COORDINATING BITS AND PRIORS: COMPACT AND FAITHFUL DECODING FOR GENERATIVE IMAGE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image compression methods have recently shown strong perceptual quality at ultra-low bitrates. Pretrained generative models provide strong natural-image priors for ultra-low-rate image compression, but they are designed for image synthesis and generation rather than reconstructing a particular source from a decoded bitstream. This motivates two considerations: different stages of the generative prior contribute unequally to reconstruction, and the decoded evidence needs to be preserved as synthesis progresses toward pixels. We propose CFIC, a compact and faithful generative image compression framework built around these observations. A training-free analysis reveals that reconstruction importance varies substantially across stages, motivating asymmetric compaction of the one-step diffusion prior rather than uniform shrinking. Multi-scale evidence recovered from the same quantized representation is then coordinated with corresponding synthesis stages, allowing source-specific information to complement the generative prior without additional side information. Finally, we alleviate the decoding bottleneck inherited from the pretrained model by preserving this evidence hierarchically along the path to pixels. Experimental results demonstrate that our proposed method achieves state-of-the-art perceptual quality with a compact model size and lower computational cost.

1 INTRODUCTION

The rapid growth of visual data makes image compression fundamental to efficient storage and transmission. Conventional standards (Wallace, 1991; Bross et al., 2021) provide practical coding solutions, while learned image compression jointly optimizes nonlinear transforms and entropy models for the rate–distortion objective (Ballé et al., 2017; Ballé et al., 2018). Subsequent work has improved entropy modeling (Minnen et al., 2018; Minnen & Singh, 2020; He et al., 2021; 2022; Qian et al., 2022; Jiang et al., 2023; 2025) and transform design (Cheng et al., 2020; Zou et al., 2022; Liu et al., 2023; Bao et al., 2025; Feng et al., 2025; Zeng et al., 2025; Ma et al., 2026; Chen et al., 2026). Yet distortion-oriented optimization tends to smooth textures at low rates, reflecting the fundamental tension among bitrate, pixel fidelity, and perceptual quality (Blau & Michaeli, 2019).

Generative compression addresses this issue by using learned image priors to synthesize plausible details that cannot be explicitly conveyed by the bitstream. Adversarial and perceptual codecs trade exact pixel recovery for visual realism (Zhang et al., 2018; Mentzer et al., 2020; Yang & Mandt, 2023), while newer methods use discrete representations, semantic guidance, diffusion models, or autoregressive priors (Careil et al., 2024; Jia et al., 2024; Körber et al., 2024; Li et al., 2025a; Ke et al., 2025; Zhang et al., 2025; Guo et al., 2026; Zhang et al., 2026b). These priors are particularly useful at ultra-low rates, where the bitstream can preserve only a small amount of source-specific information.

However, compression is not generic image generation. A pretrained generative prior captures broad image statistics, whereas a codec must reconstruct a particular source from decoded information. If the transmitted representation is denoted by Y, the remaining uncertainty is $H(X \mid Y) \leq H(X)$. This motivates using the generative prior to complement the decoded representation while retaining its source-specific constraints. This observation raises two practical questions. *First, once the source*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

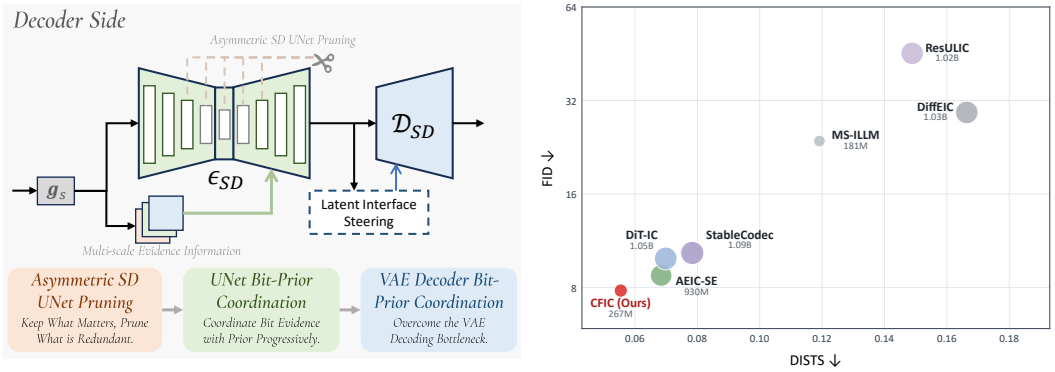


Figure 1: **Compact and faithful generative image compression.** Left: CFIC specializes a generative prior through asymmetric stable diffusion U-Net pruning, progressive bit-prior coordination, and decoder-side latent interface steering, keeping transmitted evidence available throughout reconstruction. Right: DIV2K results at 0.025 bpp within each method's measured bitrate range; bubble area indicates total parameter count, and lower-left is better for both DISTS and FID.

is constrained by decoded bits, how much generative capacity is still necessary at different stages of reconstruction? Second, how can the decoded evidence continue to constrain the generative prior as synthesis progresses toward pixels? For a fixed bitstream, these questions concern how decoded information is used, rather than an increase in the information transmitted.

Recent generative codecs have substantially improved conditioning, representation, and sampling efficiency (Li et al., 2025a; Ke et al., 2025; Xue et al., 2025; Sheng et al., 2026; Zhang et al., 2025; 2026a; Shi et al., 2026). Despite these advances, the allocation of capacity within the source-conditioned generative backbone remains less explored. This motivates two complementary directions. First, a large generative backbone continues to use capacity in stages that may contribute little once the source has been constrained. Second, decoded source features need effective access to both the generative synthesis stages and the final image decoder.

We address these issues with **CFIC, Compact and Faithful Image Compression**. CFIC starts from a simple principle: once decoded bits constrain the source, the generative prior should resolve what remains uncertain, while the transmitted evidence should continue to steer reconstruction throughout the decoding path. Following this principle, we observe the highly uneven reconstruction sensitivity of the one-step diffusion prior and compact it asymmetrically, concentrating the capacity on the reconstruction-critical stages rather than uniformly shrinking the backbone. The same quantized representation is then reused to recover multi-scale source evidence and coordinate it with corresponding synthesis stages, strengthening source conditioning without additional transmitted bits. Finally, we adapt the pretrained image-decoder interface through a trainable wide input projection and hierarchical injection to relieve its information bottleneck and carry this evidence into pixel space. Together, these designs couple model compaction with source fidelity: the prior is retained where it remains useful, while reconstruction stays anchored to the information actually conveyed by the bitstream.

Our contributions are summarized as follows:

- We reveal that generative capacity contributes highly unevenly to source reconstruction once decoded evidence is available, and exploit this asymmetry to compact the one-step diffusion prior while preserving reconstruction-critical stages.
- We introduce multi-scale residual coordination and a trainable image-decoder interface that reuse decoded features throughout reconstruction.
- Across Kodak, DIV2K, and CLIC 2020 datasets, our method achieves state-of-the-art perceptual performance with lower model size and computation relative to the compared diffusion-based methods.

2 Related Work

2.1 Learned Image Compression

Learned image compression jointly optimizes analysis and synthesis transforms and entropy models under a rate–distortion objective (Ballé et al., 2017; Ballé et al., 2018). Entropy modeling has advanced through autoregressive and channel-wise context models (Minnen et al., 2018; Minnen & Singh, 2020), checkerboard and uneven space–channel modeling (He et al., 2021; 2022), transformers (Qian et al., 2022), and multi-reference prediction (Jiang et al., 2023; 2025). Transform designs have also improved through attention and hybrid convolutional–transformer architectures (Cheng et al., 2020; Zou et al., 2022; Liu et al., 2023), shift operations (Bao et al., 2025), linear attention (Feng et al., 2025), state-space modeling (Zeng et al., 2025), adaptive frequency processing (Ma et al., 2026), and graph-adaptive coding (Chen et al., 2026). At ultra-low bitrates, however, distortion-oriented optimization tends to suppress textures, reflecting the rate–distortion–perception trade-off (Blau & Michaeli, 2019).

