PICD: Versatile Perceptual Image Compression with Diffusion Rendering

Supplementary Material

A. Implementation Details

A.1. Neural Network Architecture of Text-conditioned MLIC

In Figure 1, we illustrate the neural network architecture of text conditioned MLIC model.

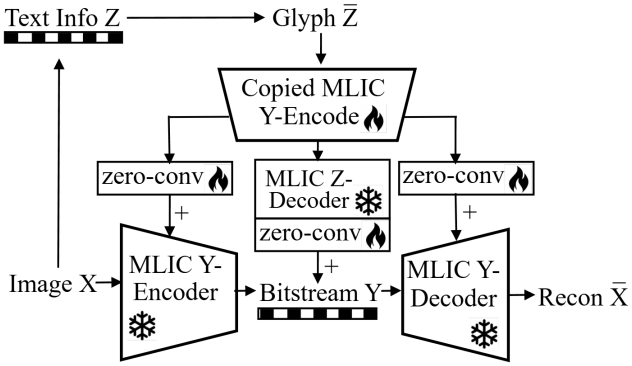

Figure 1. The neural network architecture of text-conditioned MLIC.

A.2. Neural Network Architecture of Proposed Adaptor

In Figure 2, we illustrate the adaptor's neural network architecture of vanilla ControlNet [14], StableSR [11] and our proposed approach.

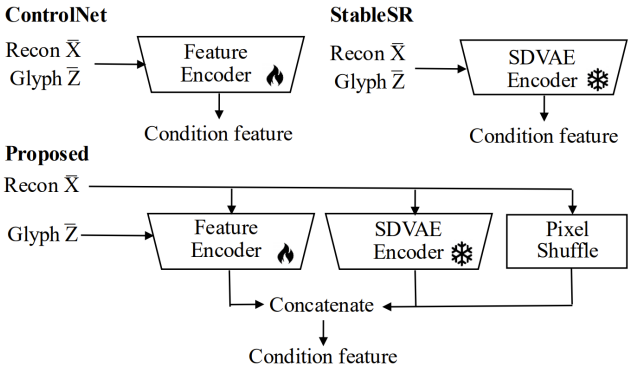

Figure 2. The neural network architecture of the proposed adaptor.

A.3. Instance Level Guidance

To implement instance level guidance, we first need to obtain $\mathbb{E}[X_0|X_t, y]$ using Tweedie's formula following Chung et al. [4]:

$$\mathbb{E}[X_0|X_t, y] = \frac{1}{\sqrt{\bar{\alpha}_t}}(X_t + (1-\bar{\alpha}_t)s_\theta(X_t, t, y)), \quad (1)$$

where $s_\theta(.,.,.)$ is the trained score estimator of diffusion model.

The instance level guidance is composed of OCR guidance and codec guidance. The codec guidance is straightforward and details can be found in Xu et al. [12]. While the OCR guidance is not that straightforward.

We adopt Tesseract OCR engine [9] to extract text from images, following Tang et al. [10]. However, this OCR engine is not differentiable. And we can not use it in instance level OCR guidance. To solve this problem, we alternatively adopt the neural network based OCR engine named PARSeq Bautista and Atienza [1], which is adopted in Lai et al. [6].

Next, we use the bounding box information in Z to cut the source image $\mathbb{E}[X_0|X_t, y]$. Then, those slice of images are feed into PARSeq. PARSeq produces the logits, which is further compared with the true text content in Z (weighted by ζ_1 in Section 3.5) as guidance for diffusion model.

A.4. Hyper-parameters of Diffusion Rendering

In Table 1, we show the hyper-parameters used for diffusion rendering.

	Diffusion rendering hyper-parameter.
SCI1K	$T = 250, \zeta_1 = 0.25, \zeta_2 = 1e-4, \omega = 0.0$
SIQAD	$T = 250, \zeta_1 = 0.25, \zeta_2 = 1e-4, \omega = 0.0$
Kodak	$T = 500, \zeta_1 = 0.25, \zeta_2 = 0.0, \omega = 3.0$
CLIC	$T = 500, \zeta_1 = 0.25, \zeta_2 = 0.0, \omega = 3.0$

Table 1. Diffusion rendering related hyper-parameters.

