# Ultra-Low Bitrate Perceptual Image Compression with Shallow Encoder

## Supplementary Material

This supplementary provides additional discussion on:

- Section A. Further ablation study and discussion.
- Section B. Additional performance comparison.
- Section C. Study of user preference for AEIC-SE.
- Section D. Network structures of AEIC models.
- Section E. Detailed training and inference procedures.
- Section F. Third-party models and evaluation methods.
- Section G. Potential future directions based on AEIC-SE.

## A. Further Ablation and Discussion

**Effect of High-Resolution Finetuning.** As a supplement to Fig. 8 and its accompanying discussion, Fig. 9 provides a visual comparison of 2K-resolution reconstructions from AEIC-SE trained with and without high-resolution finetuning (HRF). After applying HRF, AEIC-SE produces reconstructions with more faithful and visually coherent local textures while using fewer bits. For example, the stem and contour of the berry become sharper and more consistent with the original content, whereas the water textures appear more realistic and better aligned with natural patterns.

**Lightweight Encoder for Extreme Bitrate.** We further investigate whether lightweight encoders can be applied to StableCodec [31], one of the latest ultra-low bitrate image compression methods. StableCodec employs a complex multi-stage encoder that includes the Stable Diffusion VAE encoder, the ELIC encoder [11], and a latent-space transform encoder to produce a 64× downsampled latent. As shown in Fig. 3 (a), these components result in 47.16M encoder parameters in total. To examine the encoder complexity, we replace StableCodec's encoders with our moderate encoder (3.09M parameters) used in AEIC-ME, while keeping all other modules and training strategies unchanged. We test two variants with spatial compression ratios of 32 and 64, denoted as StableCodec-ME (32×) and StableCodec-ME (64×). As shown in Fig. 10, StableCodec-ME (64×) achieves performance comparable to the original StableCodec, whereas StableCodec-ME (32×) even surpasses the baseline on all four metrics. These results support our finding that ultra-low bitrate compression does not require a large or expressive encoder, since the latent information is fundamentally constrained by the bitrate budget.

**Decoder Architectural Pruning.** We next provide a detailed analysis of the decoder architecture, specifically the unconditional denoiser $\epsilon_{\text{SD}}$ and lite VAE decoder $\mathcal{D}_{\text{SD}}$ introduced in Section 3.1.2. We begin by constructing a base AEIC-ME model using the original conditional denoiser $\epsilon_{\text{SD}}$ and VAE decoder $\mathcal{D}_{\text{SD}}$ from SD-Turbo. Following [6], we remove the text encoder, timestep embeddings, and all cross-attention layers from $\epsilon_{\text{SD}}$, since textual conditions contribute negligibly reconstruction quality in image compression [22], and the timestep input degenerates to a constant in one-step denoising. As shown in Table 6 (Variant 1), this pruning removes over 75M parameters and converts $\epsilon_{\text{SD}}$ into an unconditional denoiser (from Eq. 6 to Eq. 4), while also slightly improving overall performance.

StableCodec (Table 6) [31] indicates that decoding latency is dominated by $\mathcal{D}_{\text{SD}}$. To further improve decoding efficiency, we replace the original VAE decoder with a lite version [6] that prunes 50% of its channels (Variant 2). Table 6 shows that this reduces parameters from 49.5M to 12.4M, while incurring less than 1% performance degradation relative to Variant 1. This efficiency-performance balance is reasonable because ultra-low bitrate compression (below 0.05 bpp) inherently cannot fully exploit the repre-

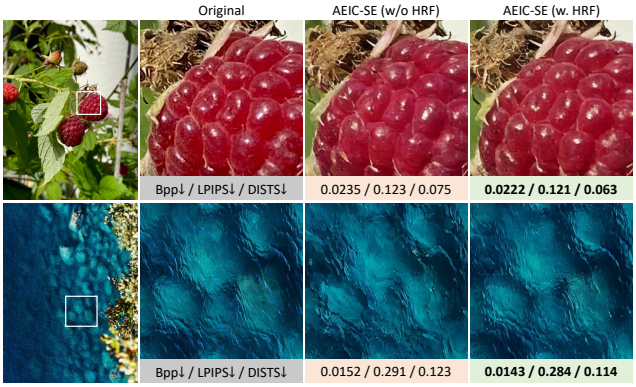



Figure 9. Qualitative comparison on AEIC-SE models trained with or without high-resolution finetuning (HRF), using 2K resolution images from the CLIC 2020 test set. Best viewed on screen.

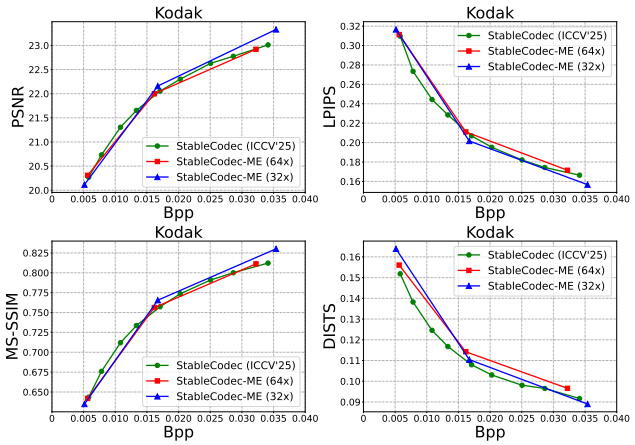



Figure 10. StableCodec [31] performance on Kodak when replacing original encoders with our moderate encoders (ME) of different spatial compression ratios (abbreviated as StableCodec-ME).