2.2 Perceptual and Generative Image Compression

Perceptual compression exploits learned image statistics to reconstruct plausible content beyond what the bitstream can fully convey. Adversarial codecs combine adversarial and perceptual objectives to improve realism (Mentzer et al., 2020), while other approaches exploit learned codebooks and generative representations (Careil et al., 2024; Jia et al., 2024; Körber et al., 2024). Conditional diffusion further demonstrates how compact representations can guide generative reconstruction (Yang & Mandt, 2023).

Recent work improves the interaction between compressed evidence and generative priors. Li et al. (2025a) use latent guidance, while Ke et al. (2025) employ semantic residual modeling. Xue et al. (2025) combine dataset-level semantics and instance-specific details through dual-generative latent fusion, and Li et al. (2025b) reconstruct residual information through relay diffusion. Other directions include content-adaptive conditioning (Sheng et al., 2026), autoregressive reconstruction (Zhang et al., 2026b), and foundation-model guidance (Gao et al., 2025).

Efficiency is another important consideration. Zhang et al. (2025); Guo et al. (2026) enable one-step generative reconstruction, while subsequent methods reduce complexity through shallow encoding (Zhang et al., 2026a) or diffusion in a compact latent space (Shi et al., 2026). Our work aims to compact the generative prior while keeping decoded source features available throughout reconstruction.

3 Method

3.1 Overview

Figure 2 illustrates the overall framework. Given an input image, the learned codec transmits the quantized main and hyper latents and reconstructs two decoder-side signals from them: a prior feature that initializes one-step generative reconstruction and multi-scale evidence features aligned with several synthesis stages. These evidence features are consumed by lightweight bit-prior coordination modules (BPCM) at matched synthesis scales. Both signals are deterministic functions of the same coded latents and require no additional coded variables. The entropy-model rate estimate, used for rate optimization during training, is

$$R = -\log_2 p(\hat{\boldsymbol{z}}) - \log_2 p(\hat{\boldsymbol{y}} \mid \hat{\boldsymbol{z}}). \tag{1}$$

The bitstream supplies source-specific features, while the pretrained prior supplies learned image statistics. We remove generative computation that has low reconstruction sensitivity under a fixed coded condition, inject the learned evidence residuals where the synthesis path reconstructs the corresponding scale, and adapt the image-decoder interface to retain access to these features. The generative prior uses one denoising step, and every added coordination path is a lightweight residual branch. This separates the amount of information transmitted by the codec from how often that information is accessed during reconstruction.

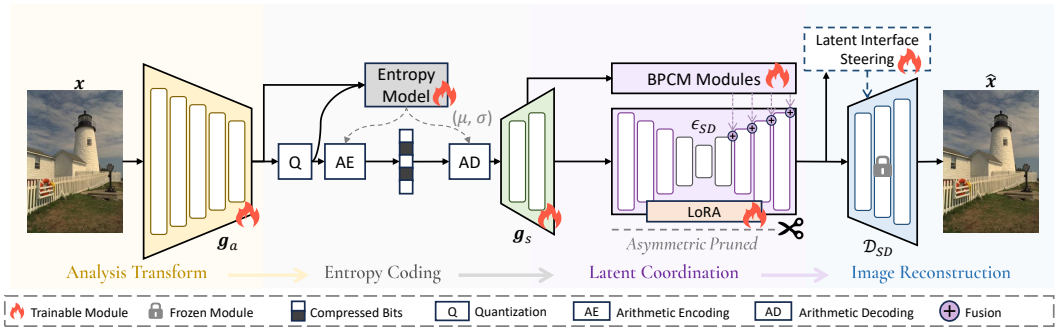


Figure 2: **Overall framework of CFIC.** The codec reconstructs both a prior feature and multi-scale source evidence from the transmitted representation. Reconstruction sensitivity guides asymmetric prior compaction; bit-prior coordination modules (BPCM) inject learned evidence residuals at matched synthesis stages; and decoder-side latent interface steering uses a trainable wide entrance with hierarchical injection to preserve the coordinated representation during image decoding.

3.2 Asymmetric Prior Compaction

A pretrained diffusion backbone distributes capacity for broad image generation, not for reconstruction under a known bitstream. We therefore measure the reconstruction sensitivity of different stages under fixed decoded conditions. For a structural perturbation q under fixed decoded latents,

$$S(q) = \mathcal{E}(\mathcal{M}_q; \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}) - \mathcal{E}(\mathcal{M}; \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}), \tag{2}$$

where $\mathcal{E}$ denotes the reconstruction criterion. All unperturbed parameters are held fixed, and no recovery training is performed during the sweep. Thus, $S(q)$ measures the immediate effect of perturbation q on reconstruction, providing an empirical guide for stage-wise pruning. Each tested stage retains 50% of its channels, starting from the full, adapted model.

The analysis in Appendix C.1 reveals a pronounced depth dependence. Deep, low-resolution stages tolerate structural perturbation with little perceptual change, whereas later synthesis stages are substantially more sensitive. Zero-step pruning trials show the same trend: retaining more synthesis capacity causes smaller immediate reconstruction errors than uniform pruning. We therefore compact the prior asymmetrically, pruning deep computation most aggressively while preserving more capacity as spatial detail is reconstructed. Selected residual and attention operators in the retained path are further replaced by lightweight depthwise/pointwise and linear-attention counterparts.

Table 3 compares the immediate and recovered effects of the asymmetric and uniform topologies under matched recovery conditions. Structural replacement is then stabilized with online heavy-to-light annealing, as illustrated in Fig. 3(d). Each lightweight operator is trained beside a frozen heavy counterpart; the heavy response is gradually annealed to zero while feature- and reconstruction-level objectives transfer the adapted behavior. All heavy branches are removed before deployment.

3.3 Bit-Prior Coordination Module

Compaction reduces redundant prior capacity, but it does not by itself guarantee that decoded source information remains visible to the synthesis path. We therefore reconstruct evidence features from the quantized representation and align them with the stages where the corresponding spatial detail is synthesized. At scale s, a bit-prior coordination module receives the current synthesis feature $\boldsymbol{h}_s^u$, the corresponding native skip feature $\boldsymbol{h}_s^d$ as reconstruction context, and bit-derived evidence $\boldsymbol{e}_s$:

$$\begin{aligned} \boldsymbol{q}_s &= Q_s(\mathrm{Norm}(\boldsymbol{h}_s^u)), \qquad \boldsymbol{k}_s = K_s(\mathrm{Norm}(\boldsymbol{h}_s^d)), \\ \boldsymbol{e}_s^{\mathrm{inn}} &= E_s(\boldsymbol{e}_s) - D_s(\boldsymbol{h}_s^d), \\ \boldsymbol{z}_s &= O_s\big(\mathrm{CrossAttn}(\boldsymbol{q}_s, \boldsymbol{k}_s, \boldsymbol{e}_s^{\mathrm{inn}})\big), \\ \tilde{\boldsymbol{h}}_s^u &= \boldsymbol{h}_s^u + m_s \tanh(\alpha_s)\boldsymbol{z}_s. \end{aligned} \tag{3}$$

Here, m_s bounds the magnitude of the residual coefficient and α_s is a learned residual scale.

