Supplementary Materials: Compact Latent Representation for Image Compression (CLRIC)

1 Additional Results

1.1 Classical Distortion Metrics

Classical distortion metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Multi-Scale SSIM (MS-SSIM) are not ideally suited for evaluating our approach. This is because the variational autoencoders employed in our method are trained using a perceptual image compression strategy [1]–[3]. These autoencoders typically achieve an average PSNR of around 25 dB [1]. Meanwhile, our model cannot exceed the PSNR of the autoencoder itself, as it operates on latent representations without accessing the original image data.

We use classical distortion metrics as a reference point for comparing various learned image codecs. We calculated the average PSNR for both the Kodak Dataset (see Figure 1) and the CLIC Professional Valid 2020 dataset (see Figure 2). Our approach yields lower PSNR values compared to other learned compression models based on autoencoders, given that PSNR evaluates pixel-wise differences. However, our method provides synthetic textures that closely resemble those in the original images.

The maximum achievable PSNR with our technique is constrained by the maximum PSNR attainable by the original autoencoders (SD, SD-In, and SD-XL) used during training [1].

PSNR increases with higher bitrate; however, it plateaus for models based on SD and SD-In. In contrast, achieving stability requires a higher bitrate when using SD-XL.

A similar trend is observed with MS-SSIM for both the Kodak Dataset and CLIC Professional Valid 2020 dataset (refer to Figures 1 and 2, respectively).

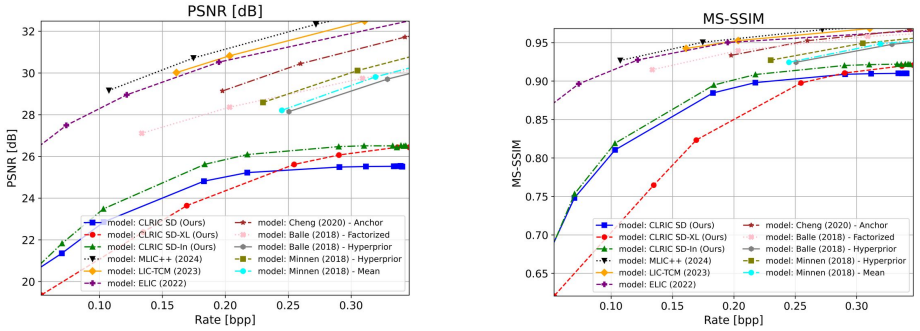

Figure 1: Comparison of PSNR and MS-SSIM metrics for our method and other learned image compression models on the Kodak dataset.

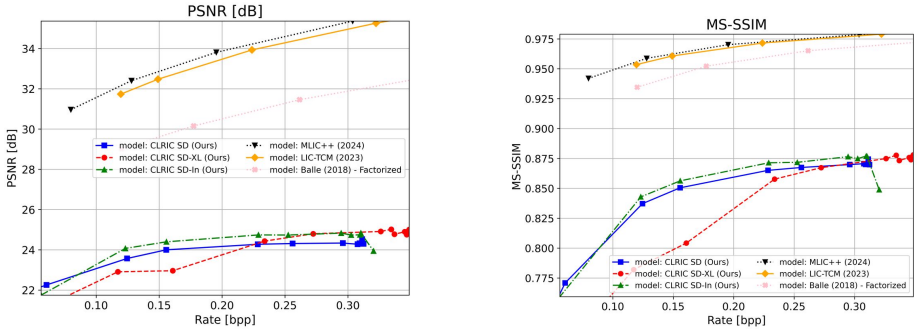

Figure 2: Evaluation of PSNR and MS-SSIM metrics for our method versus other learned image compression models on the CLIC Professional Valid 2020 dataset.

1.2 Continuous Quality Representation

In contrast to other learned image codecs based on autoencoders with a fixed number of quality levels, our method enables the representation of an image's latent variables at any desired quality level. This ranges from extremely low bitrates, where details are significantly blurred, to higher bitrates that closely resemble the original perceptual quality, as illustrated in 3

At very low bitrates, the image details appear highly blurred, and important structural information is lost. As the bitrate increases, the image quality improves until it becomes perceptually acceptable with enhanced detail.

