Single-step Diffusion for Image Compression at Ultra-Low Bitratess
Supplemental Material

1. Implementation Details

Our method adopts the SWIN-Transformer [1] U-Net architecture used in ResShift [2] for the base branch, and we train the entire model from scratch. For the quantization module, we use a pretrained VQGAN [3]. Specifically, for the two higher bitrate points, we extract features from layer f8 with codebook sizes of 64 and 256. For the two lower bitrate points, we use features from layer f16 with codebook sizes of 256 and 8192, respectively. Training is performed using the AdamW optimizer with a batch size of 8 and a learning rate of 5e-5, for a total of 60,000 iterations. All experiments are conducted on the NVIDIA TITAN RTX GPU.

Training procedure As illustrated in Algorithm 1, the training pipeline jointly optimizes the encoder, VQ module, residual fusion adapter, and diffusion-based decoder. The encoder produces latent features, which are vector-quantized with a learned codebook. These quantized latents are perturbed with Gaussian noise and refined through the residual fusion module before being denoised by the diffusion model. The reconstructed output is then decoded and optimized with a combination of reconstruction, perceptual, and VQ losses.

Inference procedure As shown in Algorithm 2, during inference, the input image is first encoded and quantized using the VQ module. The quantized latent is then perturbed by Gaussian noise and refined via residual fusion, followed by a single-step denoising process. The final quantized latent is decoded to reconstruct the image, enabling fast and high-quality decompression.

```
Algorithm 1: Training
Input: Image $X$
Output: Encoder $\mathcal{E}$, Diffusion model $D_\theta$, Adapter $A$, Codebook $\mathcal{C}$, Decoder $\mathcal{D}$
$x \leftarrow \mathcal{E}(X)$ {Encode input}
$y, \mathcal{L}_{\text{VQ}} \leftarrow \text{VQ}(x)$ {Vector quantize latent}
Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$
$\tilde{x} \leftarrow y + \eta_q \boldsymbol{\epsilon}$ {Forward process}
$z \leftarrow A(\tilde{x}, y)${Residual Fusion}
$\hat{x} \leftarrow D_\theta(z)$ {Reverse process}
$\hat{X} \leftarrow \mathcal{D}(\hat{x})$ {Decode quantized latent}
$\mathcal{L}_{\text{recon}} \leftarrow \|X - \hat{X}\|_2^2$
$\mathcal{L}_{\text{percept}} \leftarrow \mathcal{L}_{lpips}(X, \hat{X})$
$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{percept}} + \mathcal{L}_{\text{VQ}}$
Update model parameters to minimize $\mathcal{L}_{\text{total}}$
```

```
Algorithm 2: Inference
Input: Original image $X$
Output: Reconstructed image $\hat{X}$
$x \leftarrow \mathcal{E}(X)$ {Encode input}
$y \leftarrow \text{VQ}(x)$ {Quantize latent}
Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$
$\tilde{x} \leftarrow y + \eta_q \boldsymbol{\epsilon}$ {Forward process}
$z \leftarrow$A$(\tilde{x}, y)$ {Residual Fusion}
$\hat{x} \leftarrow D_\theta(z)$ {Reverse process}
$\hat{X} \leftarrow \mathcal{D}(\hat{x})$ {Decode quantized latent}
```

2. Additional Experiments

FPS-Rate comparison of image compression methods As shown as Fig. 1, we analyze the decoding speed (FPS) and rate–distortion trade-off (LPIPS BD-Rate) on the CLIC2020 dataset. Right–top indicates better performance. Our proposed single-step method (red star) achieves the best trade-off, delivering both significantly faster decoding and superior perceptual quality. Competing approaches require multiple diffusion steps and therefore suffer from lower decoding speed.