Table 6. Ablation study on the decoder architecture pruning. We first construct a base AEIC-ME model with the original conditional denoiser $\epsilon_{\text{SD}}$ and VAE decoder $\mathcal{D}_{\text{SD}}$ in SD-Turbo. Then, we construct "Variant 1" by removing text encoders, timestep embeddings, and all cross-attention layers from $\epsilon_{\text{SD}}$, transforming $\epsilon_{\text{SD}}$ into an unconditional denoiser. In "Variant 2", we further replace the original $\mathcal{D}_{\text{SD}}$ with a lite version [6] using only 50% channels.

| Model | Params. (M) | | BD-rate (↓%) on Kodak | | | |
|---|---|---|---|---|---|---|
| | $\epsilon_{\text{SD}}$ | $\mathcal{D}_{\text{SD}}$ | PSNR | MS-SSIM | LPIPS | DISTS |
| Base | 865.9 | 49.5 | 0 | 0 | 0 | 0 |
| Variant 1 | **790.6** | 49.5 | **-1.29** | **-1.39** | **-0.23** | **-0.39** |
| Variant 2 | **790.6** | **12.4** | -0.37 | -0.46 | +0.38 | +0.60 |

Table 7. Reconstruction quality of different methods on Kodak. SD VAE stands for the VAE used in SD-Turbo and SD 2.1.

| Method | PSNR↑ | MS-SSIM↑ | LPIPS↓ | DISTS↓ |
|---|---|---|---|---|
| SD VAE | **26.65** | **0.932** | **0.073** | **0.041** |
| SD VAE (w. lite $\mathcal{D}_{\text{SD}}$) | 26.56 | 0.930 | 0.079 | 0.046 |
| AEIC-ME (0.038 bpp) | 23.30 | 0.832 | 0.143 | 0.082 |

sentational capacity of the original VAE decoder. Table 7 further compares reconstruction performance across methods, indicating that even the highest bitrate setting of AEIC-ME produces reconstructions substantially worse than the SD VAE itself. Therefore, a lite decoder is sufficient for maintaining quality while enabling faster decoding.

**Selection of Perceptual Loss.** Table 8 reports the impact of different perceptual losses when finetuning AEIC-ME under ultra-low bitrates. Unlike commonly adopted LPIPS, which measures latent-level distortion using VGG features, we employ DISTS [8], which imposes statistical constraints and provides more effective supervision for texture fidelity under extreme bitrates. In practice, we adopt the overlap-chunked edge-aware DISTS (OC-EA-DISTS) [16, 25], a recent variant tailored for different patch sizes and designed to jointly evaluate structure and texture similarity. As shown in Table 8, using OC-EA-DISTS sacrifices distortion fidelity slightly but leads to improved perceptual quality, which is more critical in ultra-low bitrate scenarios.

## B. Additional Performance Comparison

**Rate-Distortion-Perception Comparison on Kodak.** Fig. 12 shows the rate-perception and rate-distortion comparisons on the Kodak dataset [7]. Since Kodak contains only 24 images at a resolution of 768×512, we follow prior works [13, 26, 31] and omit FID [12] and KID [3] due to their unreliability on small datasets. We compare AEIC with the traditional codec H.266/VVC [4] using VTM-23.13 intra mode, a distortion-oriented neural codec ELIC [11], and several state-of-the-art perceptual-oriented ultra-low bitrate methods including MS-ILLM [21], GLC [13], PerCo [5], DiffEIC [18], DLF [26], and StableCodec [31]. Both AEIC-ME and AEIC-SE achieve the best perceptual performance (e.g., LPIPS and DISTS) across all bitrates. In terms of distortion, AEIC-ME and AEIC-SE remain competitive among advanced perceptual-oriented codec, while significantly outperforming them in perception.

Table 8. Ablation study on the perceptual loss $\mathcal{L}_p$. We train AEIC-ME models using similar strategies as described in Section 3.3, only vary the selection of $\mathcal{L}_p$ in Stage 2 between LPIPS [30] and overlap-chunked edge-aware DISTS (OC-EA-DISTS) [16, 25].

| Model | $\mathcal{L}_p$ Selection | BD-rate (↓%) on Kodak | | | |
|---|---|---|---|---|---|
| | | PSNR | MS-SSIM | LPIPS | DISTS |
| AEIC-ME | LPIPS | **0** | **0** | 0 | 0 |
| | OC-EA-DISTS | +9.90 | +7.14 | **-5.19** | **-20.35** |

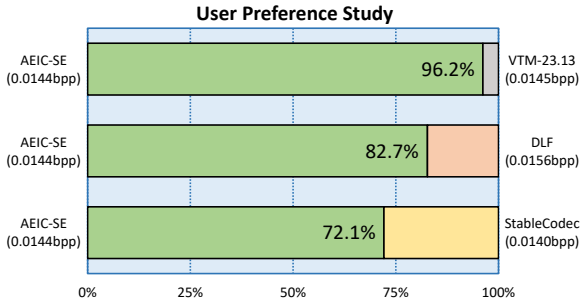



Figure 11. User preference study on Kodak comparing AEIC-SE against traditional codec H.266/VVC [4] and advanced learning-based generative codec DLF [26] and StableCodec [31].