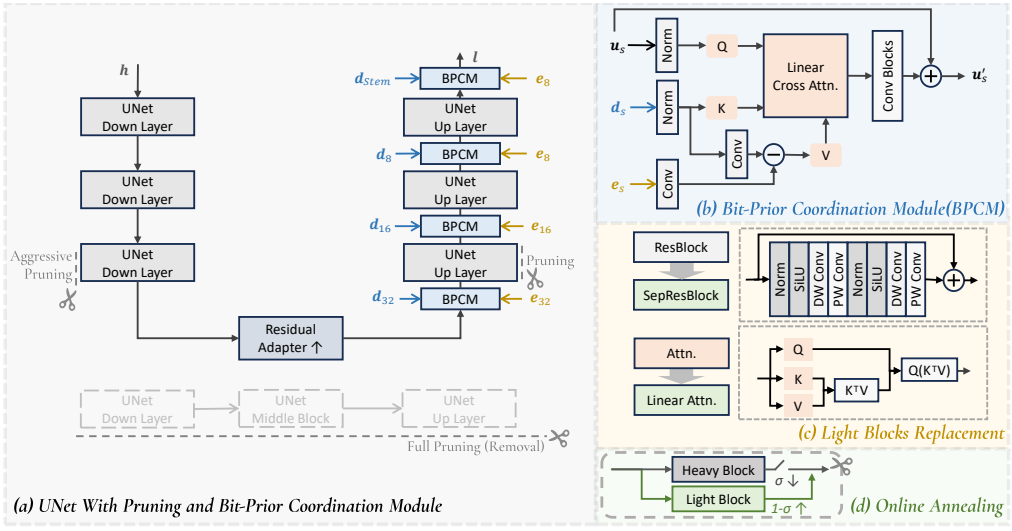


Figure 3: **Asymmetric prior compaction and Bit-Prior Coordination Module.** (a) Depth-aware pruning preserves reconstruction-sensitive synthesis while reducing deep computation; a residual adapter reconnects the retained path after full deep-block removal. (b) The coordination module combines the current synthesis feature, matched context, and bit-derived evidence through linear cross-attention. (c) Heavy residual and attention operators are replaced by separable residual blocks and linear-attention counterparts. (d) Online heavy-to-light annealing transfers the adapted response before training-only branches are removed.

The key operation is the innovation term e_s^{inn}, a learned residual between projected evidence and native context, allowing the module to learn corrections relative to the context already available at that stage. Linear cross-attention then aligns this feature residual with the current synthesis state before a residual update with a bounded coefficient. This differs from feeding all coded information only at the network entrance: coarse evidence can correct coarse structure early, while finer evidence remains available when edges and textures are actually reconstructed.

Because every evidence feature is decoded from the same entropy-coded latents, this coordination requires no additional side information. The bitstream and the prior retain distinct roles: transmitted evidence constrains source-specific content, while the compact prior fills in details that remain uncertain after quantization. Appendix A.3 gives the deployed module structure.

3.4 Decoder-Side Bit-Prior Coordination via Latent Interface Steering

The coordinated latent must still pass through the pretrained image decoder before pixels are produced. We refer to our decoder-side coordination mechanism as Latent Interface Steering (LIS). It combines a trainable wide entrance with lightweight hierarchical injection and introduces no additional coded representation.

A narrow frozen entrance can limit how enriched latent features reach the pretrained decoder. Repeating the same pretrained kernel across input-channel groups while keeping its weights frozen preserves this constraint. Let $\bar{W} = [W, \ldots, W]$ and $A(\boldsymbol{h}) = \sum_g \boldsymbol{h}_g$, where each $\boldsymbol{h}_g$ has d_{in} channels. Then

$$\bar{W} * \boldsymbol{h} = W * A(\boldsymbol{h}), \qquad \operatorname{rank}(A) \le d_{\text{in}}, \tag{4}$$

where the rank is defined over channels at each spatial location. Thus, the frozen repeated projection cannot distinguish representations with the same channel-group sum. This restriction concerns the frozen repeated projection; it does not apply to a general wide projection with independently learned channel weights.

We instead learn a wide input projection while keeping the main pretrained decoder body fixed. A projected state derived from the source-conditioned wide latent is injected at several decoder depths. Each residual path combines the local decoder state, the projected latent state, high-frequency

responses, and a lightweight evidence-magnitude statistic:

$$\boldsymbol{z}_k^+ = \boldsymbol{z}_k + \gamma_k \tanh(\alpha_k)\, \Delta_k(\boldsymbol{z}_k, \boldsymbol{\rho}_k, \boldsymbol{u}_k), \tag{5}$$

where Δ_k is a lightweight gated residual and γ_k bounds the residual coefficient.

The trainable wide entrance removes the fixed channel-group aggregation constraint, while hierarchical injection keeps source-conditioned features accessible as spatial detail is reconstructed. This is especially relevant for small edges, repeated textures, and local color structure that may otherwise be attenuated after the first projection. Within the image decoder, the entrance and lightweight residual paths are optimized during final decoder adaptation; the pretrained decoder body remains fixed. Appendix A.4 provides implementation details.

4 Experiments

4.1 Experimental Setup

Training. Training proceeds in three stages. We first train the full compression–generative model on MLIC-Train-100K for 250K steps, then obtain the compact prior through asymmetric pruning and heavy-to-light annealed distillation. The compact student is fine-tuned for each target rate, followed by high-resolution adaptation on DF2K and CLIC Train, following the data protocol of Zhang et al. (2026a). All trainable components use AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017). During rate-specific fine-tuning, the pretrained base weights remain frozen, while low-rank adapters and the designated codec and coordination modules are optimized. Appendix B.1 lists the stage-wise objectives, crop sizes, and trainable components.

Datasets and metrics. We evaluate on Kodak, DIV2K validation, and CLIC2020 Professional Test, following recent generative compression protocols (Careil et al., 2024; Zhang et al., 2025; 2026a; Shi et al., 2026). All methods are evaluated using full-resolution images under the same evaluation protocol. None of the evaluation images are used for training. Rate is reported in bits per pixel(bpp), computed from the actual coded bitstream length. We use PSNR and MS-SSIM (Wang et al., 2003) for distortion, AlexNet-LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) for perceptual similarity, and FID (Heusel et al., 2017) on DIV2K and CLIC2020. FID is computed from overlapping 256×256 patches extracted from full-image reconstructions and references, using the same patch-extraction protocol as Muckley et al. (2023); Zhang et al. (2026a). BD-rate is computed by fitting $\ln R$ as a cubic function of the metric value and integrating only over the overlapping measured range (Bjontegaard, 2001); negative values indicate rate savings relative to the anchor. Appendix B.4 gives the exact definition.

Baselines and implementation. We compare against representative perceptual and generative codecs covering Muckley et al. (2023); Li et al. (2025a); Ke et al. (2025); Zhang et al. (2025); Xue et al. (2025); Li et al. (2025b); Shi et al. (2026); Zhang et al. (2026a;b). Baseline operating points are obtained either by evaluating the publicly released checkpoints (Muckley et al., 2023; Li et al., 2025a; Zhang et al., 2025; Xue et al., 2025; Li et al., 2025b; Zhang et al., 2026a;b) or from the corresponding papers' officially released results (Ke et al., 2025; Shi et al., 2026), under the same full-image evaluation protocol. Methods without sufficient quality-range overlap remain visible in the curves but are omitted from Bjøntegaard-rate comparisons. The one-step prior is initialized from SD-Turbo (Sauer et al., 2024), and the lightweight image decoder from AdcSR (Chen et al., 2025). Training-only branches are removed and low-rank adapters are merged before evaluation.

Complexity protocol. All latency and multiply–accumulate measurements in Table 1 use 1024×1024 inputs. End-to-end latency (including entropy encoding and decoding) is measured on a single NVIDIA GeForce RTX 4090 and reported in milliseconds; multiply–accumulate counts are reported in TMACs. All reported latencies are measured by us on the same GPU. High-resolution reconstruction uses a fixed tiled-prior protocol with spatially aligned evidence crops; details are provided in Appendix B.2. Evaluation uses the resulting full-resolution reconstructions. The patches used for FID are extracted after reconstruction and are separate from the latent tiles used during inference.

