CADC: Content Adaptive Diffusion-Based Generative Image Compression

Supplementary Material

8. Implementation Details

8.1. Network Structure

The detailed architectures of the main modules are illustrated in Fig. 8, including the main analysis transform g_a, main synthesis transform g_s, hyper analysis transform h_a, hyper synthesis transform h_s, auxiliary decoder g_{aux}, uncertainty estimation network f_u, and the context model. For efficient network construction, we primarily rely on modified versions of InceptionNeXt [62], GatedCNN [61], and their combination (StarBlock).

We also illustrate the overall entropy modeling process in Fig. 9. Our entropy model integrates a hyperprior module with an autoregressive context model. The hyperprior representation $\mathbf{c_h}$ is derived as:

$$\mathbf{z} = h_a(\mathbf{y}), \quad \hat{\mathbf{z}} = \lfloor \mathbf{z} \rceil, \quad \mathbf{c_h} = h_s(\hat{\mathbf{z}}), \tag{14}$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer. Here, $\mathbf{y}$ has 320 channels with a spatial downsampling factor of 64, while $\mathbf{z}$ and $\hat{\mathbf{z}}$ also have 320 channels but with a higher spatial compression ratio of 256×. To balance coding performance and computational efficiency, we employ a 4-step autoregressive process based on quadtree partitioning [35]. As shown in Fig. 9, this process estimates the Gaussian parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for the quantized latent $\hat{\mathbf{y}}$. Subsequently, arithmetic coding is applied to encode or decode $\hat{\mathbf{y}}$ into/from the bitstream.

For BLIP, we use the lightweight *blip-image-captioning-base* model (247.41M parameters, with inputs resized to 384×384 by default).

To enable a fair evaluation of our three proposed methods, we construct a variant StableCodec by removing its original VAE encoder and auxiliary encoder, and replacing the main analysis/synthesis transforms, hyper analysis/synthesis transforms, and context model with our implementations. This reduces computational complexity and eliminates architectural discrepancies that could otherwise confound the assessment of our contributions. We refer to this variant as our baseline M_0.

8.2. Training Process

We train our codec based on the standard rate-distortion loss:

$$\mathcal{L} = \lambda \mathcal{R} + \mathcal{D}, \tag{15}$$

where $\mathcal{R}$ denotes the bitrate, $\mathcal{D}$ is the distortion measure, and λ is a Lagrange multiplier that balances the two terms. We employ a two-stage training strategy [66]. In the first stage, a base model is trained using a relatively small λ_{base}. In the second stage, this pre-trained model is fine-tuned with a larger λ_{target} to achieve ultra-low target bitrates. The distortion term $\mathcal{D}$ incorporates multiple components: MSE, LPIPS (computed with VGG features), a CLIP-based loss $\mathcal{L}_{\text{CLIP}}$ [66]—defined as the ℓ_2-distance between CLIP embeddings of $\mathbf{x}$ and $\hat{\mathbf{x}}$, an adversarial loss $\mathcal{L}_{\text{adv}}$, and our proposed auxiliary reconstruction loss $\mathcal{L}_{\text{aux}}$. We use DINOv2 [46] with registers [14] as the discriminator backbone [31]. Note that $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{aux}}$ are only activated in the second training stage. The complete objective is formulated as follows:

$$\begin{aligned}
&\text{Stage I}: \arg\min_{\theta} L_1 = \lambda_{\text{base}}\, \mathcal{R} + \mathcal{D}_1, \\
&\text{Stage II}: \arg\min_{\theta} L_2 = \lambda_{\text{target}}\, \mathcal{R} + \mathcal{D}_2, \\
&\quad \mathcal{D}_1 = d_1\, \text{MSE}(x, \hat{x}) + d_2\, \text{LPIPS}(x, \hat{x}) + d_3 \mathcal{L}_{CLIP}(x, \hat{x}) \\
&\quad \mathcal{D}_2 = d_1\, \text{MSE}(x, \hat{x}) + d_2\, \text{LPIPS}(x, \hat{x}) + d_3 \mathcal{L}_{CLIP}(x, \hat{x}) \\
&\qquad + d_4 \mathcal{L}_{aux}(x, \hat{x}_{aux}) + d_5 \mathcal{L}_{adv},
\end{aligned} \tag{16}$$

where θ denotes all trainable parameters in the codec, and d_1–d_5 are weighting coefficients that balance the distortion terms.