Additive analysis. We further analyze the effect of noise modulation and the number of diffusion steps on reconstruction quality. As shown in Figure 2, increasing the noise

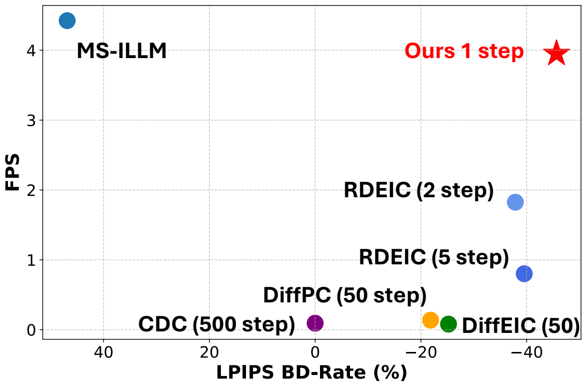


Figure 1. FPS-Rate comparison of image compression methods on CLIC2020. Right-top is better.

modulation level η consistently improves perceptual quality (lower LPIPS), with the best performance achieved at $\eta = 0.9$. Visual comparisons confirm that higher η values preserve perceptual details more faithfully, reducing oversmoothing and enhancing fidelity.

In addition, Figure 3 compares reconstructions under different numbers of diffusion steps. Without denoising (0 step), the quantized latent exhibits strong artifacts, reflected by low PSNR (16.5) and high LPIPS (0.50). A single denoising step significantly improves quality, reducing LPIPS to 0.25 and DISTS to 0.14, while also boosting PSNR to 21.7. However, applying excessive denoising (15 steps) oversmooths the image, leading to a higher LPIPS (0.29) despite slightly increased PSNR. These results highlight the effectiveness of our single-step diffusion approach in balancing fidelity and perceptual similarity.

Additive qualitative results. We provide qualitative comparisons on the Kodak and CLIC2020 datasets. As shown in Figure 4 and Figure 5, our method preserves sharper textures and clearer structural details compared to both autoencoder-based (ELIC) and diffusion-based (DiffEIC) codecs, while operating at lower bitrates and significantly higher decoding speed. For instance, fine structures such as roof tiles and wood grain are faithfully reconstructed by our approach, whereas competing methods produce either over-smoothed or distorted patterns.

In addition, Figure 6 presents qualitative results on the CLIC2020 dataset. Our single-step model achieves perceptually faithful reconstructions at ultra-low bitrates (e.g., 0.042 bpp), successfully maintaining subtle textures such as fabric details and skin tones. In contrast, ELIC exhibits noticeable blur, and DiffEIC often introduces artifacts. These results highlight the ability of our method to deliver high perceptual quality across diverse datasets, even under challenging compression settings.

References

[1] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 1

[2] Z. Yue, J. Wang, and C. C. Loy, “Resshift: Efficient diffusion model for image super-resolution by residual shifting,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 13 294–13 307, 2023. 1