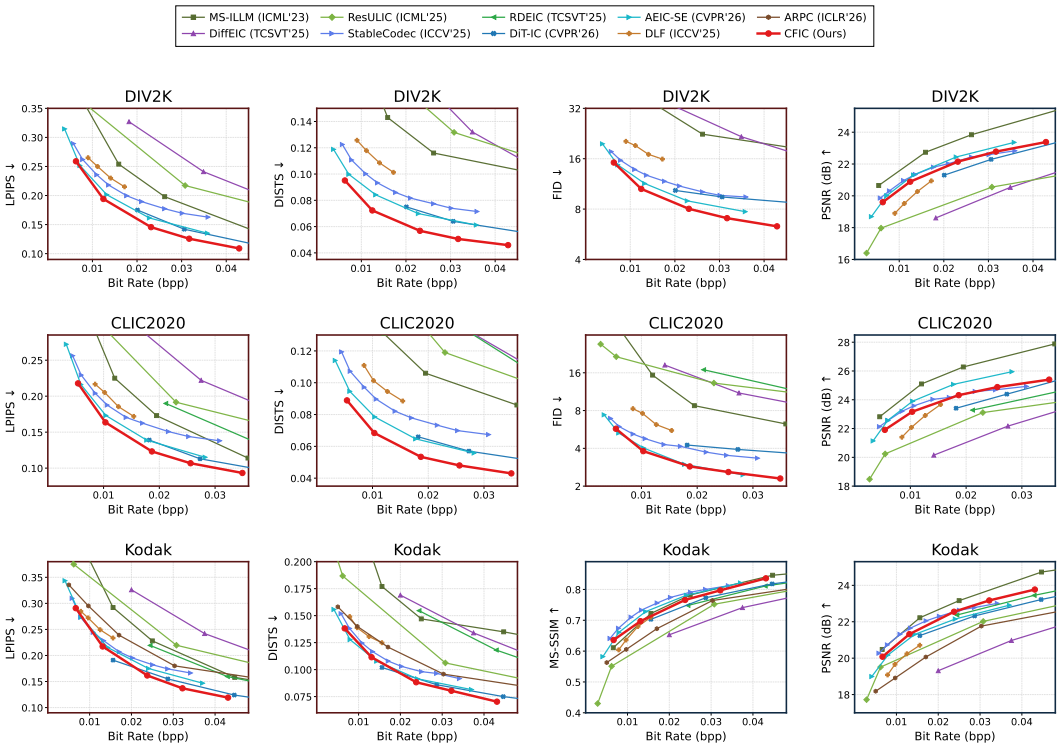


Figure 4: **Rate–distortion–perception curves.** Results on DIV2K, CLIC2020, and Kodak. Lower is better for LPIPS, DISTS, and FID; higher is better for PSNR and MS-SSIM.

Table 1: **Model complexity and perceptual BD-rate (%).** The compact-latent diffusion codec of Shi et al. (2026) is the anchor. Negative values denote rate savings at matched metric value; '–' denotes unavailable data or insufficient overlap. Best results are bold and second-best results are underlined.

Methods	Complexity			Kodak		DIV2K		CLIC2020	
	Total params↓	Latency (ms)↓	TMACs↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓	LPIPS↓	DISTS↓
MS-ILLM (Muckley et al., 2023)	**181M**	321.3	**1.02**	53.23	374.09	38.14	272.08	26.22	200.34
DiffEIC (Li et al., 2025a)	1.03B	11702.7	76.11	162.69	176.37	152.11	197.37	180.04	254.40
RDEIC (Li et al., 2025b)	1.38B	1262.0	10.28	52.82	121.37	51.76	127.28	65.97	260.04
ResULIC (Ke et al., 2025)	1.02B	6829.5	11.37	131.35	79.07	143.22	324.71	156.47	252.84
StableCodec (Zhang et al., 2025)	1.09B	729.8	9.10	37.73	38.81	43.25	50.06	63.19	–
DiT-IC (Shi et al., 2026) *(Anchor)*	1.05B	524.8	7.61	0.00	0.00	0.00	0.00	0.00	0.00
AEIC-SE (Zhang et al., 2026a)	930M	289.7	2.82	15.57	3.41	-0.72	-2.65	0.04	-5.70
CFIC (Ours)	267M	**285.9**	2.40	**-7.50**	**-9.26**	**-19.39**	**-44.81**	**-18.93**	**-42.05**

4.2 Rate–Distortion–Perception Performance

Across all three datasets, CFIC improves the perceptual rate–distortion trade-off over the compared baselines. The separation is most pronounced on DIV2K and CLIC2020, where the DISTS and FID curves remain below the compared diffusion-based baselines through most of the shared low-rate range. PSNR and MS-SSIM remain competitive at comparable rates, which is consistent with the expected perception–distortion trade-off.

Table 1 summarizes perceptual BD-rate using the compact-latent diffusion codec of Shi et al. (2026) as the common anchor. CFIC improves all six dataset–metric combinations. Relative to this anchor, the rate savings at matched LPIPS are 7.50%, 19.39%, and 18.93% on Kodak, DIV2K, and CLIC2020, respectively; the corresponding savings at matched DISTS are 9.26%, 44.81%, and 42.05%. The larger DISTS savings on the two high-resolution sets indicate that the improvement is particularly pronounced under the structure-aware perceptual metric.

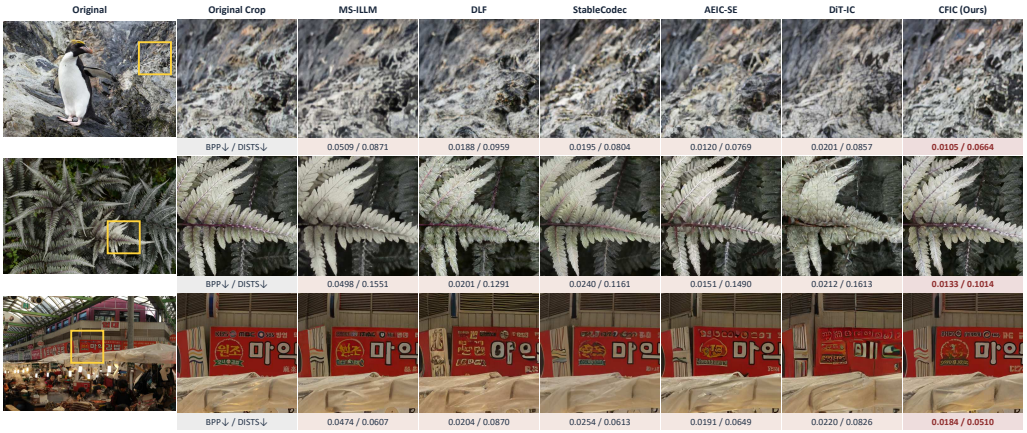


Figure 5: **Qualitative comparison.** Each crop is paired with its image-level bits per pixel and DISTS value; lower is better for both. High-resolution full-frame comparisons are provided in Appendix D.

Table 2: **Evidence-pathway ablation on Kodak.** BD-rate (%) is computed with the full CFIC model as the anchor; lower is better. All variants are retrained under matched training conditions.

	BD-rate (%) ↓		
Variant	PSNR	LPIPS	DISTS
CFIC (Ours)	**0.00**	**0.00**	**0.00**
Generative-prior ablation			
w/o BPCM	+5.56	+5.74	+8.50
Image-decoder ablation			
w/o hierarchical injection	+8.23	+12.94	+21.92
frozen 4-ch conv-in	+2.83	+15.27	+25.77
frozen repeated conv-in (Zhang et al., 2026a)	+14.06	+13.42	+25.58

Deployment efficiency. The compactness gain is visible in computation as well as parameter count. At 1024×1024, CFIC uses 2.40 TMACs and 285.9 ms. Relative to the compact-latent diffusion anchor, this corresponds to 74.6% fewer parameters, 68.5% fewer multiply–accumulates, and 45.5% lower measured latency. Compared with AEIC-SE (Zhang et al., 2026a), CFIC reduces parameters by 71.3% and MACs by 14.9% while maintaining comparable measured latency. These measurements complement the pruning ablation below: the goal is not only to shrink the network, but to retain the synthesis capacity that matters for source-conditioned reconstruction.