9. More Experimental Results

9.1. More Quantitative Results

A comprehensive evaluation of rate-distortion performance is conducted on Kodak, the validation set of DIV2K [2], and the test set of CLIC 2020 Professional [57], comparing various codecs (HiFiC [43], DiffEIC [38], GLC [26], DLF [59], ResULIC [28], MKIC [18], OSCAR [21], StableCodec [66]) across six metrics: LPIPS, DISTS, FID, KID, MS-SSIM, and PSNR. Figure 10 shows that our codec consistently delivers state-of-the-art perceptual quality, particularly under ultra-low bitrate conditions.

9.2. More Qualitative Results

Following the quantitative analysis, we also provide a more detailed subjective comparison of the three top-performing codecs: our codec, StableCodec [66], and DLF [59], under ultra-low bitrate conditions. Visual comparisons are presented in Figs. 11 to 14. The results show that our codec consistently achieves the best visual quality, corroborating the quantitative findings and confirming its superior perceptual performance at ultra-low bitrates.

9.3. Runtime Comparison

Table 2 compares the runtime efficiency of our codec against the variant StableCodec and DLF on the Kodak

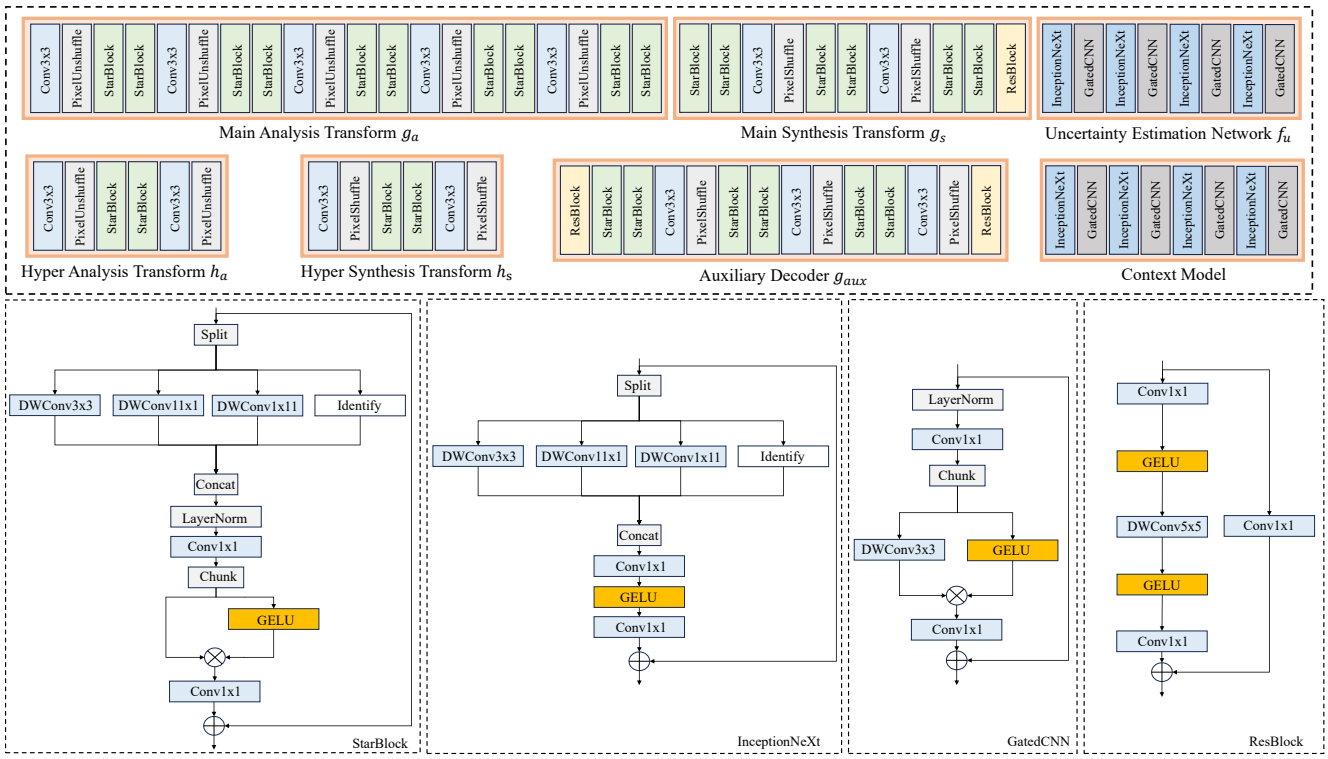

Figure 8. Network structures of the main modules, including the main analysis transform g_a, main synthesis transform g_s, hyper analysis transform h_a, hyper synthesis transform h_s, auxiliary decoder g_{aux}, uncertainty estimation network f_u, and the context model. For efficient network construction, we primarily rely on modified versions of InceptionNeXt [62], GatedCNN [61], and their combination (StarBlock).