[3] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883. 1

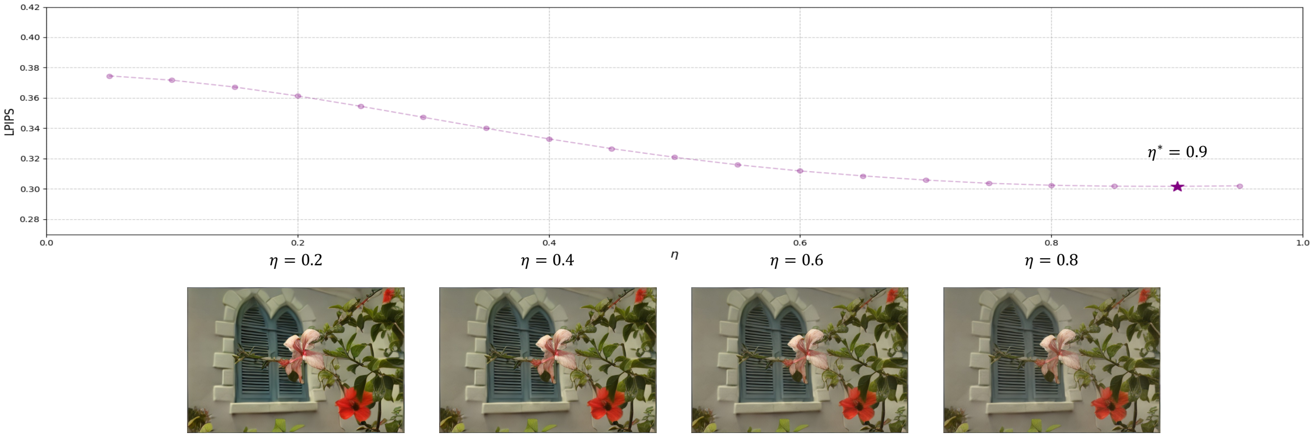
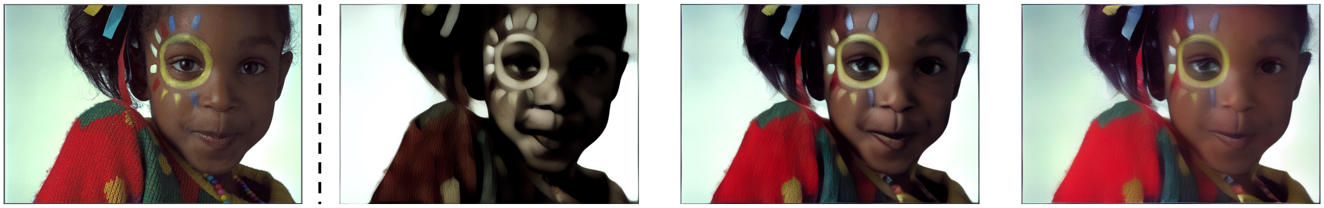


Figure 2. LPIPS score and visual comparisons under varying noise modulation levels η. The plot (top) shows that higher values of η yield better perceptual quality (lower LPIPS). The reconstructions (bottom) demonstrate that as η^* increases, perceptual details are more faithfully preserved, reducing oversmoothing and enhancing visual fidelity. The best performance is achieved at $\eta = 0.9$, as marked in the plot.



Original	0 step	1 step	15 step
PSNR	16.5	21.7	23.6
LPIPS	0.50	0.25	0.29
DISTS	0.28	0.14	0.17

Figure 3. 0 step: The reconstruction is directly decoded from the quantized latent without any refinement. It exhibits strong quantization artifacts, with low PSNR (16.5) and high LPIPS (0.50), indicating poor perceptual quality. 1 step: A single-step denoising significantly improves perceptual fidelity, reducing LPIPS to 0.25 and DISTS to 0.14, with a large PSNR gain to 21.7., 15 steps: Further denoising slightly improves PSNR (23.6) but increases LPIPS to 0.29, suggesting that excessive denoising may oversmooth fine textures and harm perceptual similarity.