4.3 Qualitative Comparison

Figure 5 compares representative reconstructions at nearby low-rate operating points. CFIC retains repeated texture, thin foliage, and boundaries with less local shape drift in the highlighted regions. The additional high-resolution panels in Appendix D show the same behavior on faces, indoor fixtures, buildings, and circular glassware. Because the methods are not evaluated at exactly identical rates in every panel, the visual comparisons should be read together with the exact bits-per-pixel labels and the quantitative curves in Fig. 4.

4.4 Ablation Studies

Bit-prior coordination and the image-decoder interface. Table 2 evaluates synthesis-side coordination and the image-decoder interface through retrained component variants. The first group removes synthesis-side bit-prior coordination; the second changes the decoder-side steering interface while retaining the same bitstream structure.

Table 3: **Pruning strategy before and after recovery training on Kodak at 0.0549 bpp.** All parameter counts refer to the complete model.

Variant	Total params↓	PSNR↑	LPIPS↓	DISTS↓
Unpruned model	859M	**24.71**	**0.107**	**0.067**
Before recovery				
Asymmetric Pruned	267M	24.59	0.116	0.085
Uniform Pruned	273M	16.69	0.357	0.264
After recovery				
Asymmetric Pruned	267M	24.67	0.108	0.069
Uniform Pruned	273M	23.93	0.125	0.076

The retrained variant without synthesis-side coordination increases LPIPS and DISTS BD-rate by 5.74% and 8.50%, respectively. This shows that synthesis-side evidence coordination provides an additional benefit beyond the remaining conditioning pathways.

The decoder ablations show an additional benefit from adapting the image-decoder interface. The retrained variant with the original frozen four-channel entrance produces the largest LPIPS and DISTS BD-rate penalties, while the frozen repeated entrance also incurs substantial losses. Repeating the frozen kernels retains the channel-aggregation restriction in Eq. 4. The trainable wide entrance removes that restriction, and hierarchical injection provides a second benefit by refreshing source information after the first projection; removing it increases DISTS BD-rate by 21.92%. Together, the two ablation groups show that reconstruction benefits from access to source evidence both inside the compact prior and through the final image decoder.

Asymmetric pruning. Table 3 contrasts asymmetric pruning with uniform pruning at a similar compactness target. Reporting both stages shows how the initial topology difference evolves under matched recovery training.

Both compact variants use the same lightweight operators and heavy-to-light transfer procedure (Li et al., 2017; Chollet, 2017; Feng et al., 2025; Chen et al., 2025; Sun et al., 2026); they differ only in where capacity is removed. Before any recovery training, asymmetric pruning retains a 24.59 dB reconstruction with 0.116 LPIPS, whereas the similarly sized uniform topology yields 16.69 dB and 0.357 LPIPS. Recovery narrows the gap but does not erase it: the asymmetric model returns to 24.67 dB and 0.108 LPIPS, close to the unpruned teacher, while the uniform model remains worse on all three metrics. After matched recovery training, the asymmetric model exceeds the uniform model by 0.74 dB PSNR and reduces LPIPS and DISTS by 0.017 and 0.007, respectively. Relative to the unpruned model, the recovered asymmetric model loses 0.04 dB PSNR, with LPIPS and DISTS increasing by 0.001 and 0.002, respectively.

In the studied source-conditioned decoder, these ablations support two complementary design choices: non-uniform capacity reduction guided by reconstruction sensitivity, and direct access to bitstream evidence during synthesis and image decoding.

5 Conclusion

We presented CFIC, which combines asymmetric prior compaction with multi-scale reuse of decoded source features. Reconstruction sensitivity provides an empirical basis for the compact topology; multi-scale bit-prior coordination injects learned evidence residuals into the synthesis path; and decoder-side latent interface steering preserves access to source-conditioned features through a trainable wide entrance and hierarchical injection. Together, these components yield a compact deployed model with strong perceptual rate–distortion performance across Kodak, DIV2K, and CLIC 2020, without additional side information. The ablations further support the complementary benefits of asymmetric capacity allocation and continued access to decoded evidence. These results suggest a practical design principle for source-conditioned generative compression: allocate prior capacity according to reconstruction sensitivity while keeping transmitted information accessible throughout synthesis and image decoding.

Reproducibility Statement

We have made every effort to improve the reproducibility of our work. The main text and appendices describe the method, training procedure, and evaluation protocols. The core implementation is included in the supplementary materials, and we will publicly release the code and supporting materials upon acceptance.

AI use statement

Generative AI tools were used to aid and polish writing, including manuscript drafting and language editing, figure layout, and LaTeX formatting. AI-assisted outputs were manually reviewed by the authors. The authors take responsibility for the final content of this work, including text, claims, citations, figures and other artifacts.

References

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJxdQ3jeg.

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. URL https://openreview.net/forum?id=rkcQFMZRb.

Youneng Bao, Wen Tan, Chuanmin Jia, Mu Li, Yongsheng Liang, and Yonghong Tian. Shiftlic: Lightweight learned image compression with spatial-channel shift operations. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):9428–9442, 2025.

Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, 2019.

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(10):3736–3764, 2021.

Marlene Careil, Matthew J Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *International conference on learning representations*, volume 2024, pp. 15544–15564, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28208–28220. IEEE, 2025.

Yunuo Chen, Bing He, Zezheng Lyu, Hongwei Hu, Qunshan Gu, Yuan Tian, and Guo Lu. Adaptive learned image compression with graph neural networks. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7936–7945. IEEE, 2020.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1800–1807. Ieee, 2017.

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.

Donghui Feng, Zhengxue Cheng, Shen Wang, Ronghua Wu, Hongwei Hu, Guo Lu, and Li Song. Linear attention modeling for learned image compression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7623–7632. IEEE, 2025.

Junlong Gao, Zhimeng Huang, Qi Mao, Siwei Ma, and Chuanmin Jia. Exploring multimodal knowledge for image compression via large foundation models. *IEEE Transactions on Image Processing*, 2025.

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. *Advances in Neural Information Processing Systems*, 38:85267–85286, 2026.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14766–14775. IEEE, 2021.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5708–5717. IEEE, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image compression. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26088–26098. IEEE, 2024.

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*, pp. 7618–7627, 2023. doi: 10.1145/3581783.3611694.

Wei Jiang, Jiayu Yang, Yongqi Zhai, Feng Gao, and Ronggang Wang. Mlic++: Linear complexity multi-reference entropy modeling for learned image compression. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(5):1–25, 2025.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29626–29650. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/ke25c.html`.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nikolai Körber, Eduard Kromer, Andreas Siebert, Sascha Hauke, Daniel Mueller-Gritschneder, and Björn Schuller. Egic: enhanced low-bit-rate generative image compression guided by semantic segmentation. In *European Conference on Computer Vision*, pp. 202–220. Springer, 2024.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJqFGTslg.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Jingwen Jiang. Toward extreme image compression with latent feature guidance and diffusion prior. *IEEE Trans. Cir. and Sys. for Video Technol.*, 35(1):888–899, January 2025a. ISSN 1051-8215. doi: 10.1109/TCSVT.2024.3455576. URL https://doi.org/10.1109/TCSVT.2024.3455576.

Zhiyuan Li, Yanhui Zhou, Hao Wei, Chenyang Ge, and Ajmal Mian. Rdeic: Accelerating diffusion-based extreme image compression with relay residual diffusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11):11540–11552, 2025b.

Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14388–14397. IEEE, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Huidong Ma, Xinyan Shi, Hui Sun, Xiaofei Yue, Xiaoguang Liu, Gang Wang, and Wentong Cai. Learned image compression via sparse attention and adaptive frequency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41278–41287, 2026.

Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in neural information processing systems*, 33:11913–11924, 2020.

David Minnen and Saurabh Singh. Channel-wise autoregressive entropy models for learned image compression. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 3339–3343. IEEE, 2020.

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.

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. In *International Conference on Machine Learning*, pp. 25426–25443. PMLR, 2023.

Yichen Qian, Xiuyu Sun, Ming Lin, Zhiyu Tan, and Rong Jin. Entroformer: A transformer-based entropy model for learned image compression. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=VrjOFfcnSV8.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXVI*, pp. 87–103, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73015-3. doi: 10.1007/978-3-031-73016-0_6. URL https://doi.org/10.1007/978-3-031-73016-0_6.

Xihua Sheng, Lingyu Zhu, Tianyu Zhang, Dong Liu, Shiqi Wang, and Jing Wang. CADC: Content adaptive diffusion-based generative image compression. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026.

Junqi Shi, Ming Lu, Xingchen Li, Anle Ke, Ruiqi Zhang, and Zhan Ma. Dit-IC: Aligned diffusion transformer for efficient image compression. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026. URL https://openreview.net/forum?id=CpOqQWhLZi.

Haoze Sun, Linfeng Jiang, Fan Li, Renjing Pei, Zhixin Wang, Yong Guo, Jiaqi Xu, Haoyu Chen, Jin Han, Fenglong Song, et al. Pocketsr: The super-resolution expert in your pocket mobiles. *Advances in Neural Information Processing Systems*, 38:93250–93268, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4): 30–44, 1991.

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*, volume 2, pp. 1398–1402. Ieee, 2003.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European conference on computer vision*, pp. 3–19. Springer, 2018.

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

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models. *Advances in Neural Information Processing Systems*, 36:64971–64995, 2023.

Fanhu Zeng, Hao Tang, Yihua Shao, Siyu Chen, Ling Shao, and Yan Wang. Mambaic: State space models for high-performance learned image compression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18041–18050. IEEE, 2025.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Jean-Daniel Boissonnat, Patrick Chenin, Albert Cohen, Christian Gout, Tom Lyche, Marie-Laurence Mazure, and Larry Schumaker (eds.), *Curves and Surfaces*, pp. 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-27413-8.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. Stablecodec: Taming one-step diffusion for extreme image compression. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17379–17389. IEEE, 2025.

Tianyu Zhang, Dong Liu, and Chang Wen Chen. Ultra-low bitrate perceptual image compression with shallow encoder. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026a.

Ziyuan Zhang, Yichong Xia, Bin Chen, Tianwei Zhang, Hao Wang, and Han Qiu. Autoregressive-based progressive coding for ultra-low bitrate image compression. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=FXu4G5T5QZ.

Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17471–17480. IEEE, 2022.

A Model Architecture

This appendix describes the codec, compact generative prior, synthesis-side coordination, and image-decoder interface. Appendix B presents the training and evaluation procedures, followed by additional analyses in Appendix C.

A.1 Codec and Shared Bitstream

Figure A.1 summarizes the analysis transform, hyperprior, entropy model, and synthesis transform. The codec recovers both the prior feature and multi-scale evidence from the quantized main and hyper latents. These decoder-side features share the same transmitted bitstream.

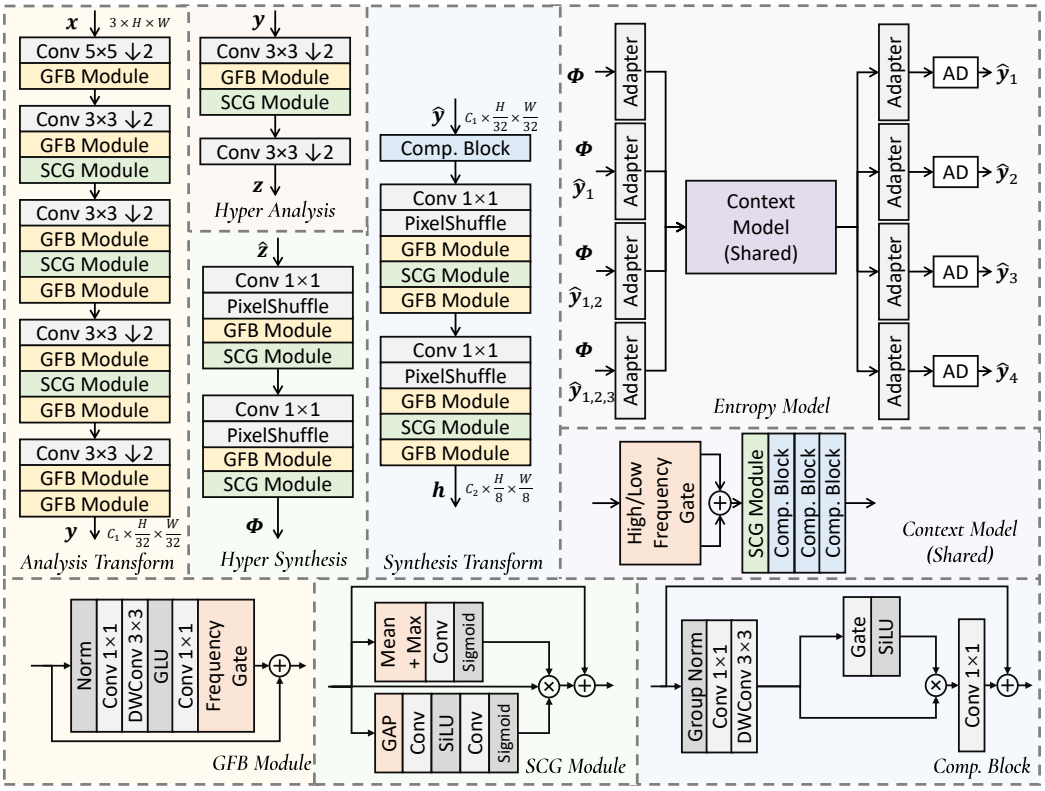


Figure A.1: **Codec architecture.** The analysis transform produces the main latent, the hyperprior supplies entropy context, and the synthesis transform recovers features for the generative prior and evidence heads. The transform design follows the sparse spatial interaction and adaptive frequency modeling principles of SAAF (Ma et al., 2026). All evidence used by CFIC is recovered from quantized latents.

Analysis transform. The analysis transform g_a progressively downsamples the image using gated frequency processing, lightweight spatial saliency inspired by channel and spatial attention (Woo et al., 2018), and low-resolution local–global attention. Its output is quantized to form the main transmitted latent $\hat{\boldsymbol{y}}$. Gated frequency blocks separate a local low-pass component from a high-frequency residual and modulate both with content-dependent gates; they do not use a fixed wavelet or Fourier transform. The lowest-resolution attention stage follows the sparse local–global interaction used in SAAF (Ma et al., 2026).

Hyperprior and entropy model. The hyper-analysis transform produces a hyper latent, which is quantized and coded using an entropy bottleneck. Hyper-synthesis then recovers context for the main-latent probability model. The entropy model combines this hyper context with previously decoded latent context to predict Gaussian parameters (μ, σ) for arithmetic coding. We use the

AEIC-style checkerboard schedule (Zhang et al., 2026a), following checkerboard context modeling (He et al., 2021), and transmit both the main and hyper latents. A shared context transform applies adaptive frequency gating, spatial saliency, and compact context blocks before predicting the Gaussian parameters.

Synthesis transform and output heads. The synthesis transform reconstructs the prior feature and exposes intermediate features at several spatial scales. Evidence heads project these intermediate features into the coordination interfaces. The terminal evidence feature is used by the final synthesis-side coordination module and the image decoder's evidence gate. The prior head produces the input feature for one-step generative reconstruction. No additional coded variables are introduced by these heads.