B. Additional Experimental Results

B.1. Additional Experimental Setup

All the experiments are conducted on a computer with 1 A100 GPU. For the domain level finetuning, we train the LoRA augumented Stable Diffusion 2.0 model with batchsize 64 and 10,000 steps of gradient ascent. We use a learning rate of 1e-4 and a LoRA with rank 256. The training costs around 2 days. For the adaptor training, we adopt a batchsize 64 and 5,000 steps of gradient ascent with learning rate 1e-4 and batchsize 64. The training cost around 1 day. Note that the domain level finetuning only happens

once. While for each bitrate, we need to train a different adaptor.

B.2. Additional Quantitive Results

For RD performance, we also evaluate the LPIPS metric for screen contents, which is shown in Table 2. And in Figure 3, we present the RD curve on SIQAD and CLIC dataset.

	SCI1K (Screen)	SIQAD (Screen)
	BD-LPIPS↓	BD-LPIPS↓
MSE Optimized Codec		
MLIC [5] (Baseline)	0.000	0.000
VTM-SCC [2]	0.055	0.021
Perceptual Optimized Codec		
Text-Sketch [7]	0.135	0.087
CDC [13]	0.100	0.024
MS-ILLM [8]	**-0.023**	**-0.082**
PerCo [3]	0.001	-0.070
PICD (Proposed)	<u>-0.005</u>	<u>-0.080</u>

Table 2. LPIPS results on screen images. **Bold** and <u>Underline</u>: Best and second best performance in perceptual codec.

B.3. Additional Qualitative Results

We present more qualitative results in Figure 5-6.

B.4. Additional Ablation Studies

Classifier-free Guidance Additionally, in the context of PICD for natural image compression, we discovered the significant importance of classifier-free guidance (CFG). Table 3 illustrates that varying levels of CFG markedly affect the FID and PSNR. Through empirical evaluation, we determined that a CFG value of 3.0 optimizes results, yielding the best FID, CLIP similarity, and LPIPS. This finding is consistent with observations reported by Careil et al. [3].

B.5. MS-SSIM as Perceptual Metric

In both our setting and other papers (ILLM), MS-SSIM aligns more with PSNR than visual quality. In our case, for SCI1K dataset, the BD-MS-SSIM is: MLIC (0.01) > VTM (0.00) > ILLM (-0.003) > PICD (-0.006). We are reluctant to use MS-SSIM as perceptual metric, as it is obviously not aligned with visual quality. In CLIC codec competition [50], the best human rated codec has almost worst MS-SSIM. We will emphasis that MS-SSIM is not a perceptual metric, and include those results.

CFG	FID↓	PSNR↑	CLIP↑	LPIPS↓
0.0	71.37	24.47	0.9247	0.1498
3.0	63.76	24.25	0.9356	0.1464
5.0	68.00	23.74	0.9274	0.1555
7.0	70.14	23.67	0.9213	0.1575

Table 3. Ablation study on classifier-free guidance (CFG) for natural images.

B.6. Failure Case

Our text rendering fails if the OCR algorithm fails. Typically, an OCR failure brings distortion and mis-rendering of text content. A visual example is shown in Fig. 4.