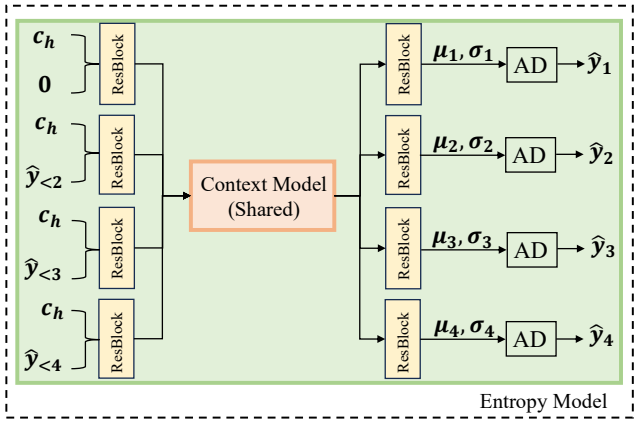

Figure 9. Illustration of our entropy modeling process.

dataset using a single RTX 3090 GPU. The encoding time of our codec and the variant StableCodec is significantly reduced [66], as they eliminate the need for the pre-trained VAE encoder used in Stable Diffusion and the auxiliary encoder along with the latent residual prediction (LRP) module. Our proposed uncertainty-guided adaptive quantization only slightly increases the encoding complexity compared to the variant StableCodec. For decoding, since our codec

Table 2. Runtime comparison of in seconds averaged on the Kodak dataset.

Method	Encoding Time (s)	Decoding Time (s)
Ours	0.034	0.355
StableCodec	0.030	0.263
DLF	0.181	0.195

incorporates an auxiliary decoder and a BLIP image captioning model, the overall decoding time brings a moderate increase. In addition, due to the inherent complexity of the diffusion model, our decoding time remains higher than that of DLF. However, it is important to note that in ultra-low bitrate application scenarios, such as mobile-satellite communications, the uplink bandwidth (from the mobile device to the satellite) is often more constrained and valuable than the downlink. This places a premium on low encoding complexity at the client side, while the more computationally intensive decoding can be offloaded to powerful cloud servers. Therefore, our codec—with its fast encoding and competitively performing decoding—retains strong practical value for real-world deployment.