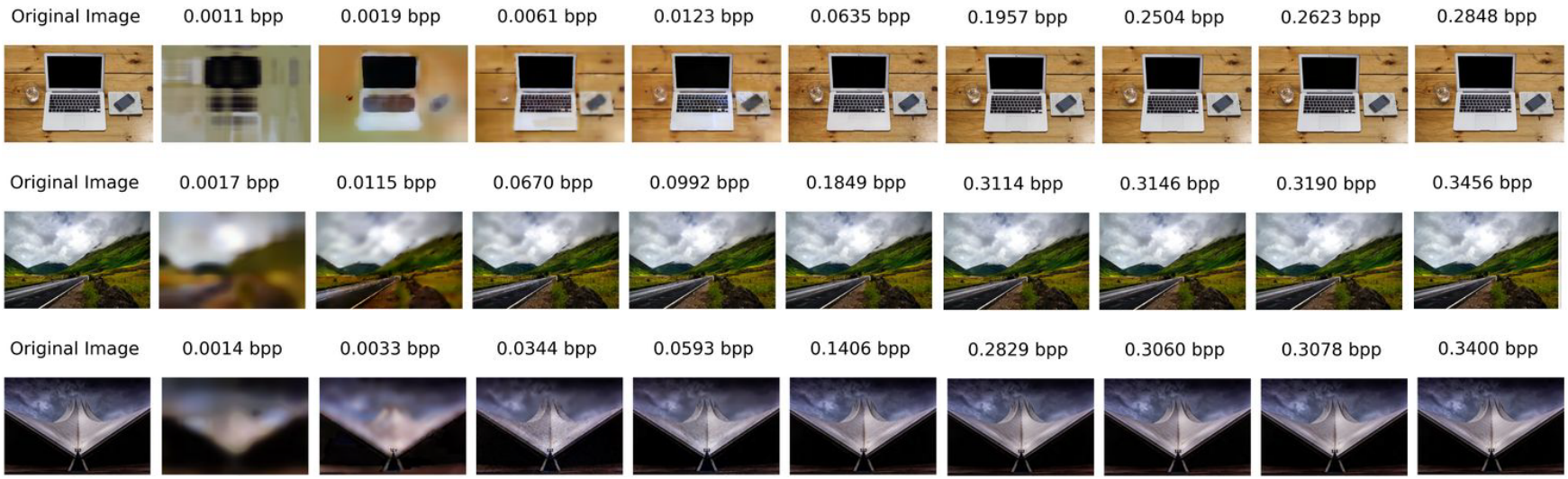

Figure 3: Examples from the CLIC Professional Validation 2020 dataset demonstrating varying bitrates and corresponding quality levels.

1.3 Decoding Complexity Analysis of the overfitted neural function

In our study, we conducted a decoding complexity analysis of an overfitted neural function trained on latent from SD autoencoder. The results demonstrate that the decoding computational complexity of the overfitted function accounts for approximately 26 multiply-accumulate operations per pixel (MAC/Pixel). This complexity slightly decreases as the image size increases, as detailed in Table 1.

This reduction in the overfitted neural function complexity as the number of pixels increases can be attributed to our approach's operation in latent space rather than directly in image space, which reduces the number of parameters involved.

Image size	Latent size	Overfitted Fun. (MAC/Pixel)
512, 768	64 , 96	25.9
1363, 2048	170, 256	25.48

Table 1: Analysis of decoding complexity of two images with different sizes.

1.4 Limitations

Despite our method achieves state-of-the-art performance in perceptual image quality and low decoding complexity, it presents significant challenges related to encoding complexity and time. Encoding a single image takes approximately 10 minutes on an NVIDIA GeForce GTX 1080 Ti (11 GB) GPU, which limits its practicality for mobile devices. However, this approach could be advantageous for high-demand images or videos (in future work), as the training can be executed once on a server, making evaluation feasible across various devices. Another limitation is the challenge of accurate pixel-wise reconstruction. Although our method excels in perceptual similarity, it is not suitable for compression tasks that require precise pixel-wise fidelity.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 10 674–10 685, ISBN: 978-1-66546-946-3.