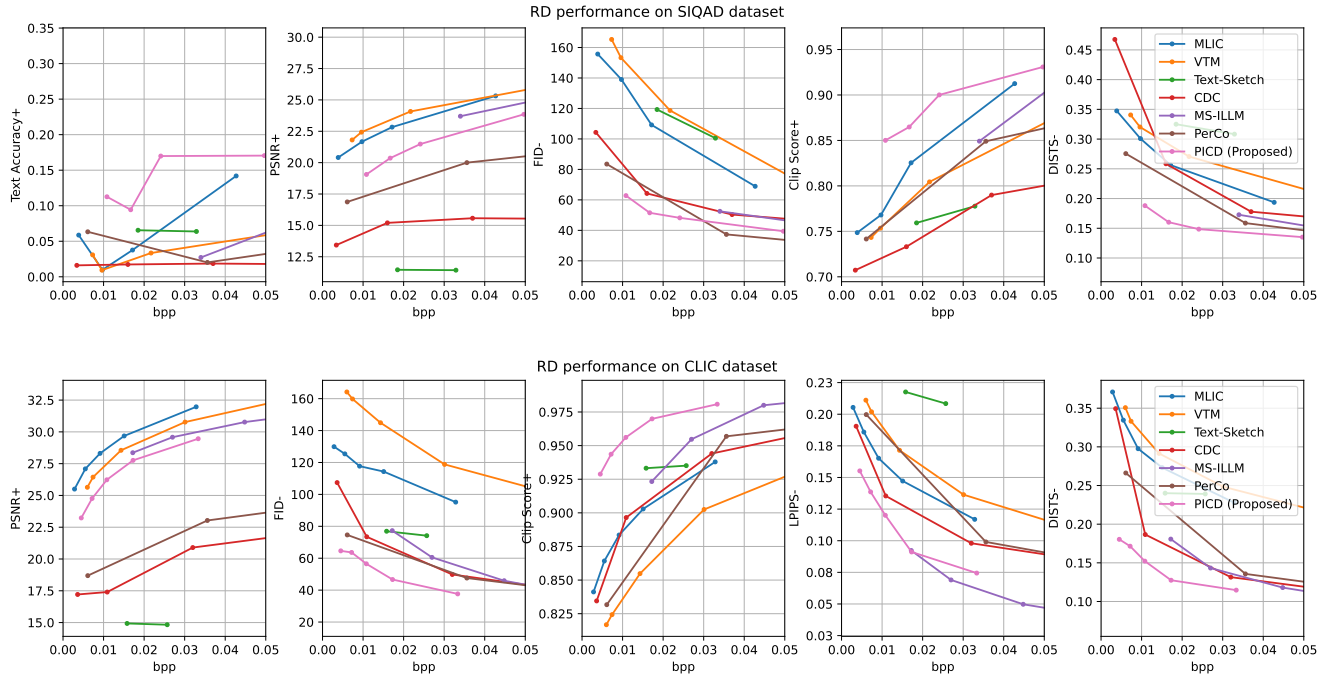

Figure 3. The rate distortion (RD) curve on screen and natural images.

Figure 4. An example of OCR failure.

References

[1] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, 2022. 1

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:3736–3764, 2021. 2

[3] Marlene Careil, Matthew Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. *ArXiv*, abs/2310.10325, 2023. 2

[4] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 1

[5] 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 31st ACM International Conference on Multimedia*, pages 7618–7627, 2023. 2

[6] Chih-Yu Lai, Dung N. Tran, and Kazuhito Koishida. Learned image compression with text quality enhancement. *ArXiv*, abs/2402.08643, 2024. 1

[7] Eric Lei, Yiugit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text + sketch: Image compression at ultra low rates. *ArXiv*, abs/2307.01944, 2023. 2

[8] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. 2023. 2

[9] Raymond W. Smith. An overview of the tesseract ocr engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2:629–633, 2007. 1

[10] Tong Tang, Ling X. Li, Xiao Wen Wu, Ruizhi Chen, Haochen Li, Guo Lu, and Limin Cheng. Tsa-scc: Text semantic-aware screen content coding with ultra low bitrate. *IEEE Transactions on Image Processing*, 31:2463–2477, 2022. 1

[11] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *ArXiv*, abs/2305.07015, 2023. 1

[12] Tongda Xu, Ziran Zhu, Dailan He, Yanghao Li, Lina Guo, Yuanyuan Wang, Zhe Wang, Hongwei Qin, Yan Wang, Jingjing Liu, and Ya-Qin Zhang. Idempotence and perceptual image compression. *ArXiv*, abs/2401.08920, 2024. 1

Figure 5. Qualitative results on screen images.