A.2 Compact Generative Prior

The compact prior retains more capacity in the shallow contracting path and higher-resolution synthesis stages of the U-Net backbone (Ronneberger et al., 2015), while reducing deep, low-resolution computation. Down0, Up2, and Up3 retain comparatively more capacity; Down3, the middle block, and Up0 are pruned aggressively. Selected operators in the retained stages are replaced by separable residual blocks, lightweight convolutions, or linear attention. The allocation follows the reconstruction-sensitivity diagnostic in Section 3.2 and the zero-step pruning results in Table 3.

Separable residual blocks. The lightweight residual block follows

$$\begin{aligned} &\mathrm{GN} \to \mathrm{SiLU} \to \mathrm{PWConv}_{1\times1} \to \mathrm{DWConv}_{3\times3} \to \mathrm{GN} \\ &\quad \to \mathrm{SiLU} \to \mathrm{DWConv}_{3\times3} \to \mathrm{PWConv}_{1\times1} + \mathrm{shortcut}. \end{aligned} \tag{6}$$

Depthwise and pointwise convolutions follow established lightweight designs (Howard et al., 2017; Chollet, 2017). Retained channels are selected using the joint ℓ_1 importance of the original residual projections. A low-rank factorization of pretrained spatial kernels initializes the depthwise and pointwise components, and existing shortcuts are copied when their interfaces are unchanged.

Linear attention. The replacements preserve the input and output interfaces while reducing the internal width. Query, key, value, output-projection, normalization, and feed-forward weights are initialized from importance-selected submatrices. Attention is evaluated as

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}(Q;d)\left[\mathrm{softmax}(K;n)^\top V\right], \tag{7}$$

where d and n denote the feature and token dimensions. This factorized form is related to linear-attention designs (Katharopoulos et al., 2020; Feng et al., 2025) and uses the query–key–value interface of standard attention (Vaswani et al., 2017).

A.3 Synthesis-Side Bit-Prior Coordination

At each selected synthesis scale, a coordination module receives the current synthesis feature $\boldsymbol{h}^u$, the corresponding native skip feature $\boldsymbol{h}^d$, and decoded evidence $\boldsymbol{e}$. The modules share the same residual formulation, with scale-dependent capacity and residual coefficients:

$$\begin{aligned} \boldsymbol{q} &= \mathrm{Conv}_{1\times1}(\mathrm{GN}(\boldsymbol{h}^u)), \qquad \boldsymbol{k} = \mathrm{Conv}_{1\times1}(\mathrm{GN}(\boldsymbol{h}^d)), \\ \boldsymbol{e}^{\mathrm{inn}} &= \mathrm{Conv}_{1\times1}(\boldsymbol{e}) - \mathrm{Conv}_{1\times1}(\boldsymbol{h}^d), \\ \boldsymbol{z} &= \mathrm{CrossAttn}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{e}^{\mathrm{inn}}), \\ \boldsymbol{z} &\leftarrow \mathrm{Conv}_{1\times1}(\mathrm{PWConv}(\mathrm{SiLU}(\mathrm{DWConv}_{3\times3}(\boldsymbol{z})))), \\ \tilde{\boldsymbol{h}}^u &= \boldsymbol{h}^u + m_s \tanh(\alpha)\boldsymbol{z}. \end{aligned} \tag{8}$$

The projection layers have separate learned weights. The innovation $\boldsymbol{e}^{\mathrm{inn}}$ is a learned feature residual between projected evidence and native context. Cross-attention uses the factorization in Eq. 7, with queries from the synthesis feature, keys from native context, and values from the innovation. A lightweight convolutional path processes the attention output before residual addition. The scalar $m_s \tanh(\alpha)$ bounds the residual coefficient, rather than the full feature update. These modules

operate alongside the original synthesis path and skip connections (He et al., 2016); the deployed model retains only the compact coordination paths. The terminal module additionally gates the attention output and adds a separately scaled residual from linear projections of the native context and decoded evidence.

A.4 Image-Decoder Interface

Latent Interface Steering operates on the wide latent ℓ_0 produced by the compact generative prior. The image decoder's main pretrained body remains frozen, while its wide input projection and lightweight injection modules are trainable. A projection of ℓ_0 forms the injection state ρ; the evidence feature supplies a separate magnitude statistic for gating.

Hierarchical injection is applied at early and intermediate decoder stages. Each controller combines the resized injection state, the current decoder state, high-frequency responses of both, and the evidence statistic. State and texture projections, followed by depthwise and pointwise processing, produce a gated residual Δ_i:

$$z_i' = z_i + \gamma_i \tanh(\alpha_i)\Delta_i. \tag{9}$$

Here, $\gamma_i \tanh(\alpha_i)$ bounds the residual coefficient. The evidence statistic is computed from the finest evidence feature:

$$u = \frac{\mathrm{mean}_c(|e_8^{\mathrm{tail}}|)}{\mathrm{mean}_{h,w}\,\mathrm{mean}_c(|e_8^{\mathrm{tail}}|)}. \tag{10}$$

The evidence magnitude is detached when computing this statistic, so gradients through the gate do not update the evidence projection. The resulting gate modulates the residual contribution at each injection stage.

B Training, Deployment, and Evaluation

B.1 Stage-Wise Optimization

All trainable components use AdamW (Kingma & Ba, 2014; Loshchilov & Hutter, 2017), with an effective batch size of 8 throughout training. The procedure comprises full-model pretraining, asymmetric compaction, and rate-specific adaptation; the last stage includes high-resolution fine-tuning.

Stage 1: full-model pretraining. The unpruned compression–generative model is trained on MLIC-Train-100K for 250K steps using random 512×512 crops. It serves as the teacher for compaction. We adopt the OCEA-DISTS objective from Zhang et al. (2026a) and optimize

$$\mathcal{L}_{\mathrm{RDP}} = \lambda_R \mathcal{L}_{\mathrm{Rate}} + \lambda_M \mathcal{L}_{\mathrm{MSE}} + \lambda_L \mathcal{L}_{\mathrm{LPIPS}} + \lambda_D \mathcal{L}_{\text{OCEA-DISTS}}. \tag{11}$$

Stage 2: asymmetric compaction and annealed distillation. We initialize the compact topology from the full teacher. During transfer, each frozen heavy operator M_H is paired with its trainable lightweight replacement M_L:

$$y = \sigma M_H(x) + (1-\sigma) M_L(x), \qquad \sigma : 1 \to 0. \tag{12}$$

The heavy contribution is gradually reduced to zero. Distillation follows teacher–student transfer (Hinton et al., 2015) and combines local-feature, multi-scale feature, and output matching, summarized by $\mathcal{L}_{\mathrm{KD}}$. Together with reconstruction supervision, the objective is

$$\mathcal{L}_{\mathrm{Stage2}} = \lambda_M \mathcal{L}_{\mathrm{MSE}} + \lambda_L \mathcal{L}_{\mathrm{LPIPS}} + \lambda_D \mathcal{L}_{\text{OCEA-DISTS}} + \lambda_{\mathrm{KD}} \mathcal{L}_{\mathrm{KD}}. \tag{13}$$

The final recovery phase removes the teacher constraints and optimizes the lightweight student alone. Stage 2 runs for 20K steps, with 5K steps each allocated to annealing the upsampling and downsampling paths and the remaining 10K steps to recovery training.

Stage 3-1: rate-specific fine-tuning. For each target rate, we load the compact student, merge the previous adapters, and attach fresh low-rank adapters to the U-Net. Fine-tuning runs for 20K steps using 512×512 crops and the objective

$$\mathcal{L}_{\text{Stage3-1}} = \lambda_R \mathcal{L}_{\mathrm{Rate}} + \lambda_M \mathcal{L}_{\mathrm{MSE}} + \lambda_D \mathcal{L}_{\text{OCEA-DISTS}} + \lambda_P \mathcal{L}_{\mathrm{adv}}. \tag{14}$$

Adapters are applied to attention, feed-forward, convolutional, shortcut, and lightweight pointwise projections. The trainable components also include the generative-prior input and output projections, codec, evidence projections, coordination modules, wide image-decoder entrance, and hierarchical injection modules. The remaining pretrained base weights stay frozen.