**Additional Visual Comparison.** Fig. 15 presents additional qualitative results on 512×512 patches from the Kodak dataset. We compare AEIC-SE with H.266/VVC (VTM-23.13 intra), as well as strong ultra-low bitrate baselines DLF and StableCodec. AEIC-SE consistently reconstructs more visually coherent structures and textures using fewer bits. Figs. 16-21 further provide comparisons on 2K-resolution images from the CLIC 2020 test set and DIV2K validation set. Across all resolutions and content types, AEIC-SE delivers the most visually consistent results while operating at the lowest bitrate, reinforcing its superior capability for ultra-low bitrate perceptual compression.

## C. User Study

We conducted a user preference study based on side-by-side visual comparisons. In each case, we display the ground-truth image and two reconstructions at similar ultra-low bitrates: one produced by AEIC-SE and the other produced by a competitor (H.266/VVC, DLF and StableCodec). The left-right order of the two reconstruction methods was randomized to prevent positional bias. We invited 15 users. Each participant evaluated 24 cases. Fig. 11 shows a clear preference, where AEIC-SE received 96.2% of the votes against H.266/VVC, 82.7% against DLF and 72.1% against StableCodec, indicating consistently better visual quality.

## D. Model Structure

The overall structure of AEIC models are detailed in Fig. 13 and Fig. 14. Our codec consists of an analysis transform $g_a$, a synthesis transform $g_s$, and an entropy model. We follow [17, 31] to construct our entropy model with a pair of hyper transform [1] and a 4-step quadtree-partitioned autoregressive context model. The major networks and hidden dimensions are detailed in Fig. 14, exploiting efficient convolution blocks [19, 27, 28]. The synthesis transform $g_s$ produces two latents, $l_T$ and $l_{res}$, following the dual-branch decoding format [31]. Note that AEIC-ME and AEIC-SE only differ in the analysis transform $g_a$ and the entropy model as detailed in Fig. 14 and summarized in Table 1. Regarding the one-step diffusion, we set the LoRA rank in the unconditional Unet denoiser $\epsilon_{\text{SD}}$ to 32, while keeping the pretrained VAE decoder $\mathcal{D}_{\text{SD}}$ [6] frozen throughout AEIC training.

## E. Training and Inference Details

**AEIC-ME training.** Stage 1 for AEIC-ME takes over 300K iterations, using 512×512 patches and a batch size of 8. On 2× RTX 3090 GPUs (24GB memory), this process requires 4 gradient accumulation steps and an actual batch size of 1 for each GPU. The learning rate degrades from $1e^{-4}$ to $5e^{-5}$ after 280K iterations. $\{\lambda_{\text{S1}}, \gamma_1, \gamma_2, \gamma_3\}$ are set to $\{1, 2, 1, 0.1\}$, respectively. Stage 2 takes over 30K iterations, increasing the batch size to 32. The learning rate starts from $5e^{-5}$, then degrades to $\{2e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}\}$ at $\{10, 25, 28, 29\}$K iterations. $\lambda_{\text{S2}}$ chooses from $\{2, 4, 8, 16, 32\}$ for different ultra-low bitrates. $\alpha$ is set to 0.1. The total training for AEIC-ME requires approximately 9 days on 2× RTX 3090 GPUs.

**AEIC-SE training.** Stage 1 for AEIC-SE takes over 200K iterations, using 512×512 patches and a batch size of 8. This process requires 2 gradient accumulation steps and an actual batch size of 2 for each GPU. The learning rate degrades from $1e^{-4}$ to $5e^{-5}$ after 180K iterations. $\{\lambda_{\text{S1}}, \gamma_1, \gamma_2, \gamma_3, \beta_1\}$ are set to $\{1, 2, 1, 0.1, 0.5\}$, respectively. After 180K iterations, we drop $\mathcal{L}_{enc}$ and reset $\{\lambda_{\text{S1}}, \gamma_1\}$ to $\{1.1, 0.5\}$ for fast convergence. Stage 2 takes over 30K iterations, using 512×512 patches and a batch size of 32. The actual batch size and gradient accumulation step for each GPU are set to 1 and 16. The learning rate starts from $5e^{-5}$, then degrades to $\{2e^{-5}, 1e^{-5}, 5e^{-6}, 1e^{-6}\}$ at $\{10, 25, 28, 29\}$K iterations. $\lambda_{\text{S2}}$ chooses from $\{2, 4, 8, 16, 32\}$ for different ultra-low bitrates. $\{\gamma_1, \gamma_2, \gamma_3, \alpha, \beta_2\}$ are set to $\{0.5, 1, 0.05, 0.05, 0.001\}$, respectively. After 20K iterations, we drop $\mathcal{L}_{dec}$ for fast convergence. Stage 3 for AEIC-SE takes over 5K iterations, using 1024×1024 patches and a batch size of 8. The actual batch size and gradient accumulation step for each GPU are set to 1 and 4. Gradient checkpointing is activated. The learning rate starts from $2e^{-5}$, then degrades to $\{1e^{-5}, 5e^{-6}, 1e^{-6}\}$ at $\{3, 4.5, 4.8\}$K iterations. $\{\lambda_{\text{S3}}, \gamma_1, \gamma_2, \gamma_3, \alpha\}$ remains the same as Stage 2. The total training for AEIC-SE also requires about 9 days.

**Inference Strategy.** We use similar tiling and color fix strategies [23, 29, 31] for high-resolution images. Specifically, for AEIC-ME we set Unet tile size to 96 with an overlap of 32, and the VAE decoder tile size to 160. Since AEIC-SE has been finetuned on 1024×1024 patches, we set Unet tile size to 192 with an overlap of 64. 16-bit color fix [31] is employed when using tiling strategies for inference.