[13] Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *arXiv preprint arXiv:2209.06950*, 2023. 2

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023. 1

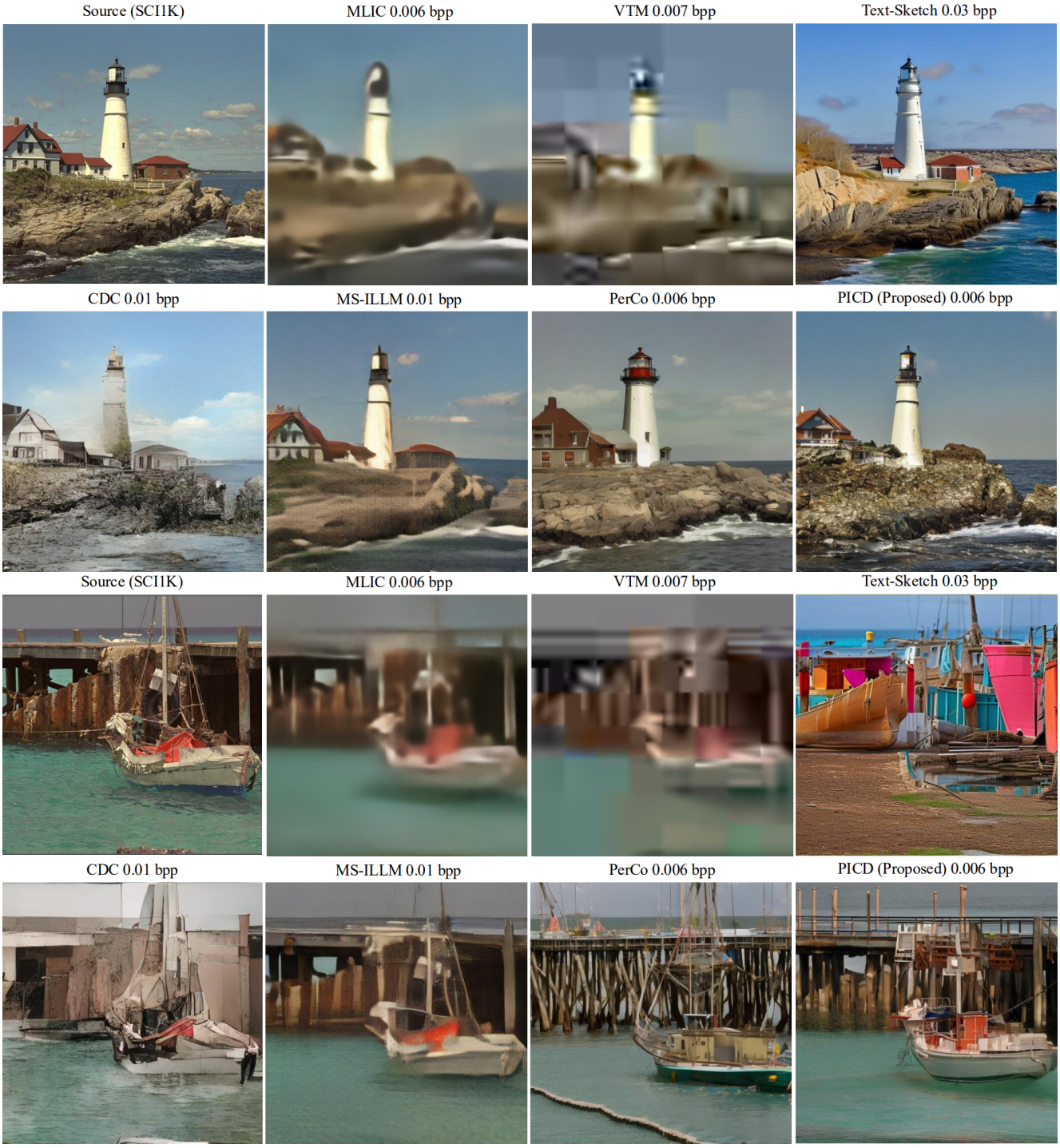

Figure 6. Qualitative results on natural images.