIC backbone. Table 2 reports PSNR (dB) at a single compression level, where our methods deliver overall performance gains with significantly fewer parameters.

Table 2: Performance and parameter comparison between our proposed probabilistic prompt and the existing factorized prompt.

Prompt Type	**Rain_L (bpp=0.23)**	**Haze (bpp=0.19)**	**Noise_50 (bpp=0.26)**	**Para. (M)**
Probabilistic	28.01	26.53	25.93	4.6
Factorized	27.95	26.39	25.76	16.0

We also assess generalization on unseen conditions, including a mixed haze-and-rain setting created by adding synthetic rain streaks to the SOTS haze test images, and the unseen DID de-raining dataset. Table 3 shows that our method achieves stronger robustness in both cases.

Table 3: Comparison of generalization performance between probabilistic and factorized prompt designs regarding bpp/ PSNR(dB).

Prompt Type	Haze + Rain (Mixed)	Unseen Domain (DID)
Probabilistic	0.183 / 24.09	0.202 / 25.15
Factorized	0.207 / 19.24	0.214 / 22.93

A.2 Effectiveness of the Information Bottleneck (IB) Constraint

We evaluate the impact of the Information Bottleneck by removing it entirely ($\beta = 0$). As reported in Table 4, the absence of the IB constraint leads to a uniform degradation in performance, supporting its role in guiding the prompt to capture essential degradation features.

Table 4: Effect of removing the IB constraint ($\beta = 0$) on rate–distortion (bpp / PSNR).

Version	Rain_L	Haze	Noise_25
Ours (Full Model)	0.235 / 28.01	0.191 / 26.53	0.289 / 27.51
Ablated ($\beta = 0$)	0.232 / 27.61	0.200 / 26.00	0.289 / 26.99

A.3 Comparison with Joint Denoising-Compression Methods

We compare against the joint denoising–compression method of Brummer *et al.* [47] (JDC-CN). For a fair comparison, we retrain both approaches on the same TIC backbone using only the *Gaussian noise* split of our training set. As shown in Table 5, our method demonstrates a notable advantage even in this single-degradation setting.

Table 5: Comparison with joint denoising–compression method (bpp/PSNR(dB)).

Method	Noise_15	Noise_25	Noise_50
Ours	0.271 / 28.23	0.294 / 28.07	0.245 / 26.53
JDC-CN [47]	0.254 / 24.71	0.283 / 24.20	0.265 / 23.01