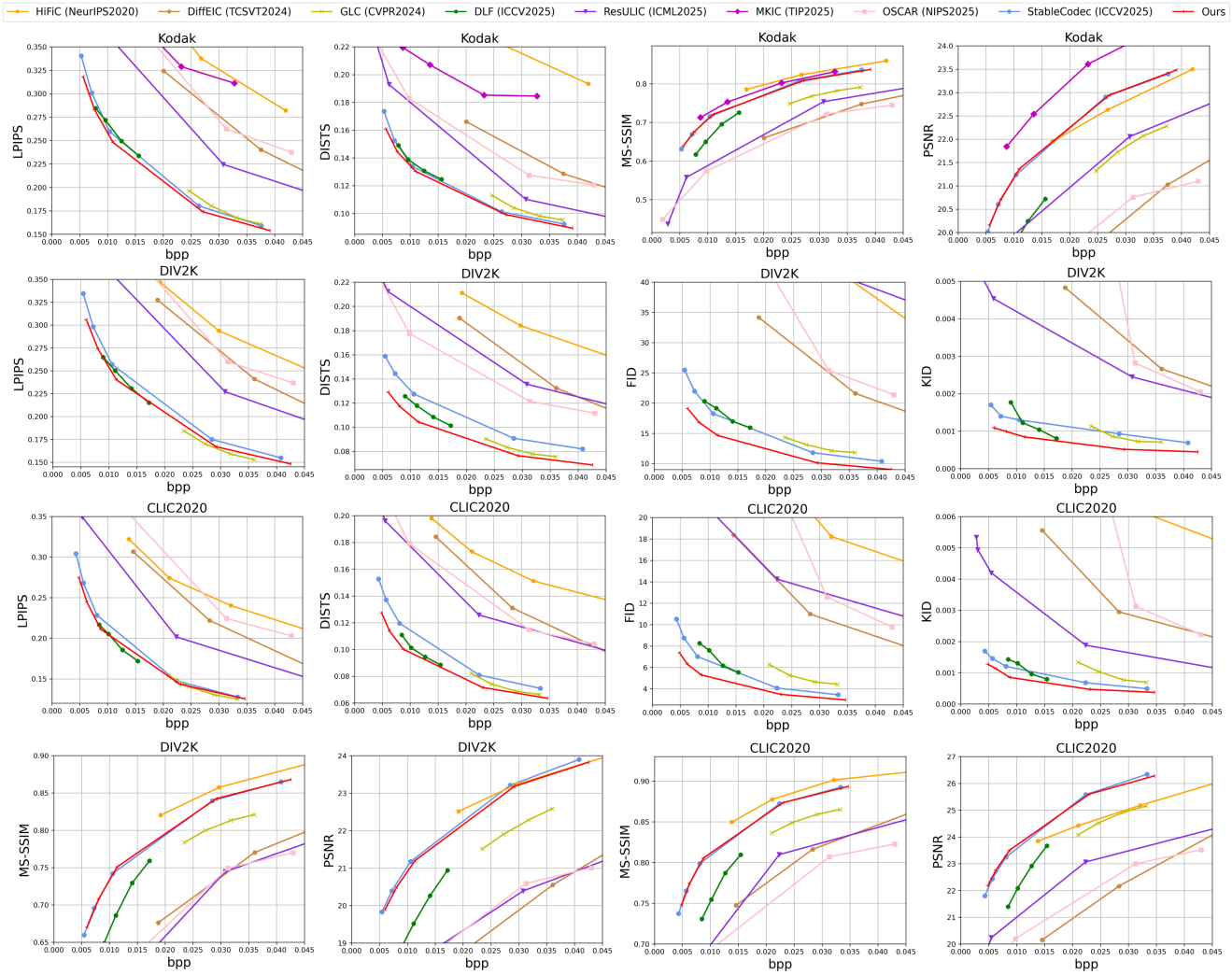

Figure 10. All rate-distortion curves of different generative image codecs on Kodak, the validation set of DIV2K, and the test set of CLIC 2020 Professional in terms of LPIPS, DISTS, FID, KID, MS-SSIM, and PSNR metrics.

9.4. User Study

To comprehensively evaluate the perceptual quality of reconstructed images at ultra-low bitrates, we conduct a user study on the Kodak dataset, comparing our codec with two top-performing codecs: StableCodec [66] and DLF [59]. The study follows a top-1 preference protocol, in which participants are asked to select the reconstruction they perceive as most visually consistent with the ground-truth image. Each participant evaluates 24 randomly ordered test cases. For each case, the ground-truth image is displayed together with the three reconstructed images in a single row of four images, with the order of the methods randomized across trials. Participants are instructed to select the reconstruction that is the most "consistent" with the original image. A total of 25 participants took part in the study, collectively contributing 600 evaluation cases. The results, summarized in Tab. 3, indicate that reconstructions from our codec were preferred in 58.5% of cases. This strong preference underscores its superior perceptual quality as judged by human observers.

Table 3. Top-1 user preference on the Kodak dataset.