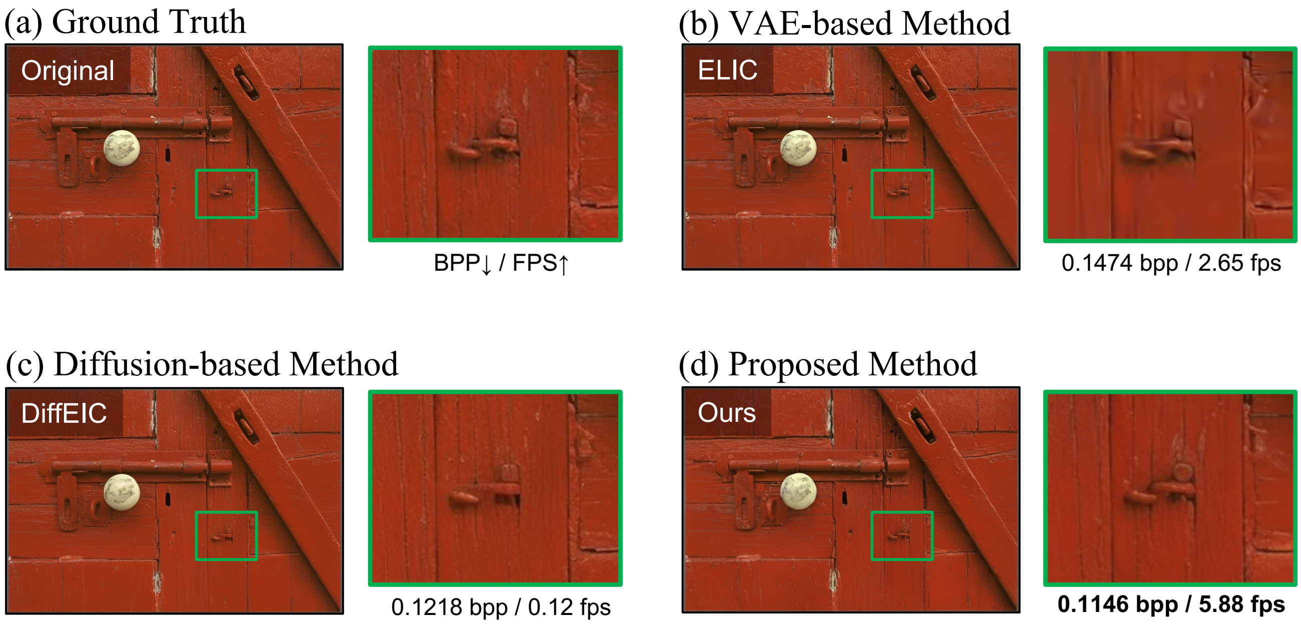


Figure 4. Qualitative examples on the Kodak dataset.

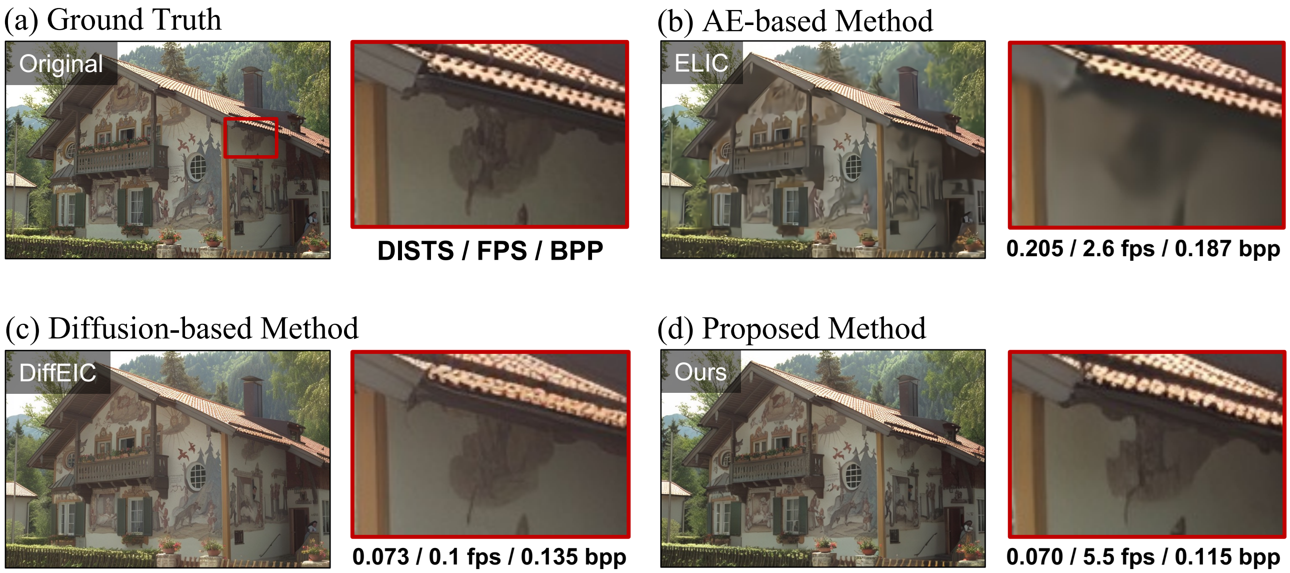

(a) Ground Truth
Original
DISTS / FPS / BPP
(b) AE-based Method
ELIC
0.205 / 2.6 fps / 0.187 bpp
(c) Diffusion-based Method
DiffEIC
0.073 / 0.1 fps / 0.135 bpp
(d) Proposed Method
Ours
0.070 / 5.5 fps / 0.115 bpp

Figure 5. Qualitative examples on the Kodak dataset.