## F. Third-Party Models and Evaluation

**Ultra-low Bitrate Image Codec.** We evaluate GLC [13], DiffEIC [18], ResULIC [14], OSCAR [9], DLF [26] and StableCodec [31] using the official code and pretrained weights. We finetune MS-ILLM [21] using the official code and the pretrained weight (at the lowest available bitrate) to reach ultra-low bitrates. For PerCo [5], we rely on a community implementation [15] and the pretrained weights as the official code is not available.

**Distortion-Oriented Neural Image Codec.** We evaluate EVC-Small [10] by the official code and pretrained weights. For ELIC [11], we follow the implementation in CompressAI [2], and train models for ultra-low bitrates.

**Traditional Codec.** VTM-23.13 is the reference software for H.266/VVC [4]. We install the software according to the official instructions. For RGB images, we manage RGB-YUV420 transformation using FFmpeg following [17].

**Implementation of Evaluation Metrics.** We construct PSNR, MS-SSIM [24] and DISTS [8] metrics using PyIQA with default settings, while implement LPIPS [30], FID [12] and KID [3] metrics using TorchMetrics. FID and KID are evaluated by splitting images into overlapped 256×256 patches following the protocol in [21].

## G. Future Work

In this work, we primarily focus on the feasibility of applying shallow encoder for source-limited ultra-low bitrate image compression senders. While the proposed AEIC-SE demonstrates strong perceptual quality, real-time practical encoding efficiency, and competitive decoding speed at ultra-low bitrates, a key challenge lies in further reducing the decoding latency, since achieving truly real-time decoding at extreme bitrates remains difficult due to the computational overhead of generative reconstruction. Future research may investigate more compact generative priors, hardware-friendly decoder designs, and novel decoder pruning mechanisms that preserve perceptual fidelity while significantly lowering computational costs. We hope these directions will inspire continued advancement toward efficient, deployable, and perceptually optimized ultra-low bitrate image compression systems.

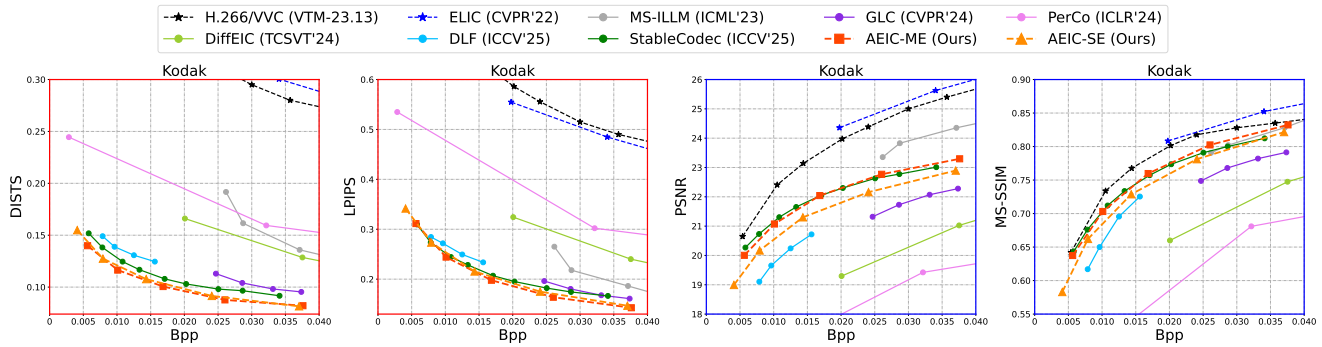



Figure 12. Rate-perception (sub-figures in red borders) and rate-distortion (sub-figures in blue borders) comparison of advanced image compression methods on the Kodak [7] dataset. Note that FID and KID are not reported on Kodak due to its small size (24 images).

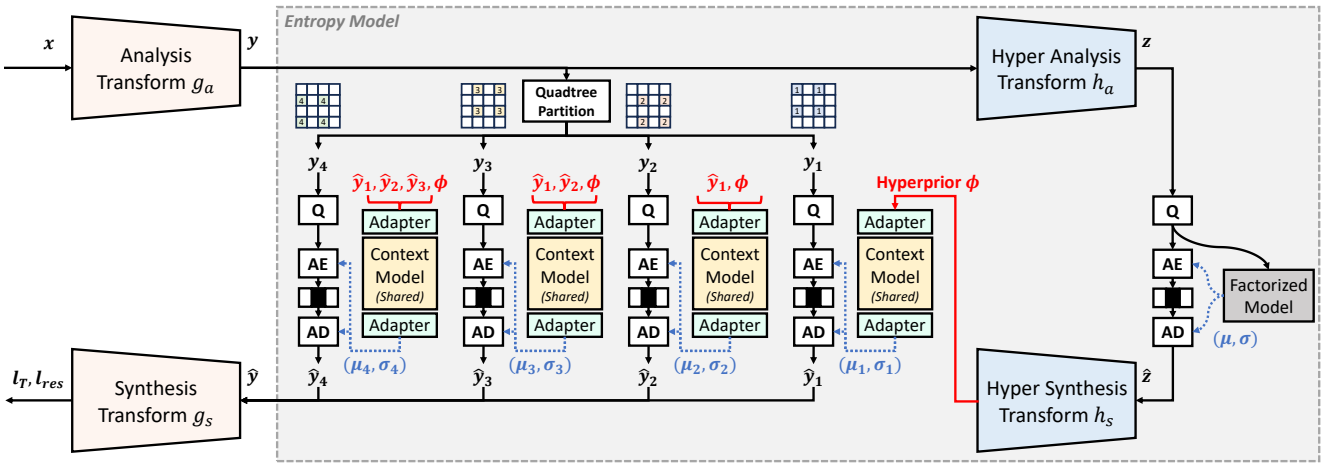
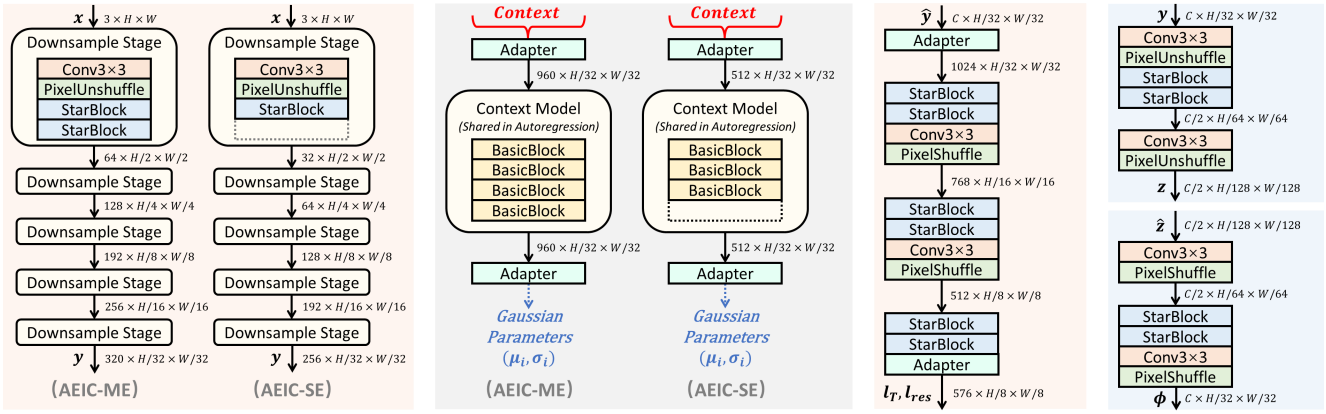