Method	Ours	StableCodec	DLF
Bitrate (bpp)	0.0076	0.0072	0.0079
Top-1 Percentage (%)	58.5	29.0	12.5

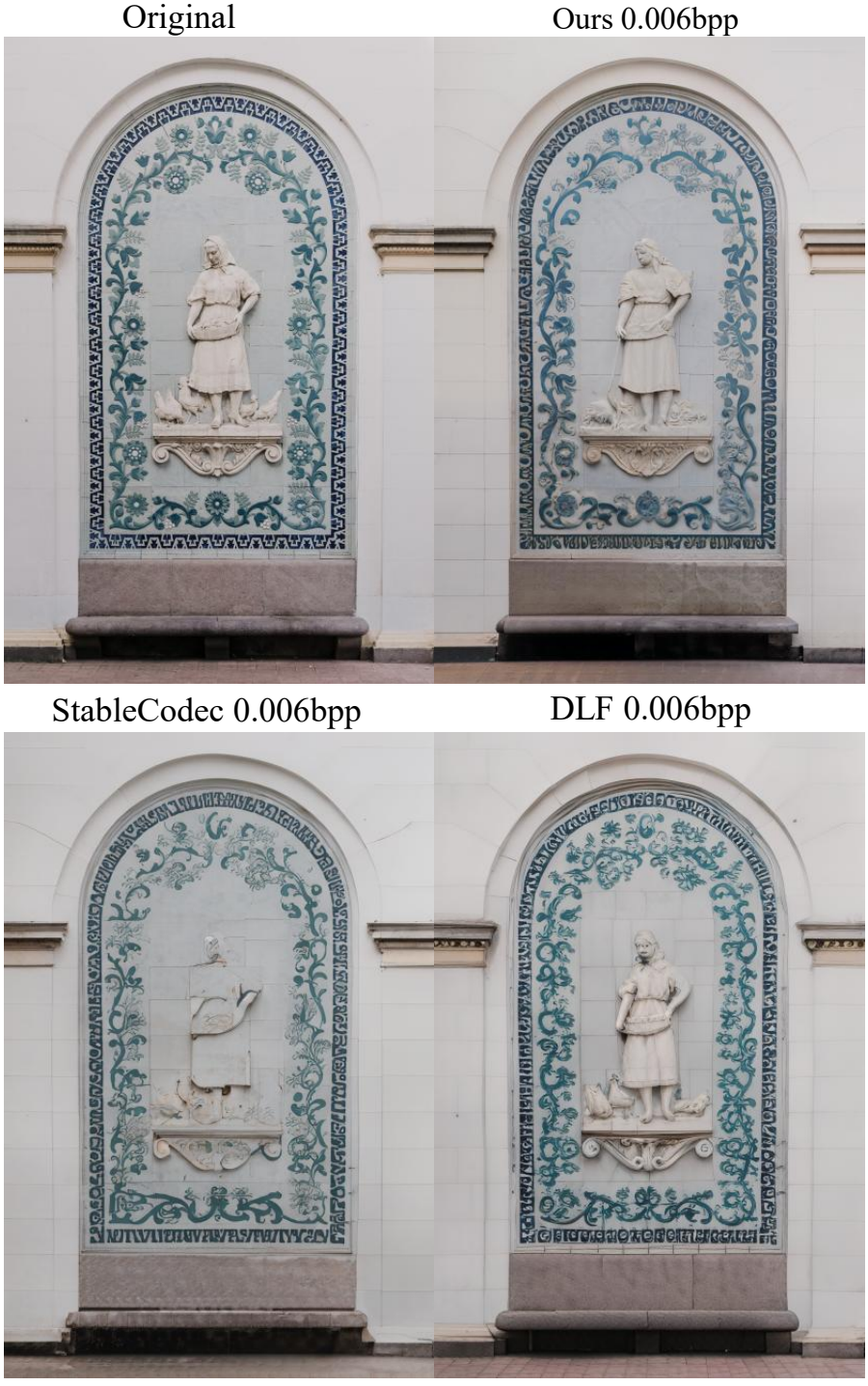

Figure 11. Visual examples and comparisons on 2K-resolution images from the test set of CLIC 2020 Professional.

Figure 12. Visual examples and comparisons on 2K-resolution images from the test set of CLIC 2020 Professional.

Figure 13. Visual examples and comparisons on 2K-resolution images from the validation set of DIV2K.

Original

Ours 0.008bpp

StableCodec 0.008bpp

DLF 0.008bpp

Figure 14. Visual examples and comparisons on 2K-resolution images from the validation set of DIV2K.