[2] A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach, *Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation*, Mar. 2024.

[3] D. Podell, Z. English, K. Lacey, *et al.*, *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*, Jul. 2023.

2 Additional Figures

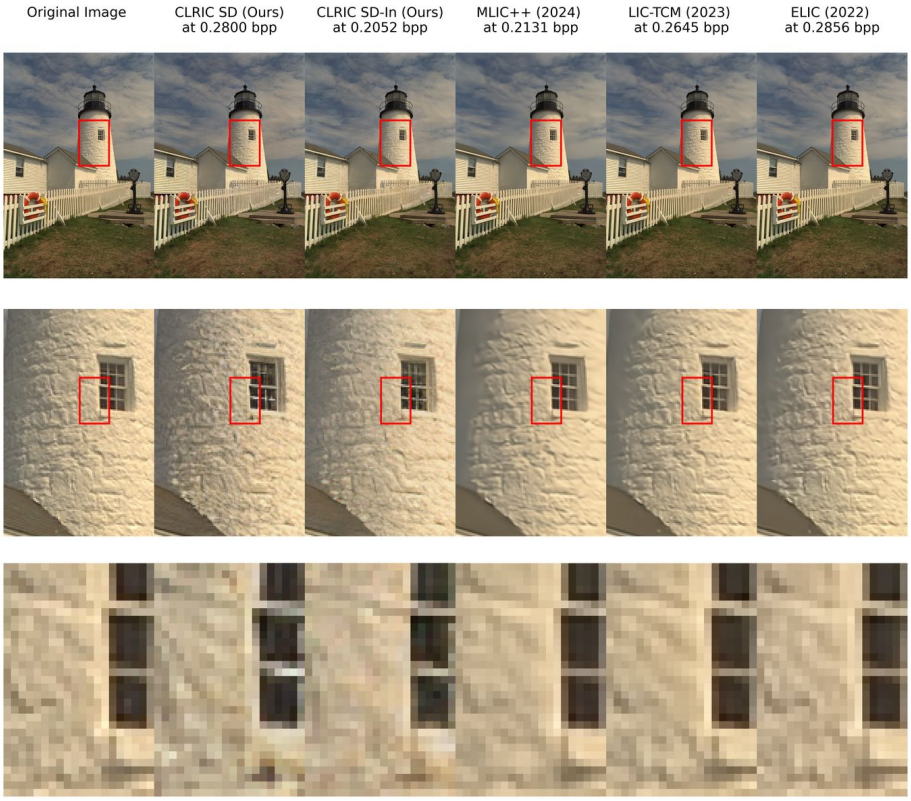

Figure 4: Comparison of our approach against various models on image number 19 from the Kodak dataset.