Figure 13. Network structure of AEIC models. We follow [20] to build our entropy model with a hyperprior [1] and an efficient 4-step autoregressive context model, exploiting quadtree partition [17] and expressive networks [31]. Our moderate encoder variant (AEIC-ME) and shallow encoder variant (AEIC-SE) mainly differ in the analysis transform $g_a$ and the entropy model, as illustrated in Fig. 14.




(a) Analysis Transform $g_a$ (b) Autoregressive Context Model with Adapters (c) Synthesis Transform $g_s$ (d) Hyper Transform $h_a$, $h_s$

Figure 14. Modules in AEIC models. Specifically, our transforms ($g_a$, $g_s$, $h_a$ and $h_a$) employ StarBlock [19], an efficient convolution network based on element-wise multiplication. The context model follows a similar implementation of StableCodec [31] with shared BasicBlocks (consists of InceptionNeXt [28] and GatedCNN [27]) and independent Adapters (a single resblock to adjust channel dimensions).

## References


[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 3, 4

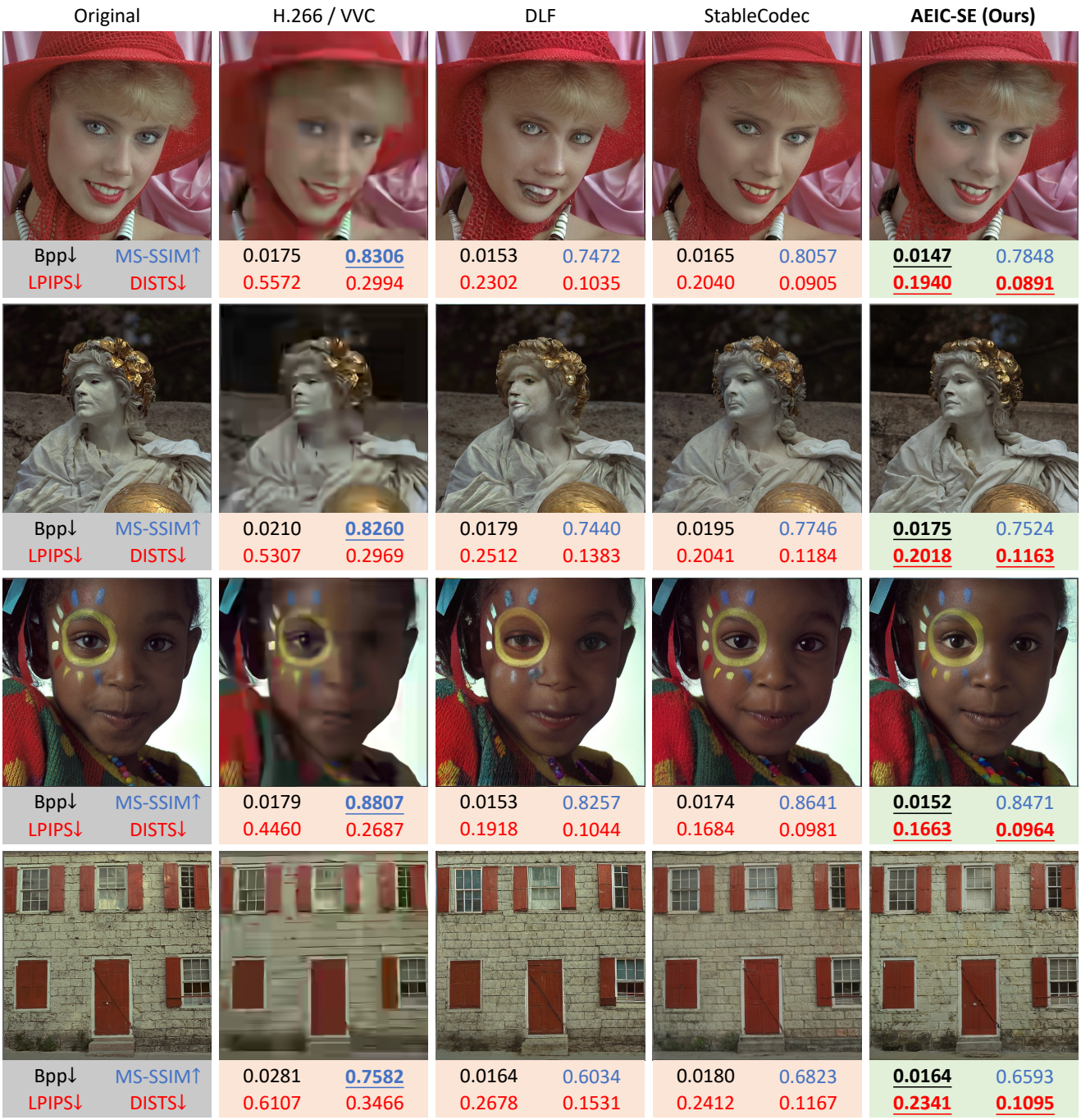



Figure 15. Qualitative comparison (512×512 patches) on the Kodak dataset. Distortion is evaluated with MS-SSIM, while perceptual quality is assessed using LPIPS and DISTS. The best results are highlighted in **bold and underlined**. AEIC-SE achieves superior perceptual reconstruction with the fewest bits. Although H.266/VVC attains the highest MS-SSIM scores, its outputs exhibit blurriness and blocking artifacts, indicating that distortion metrics like MS-SSIM become less reliable at ultra-low bitrates.


[2] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020. 3

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 2, 3

[4] 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.


Figure 16. Visual comparison (2K resolution) on the DIV2K validation set.

*IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 2, 3

[5] Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3

[6] Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28208–28220, 2025. 1, 2, 3

[7] Eastman Kodak Company. Kodak image database, 1993. Accessed: 2024-08-27. 2, 4

[8] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 2, 3

[9] Jinpei Guo, Yifei Ji, Zheng Chen, Kai Liu, Min Liu, Wang Rao, Wenbo Li, Yong Guo, and Yulun Zhang. Oscar: One-step diffusion codec across multiple bit-rates. In *The Thirty-*