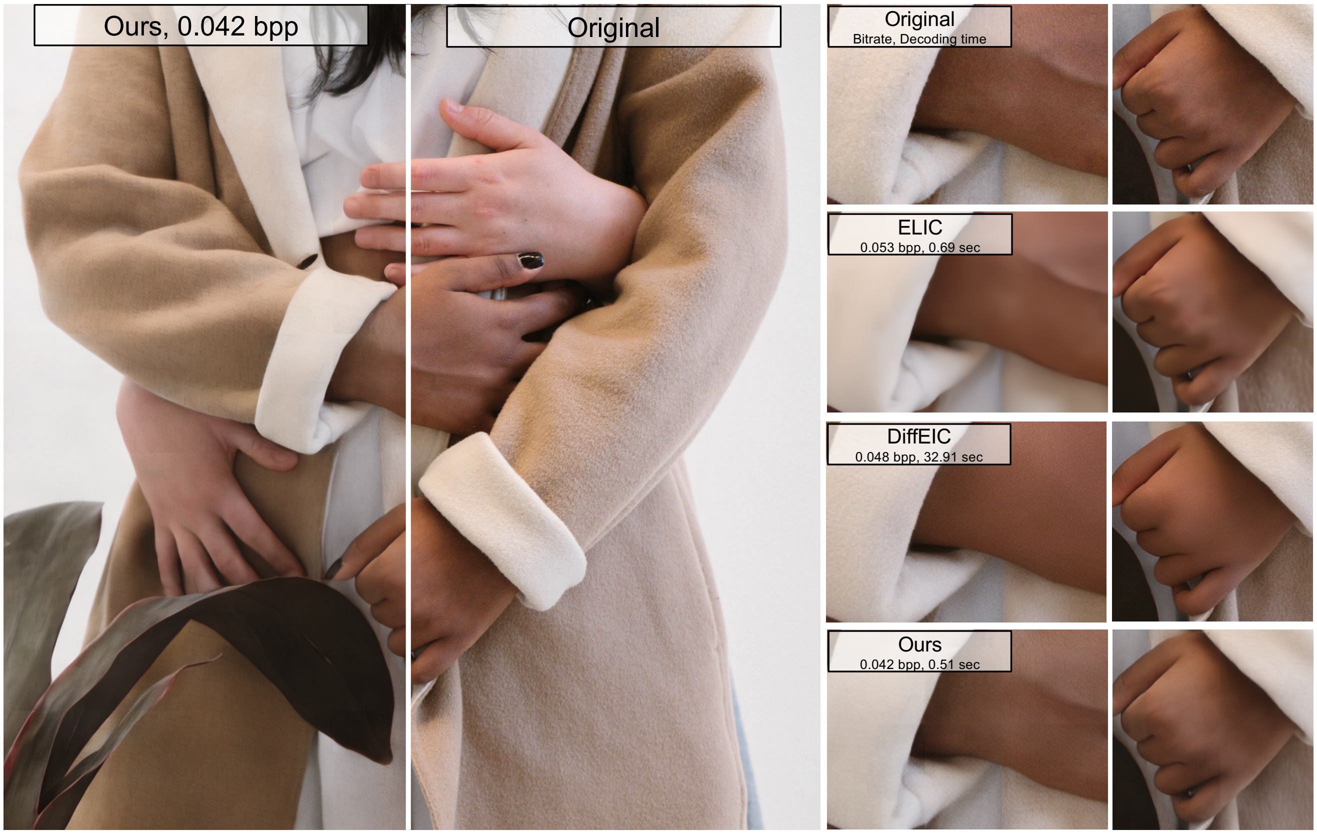


Figure 6. Qualitative examples on the CLIC2020 dataset.