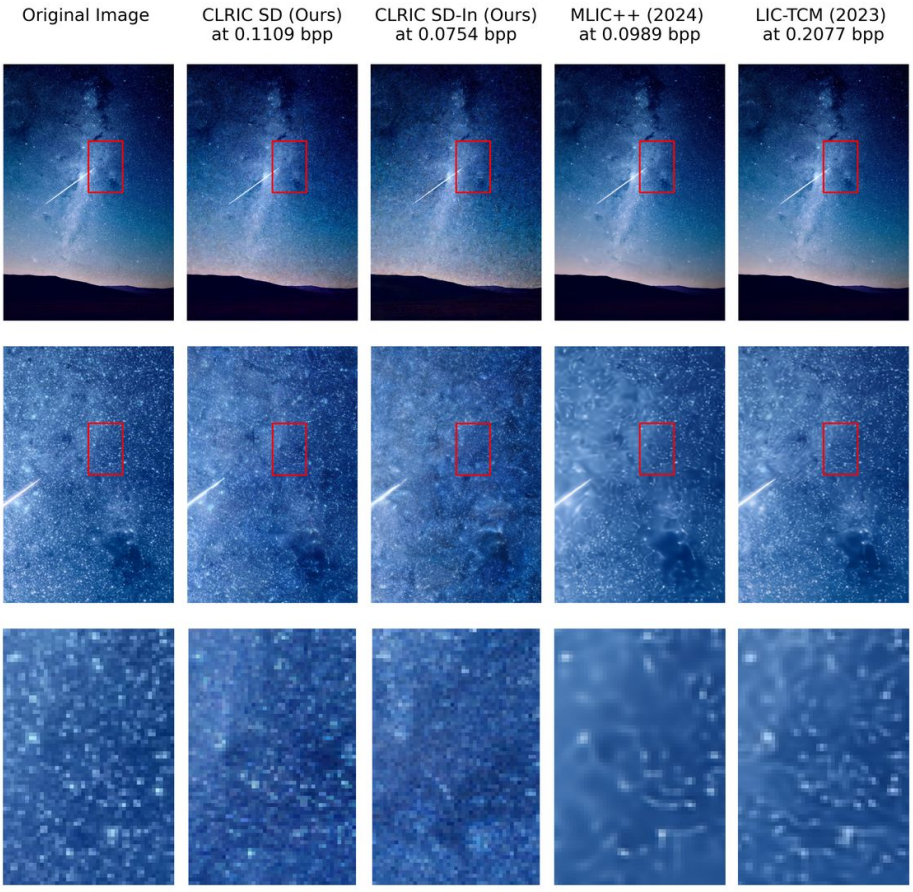

Figure 5: Comparison of our approach against various models on image number 18 from the CLIC Professional Valid 2020 dataset.

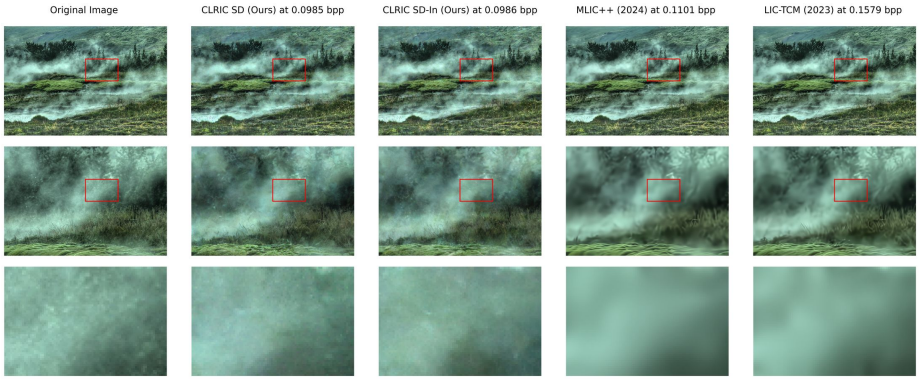

Figure 6: Comparison of our approach against various models on image number 14 from the CLIC Professional Valid 2020 dataset.

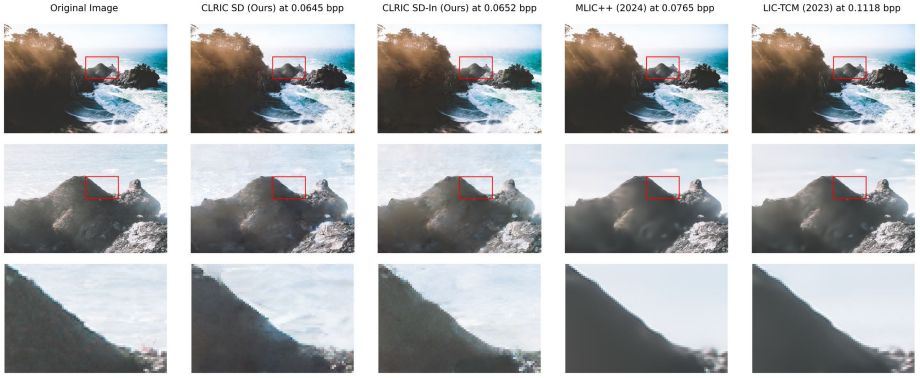

Figure 7: Comparison of our approach against various models on image number 23 from the CLIC Professional Valid 2020 dataset.