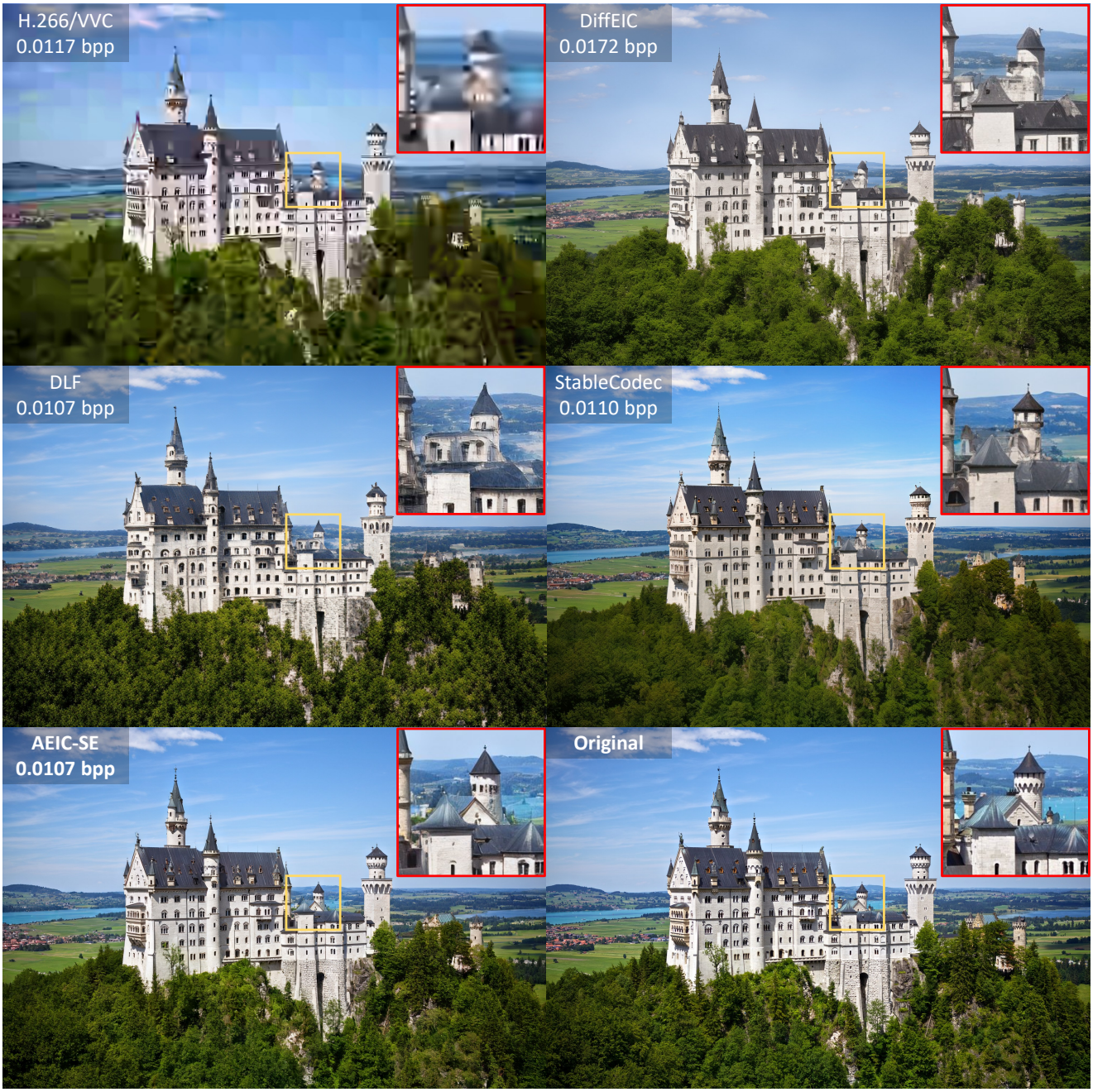



Figure 17. Visual comparison (2K resolution) on the DIV2K validation set.

*ninth Annual Conference on Neural Information Processing Systems*. 3

[10] Wang Guo-Hua, Jiahao Li, Bin Li, and Yan Lu. Evc: Towards real-time neural image compression with mask decay. In *The Eleventh International Conference on Learning Representations*. 3

[11] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5718–5727, 2022. 1, 2, 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2, 3

[13] Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *Proceedings of the IEEE/CVF Conference on*