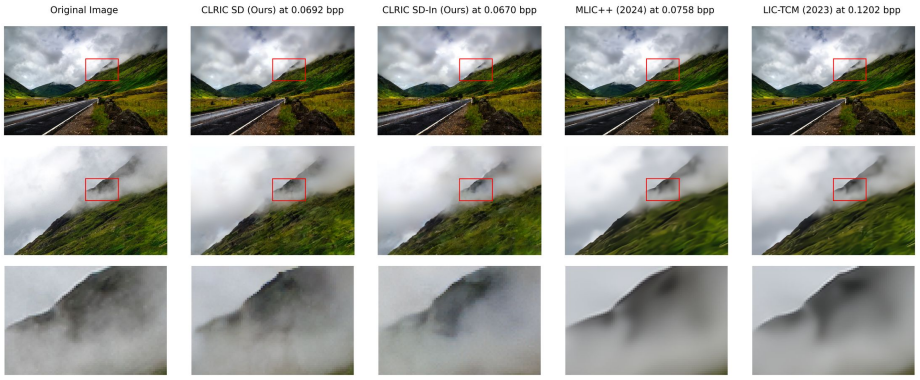

Figure 8: Comparison of our approach against various models on image number 7 from the CLIC Professional Valid 2020 dataset.

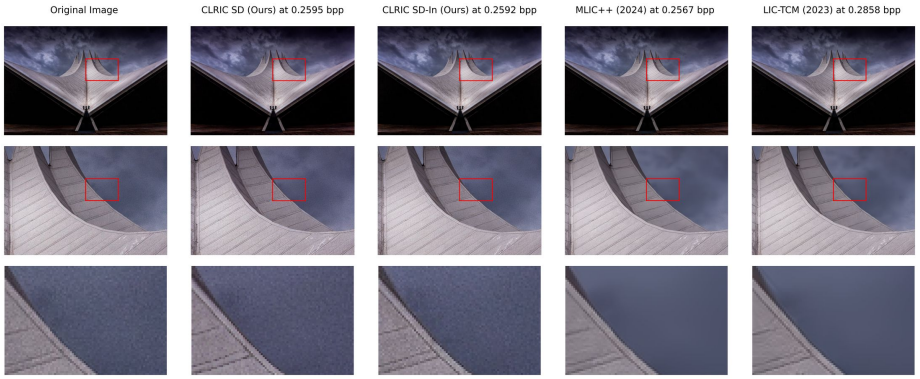

Figure 9: Comparison of our approach against various models on image number 20 from the CLIC Professional Valid 2020 dataset.

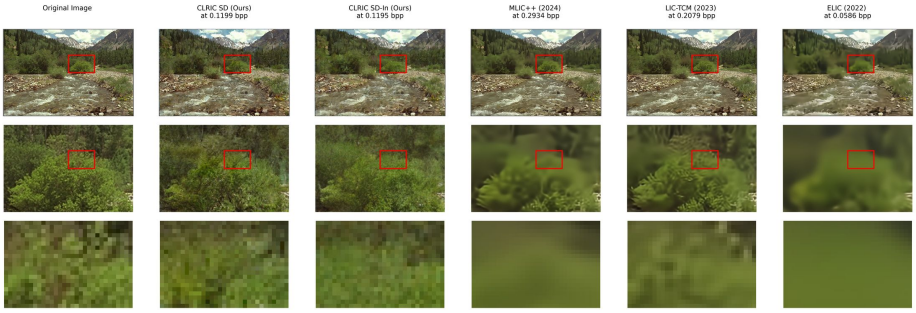

Figure 10: Comparison of our approach against various models on image number 13 from the Kodak dataset.