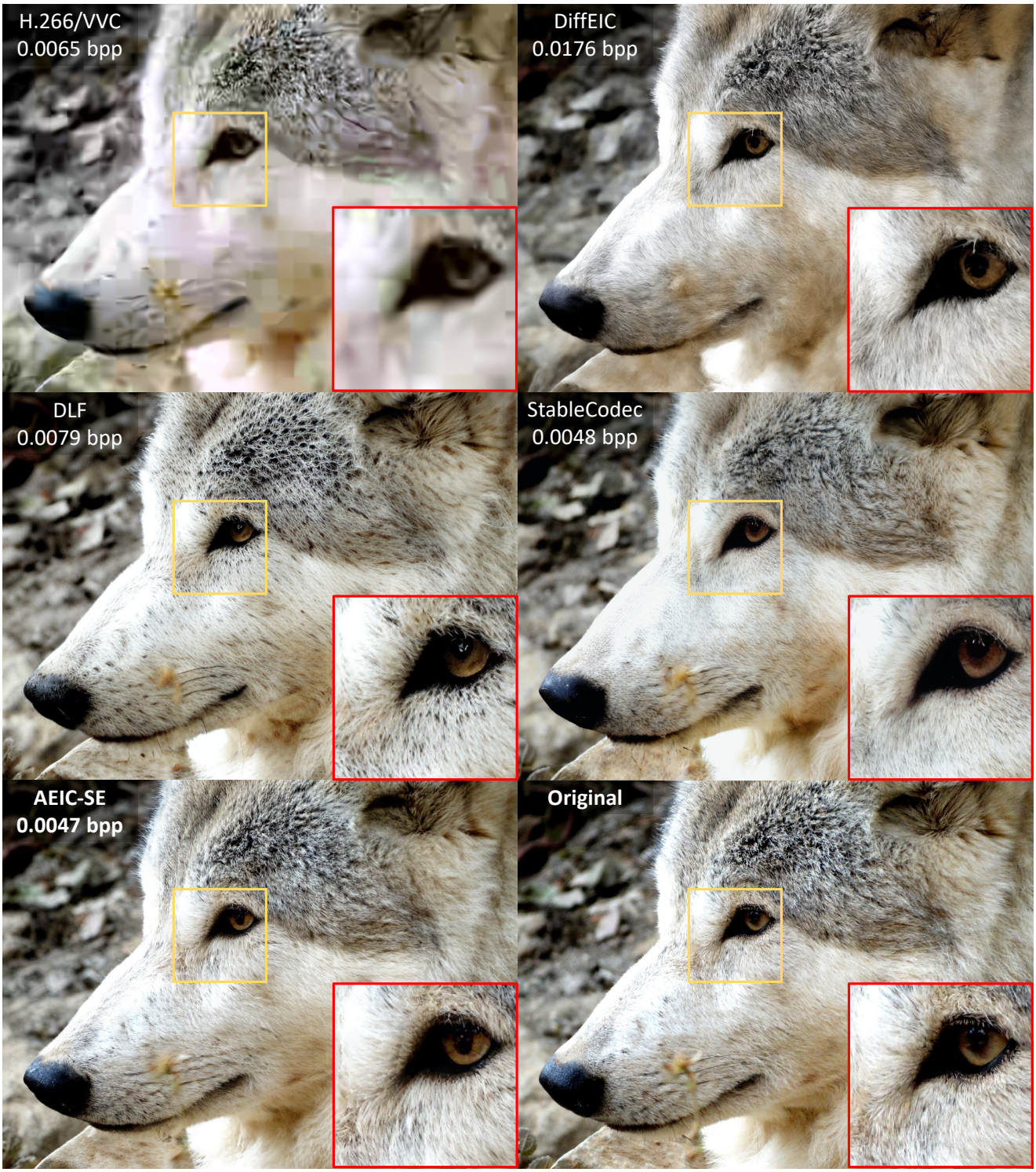



Figure 18. Visual comparison (2K resolution) on the DIV2K validation set.

*Computer Vision and Pattern Recognition*, pages 26088–26098, 2024. 2, 3

[14] Anle Ke, Xu Zhang, Tong Chen, Ming Lu, Chao Zhou, Jiawen Gu, and Zhan Ma. Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. In *Forty-second International Conference on Machine Learning*. 3

[15] Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha


Figure 19. Visual comparison (2K resolution) on the CLIC 2020 test set.

Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. Perco (sd): Open perceptual compression. *arXiv preprint arXiv:2409.20255*, 2024. 3

[16] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Unleashing the power of one-step diffusion based image super-resolution via a large-scale diffusion discriminator. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2

[17] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22616–22626, 2023. 3, 4

[18] Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Towards extreme image compression with latent feature guidance and diffusion prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 3

[19] Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, and Yun Fu. Rewrite the stars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5694–5703, 2024. 3, 4

[20] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018. 4

[21] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich,


Figure 20. Visual comparison (2K resolution) on the CLIC 2020 test set.

Hervé Jégou, and Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. In *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023. 2, 3

[22] Jeremy Vonderfecht and Feng Liu. Lossy compression with pretrained diffusion models. In *The Thirteenth International Conference on Learning Representations*. 1

[23] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 3

[24] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. Ieee, 2003. 3

[25] Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution. *arXiv preprint arXiv:2508.08227*, 2025. 2

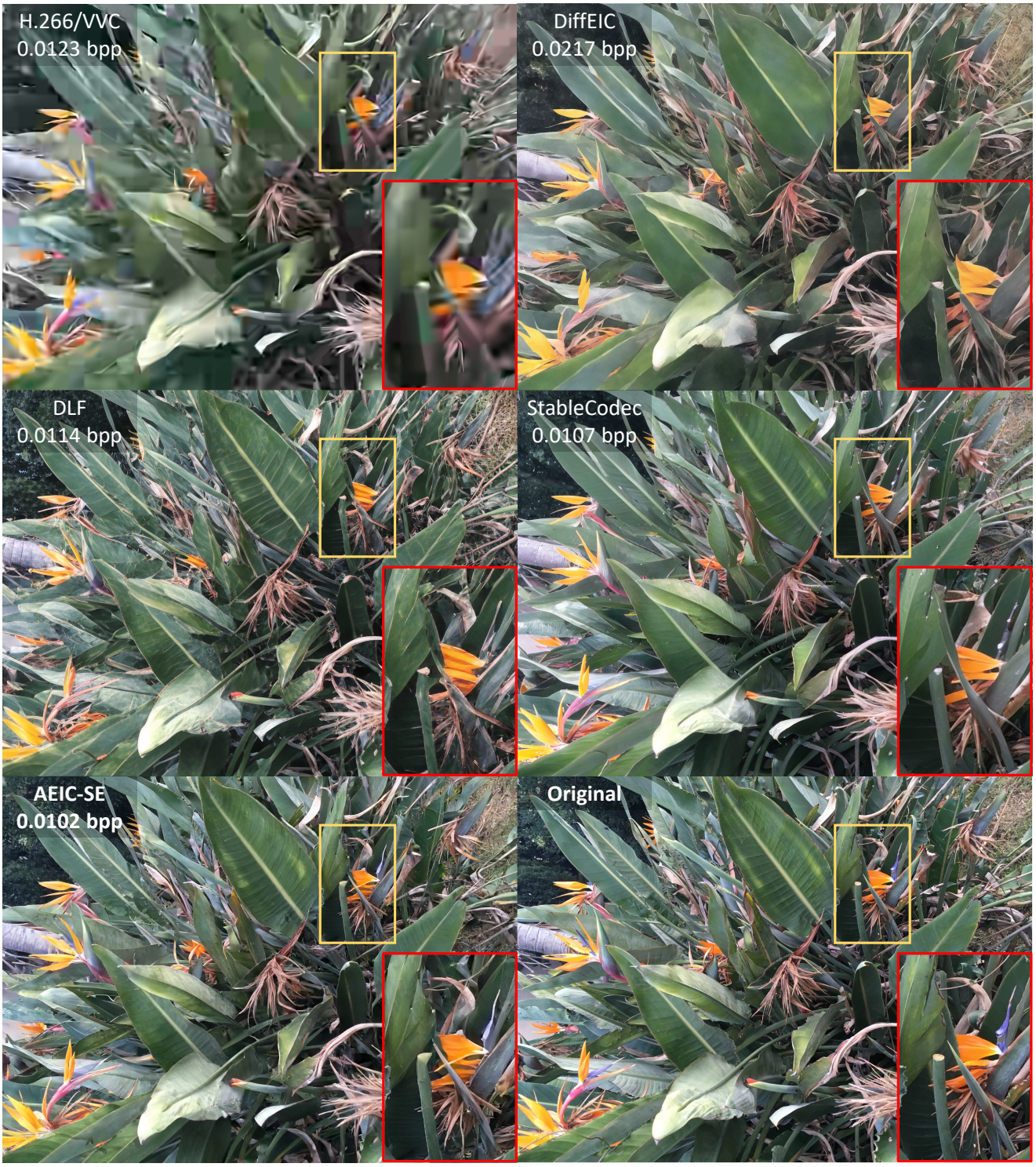



Figure 21. Visual comparison (2K resolution) on the CLIC 2020 test set.

[26] Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. Dlf: Extreme image compression with dual-generative latent fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19227–19236, 2025. 2, 3

[27] Weihao Yu and Xinchao Wang. Mambaout: Do we really need mamba for vision? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

pages 4484–4496, 2025. 3, 4

[28] Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. Inceptionnext: When inception meets convnext. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5672–5683, 2024. 3, 4

[29] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024. 3

[30] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 3

[31] Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. Stablecodec: Taming one-step diffusion for extreme image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17379–17389, 2025. 1, 2, 3, 4