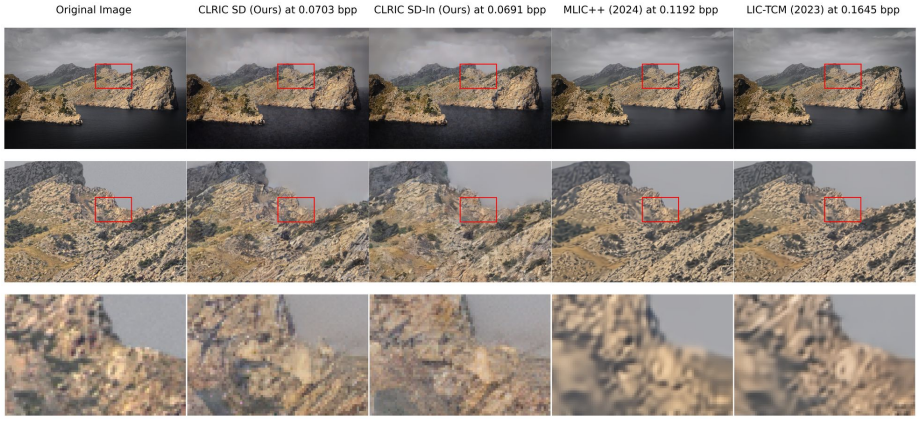

Figure 11: Comparison of our approach against various models on image number 39 from the CLIC Professional Valid 2020 dataset.

Figure 12: Comparison of our approach against various models on image number 6 from the CLIC Professional Valid 2020 dataset.

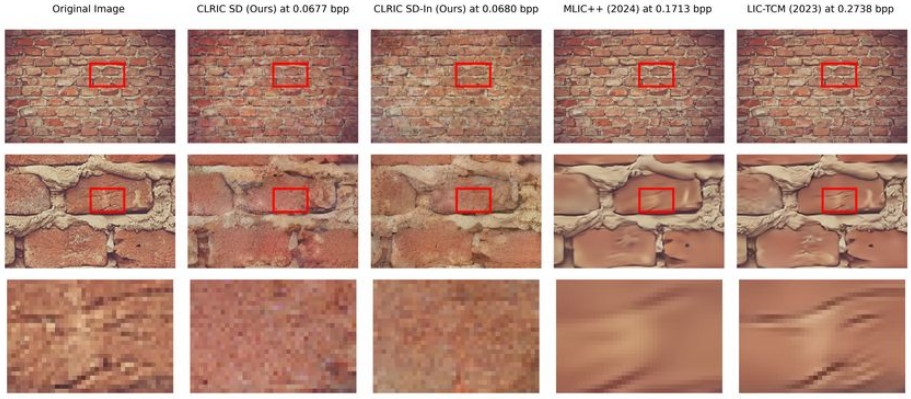

Figure 13: Comparison of our approach against various models on image number 21 from the CLIC Professional Valid 2020 dataset.

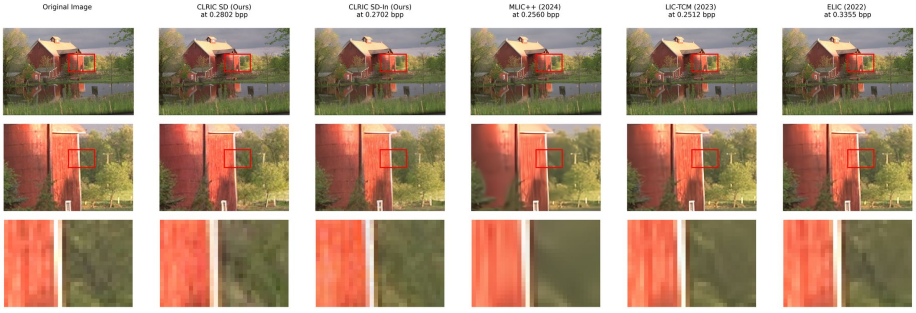

Figure 14: Comparison of our approach against various models on image number 22 from the Kodak dataset.