Stage 3-2: high-resolution fine-tuning. Starting from the corresponding Stage 3-1 checkpoint, we fine-tune on DF2K and CLIC Train for 5K steps with 1024×1024 crops, following the data protocol in Section 4.1. The effective batch size remains 8. We use the same objective,

$$\mathcal{L}_{\text{Stage3-2}} = \lambda_R \mathcal{L}_{\text{Rate}} + \lambda_M \mathcal{L}_{\text{MSE}} + \lambda_D \mathcal{L}_{\text{OCEA-DISTS}} + \lambda_P \mathcal{L}_{\text{adv}}, \tag{15}$$

with a lower learning rate and reduced relative weights on the rate and MSE terms. The resulting rate-specific checkpoints are used to evaluate all three datasets: Kodak, DIV2K, and CLIC2020.

B.2 Deployment and High-Resolution Inference

For deployment, we retain only the compact topology and lightweight replacement branches, remove all training-only heavy modules, merge the final low-rank adapters, and remove their wrappers. Reported parameter counts refer to the complete deployed models.

High-resolution inputs are padded to a compatible spatial size and cropped back to their original dimensions after reconstruction. The generative prior uses overlapping latent tiles, with evidence crops spatially aligned to each tile. The image decoder is not tiled by default. This internal tiling procedure is distinct from the patch extraction used for FID evaluation. We apply quantized AdaIN color correction following Zhang et al. (2025), with the color-statistics overhead included in the reported bitrate.

B.3 Evaluation and Ablation Protocols

All methods are evaluated on complete images from Kodak, DIV2K validation, and CLIC2020 Professional Test. Baseline results come from public checkpoints evaluated under this protocol or the corresponding papers' official results, as specified in Section 4.1. FID on DIV2K and CLIC2020 is computed from overlapping 256×256 patches extracted from full-image reconstructions and references.

Component ablations are retrained under matched training conditions and retain the same bitstream structure. The pruning comparison uses the same lightweight operators and heavy-to-light transfer procedure for asymmetric and uniform topologies at similar total parameter counts. Before-recovery measurements record the immediate effect of structural modification; after-recovery measurements compare the resulting models under matched recovery training.

B.4 BD-Rate Computation

We follow the original Bjøntegaard rate formulation (Bjontegaard, 2001). For a method m and anchor a, let $f_m(q)$ and $f_a(q)$ be cubic fits of $\ln R$ against metric value q. Over the overlapping measured interval $[q_{\min}, q_{\max}]$,

$$\Delta = \frac{1}{q_{\max} - q_{\min}} \int_{q_{\min}}^{q_{\max}} [f_m(q) - f_a(q)]\, dq, \qquad \text{BD-Rate} = \left(e^{\Delta} - 1\right) \times 100\%. \tag{16}$$

The definition applies to PSNR, LPIPS, and DISTS because rates are compared at matched metric values, regardless of whether higher or lower metric values are preferred. We do not extrapolate beyond the shared measured range. Negative BD-rate denotes rate savings relative to the anchor; methods without sufficient overlap are omitted from BD-rate comparisons but remain in the curves.

C Additional Analysis

C.1 Reconstruction-Sensitivity Diagnostic

The reconstruction-sensitivity analysis is conducted on Set14 (Zeyde et al., 2012). The CFIC sensitivity sweep starts from the full, adapted model before structural pruning. Each perturbation reduces one

stage to 50% of its channels while keeping the decoded latents and all unaffected parameters fixed. The resulting LPIPS and DISTS changes measure the immediate effect on reconstruction, without recovery training. Figure C.1 reports this stage-wise profile alongside the reference StableCodec backbone.

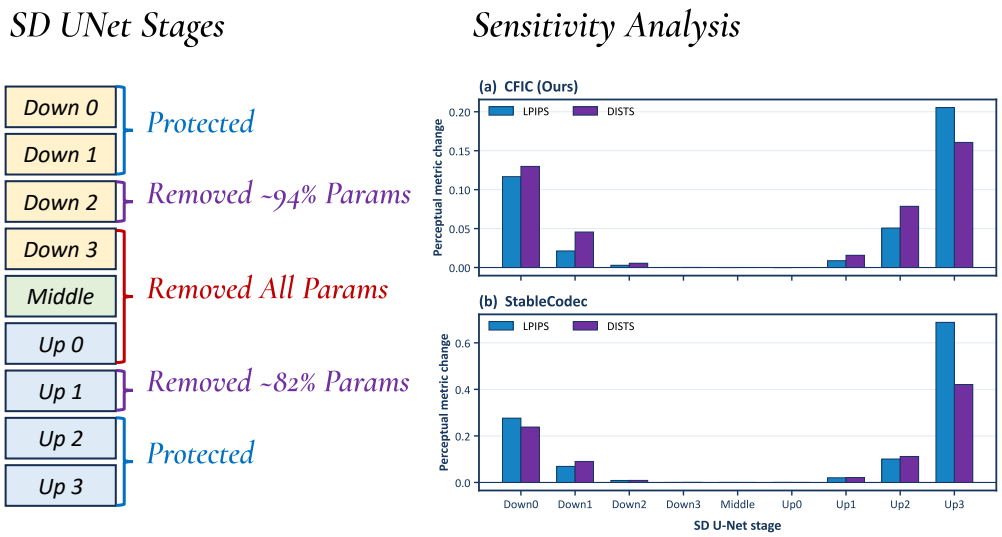


Figure C.1: **Reconstruction-sensitivity diagnostic and asymmetric allocation.** Right: stage-wise LPIPS and DISTS changes. The CFIC sweep starts from the full, adapted model, with each tested stage reduced to 50% of its channels and no recovery training. Left: the resulting asymmetric allocation, with stronger reduction in deep computation and more capacity retained in synthesis stages.

The plotted profiles vary substantially across depth, with smaller changes at deep, low-resolution stages and larger changes at later synthesis stages. We use these observations as empirical motivation for asymmetric capacity allocation. The topology is fixed for subsequent heavy-to-light transfer and recovery training, as described in Appendix B.1.

C.2 Residual-Update Visualization

Figure C.2 shows the channel-wise root-mean-square magnitude of updates from the final synthesis-side coordination module and the terminal image-decoder injection. The maps locate changes to intermediate features; stronger responses occur around structured regions rather than uniformly across the image.

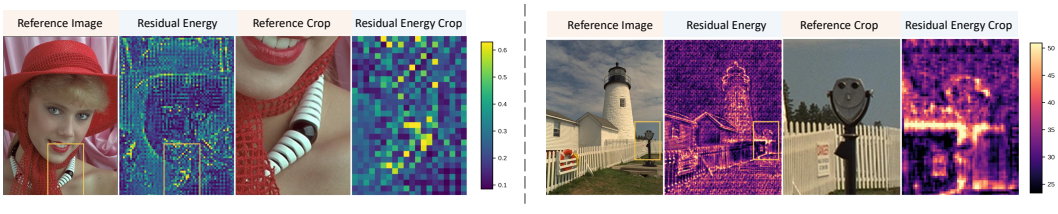


Figure C.2: **Residual-update magnitudes.** Left: the update from the final bit-prior coordination module at the $1/8$ synthesis stage. Right: the update from the terminal image-decoder injection. Brighter regions indicate larger channel-wise RMS magnitude, $\mathrm{RMS}(\Delta h)$. Zoomed crops correspond to highlighted regions in the reference images.

D HIGH-RESOLUTION QUALITATIVE COMPARISONS

Figures D.1–D.4 provide full-frame comparisons at low bitrates. Each example uses the same zoom region across methods, with the exact image-level bits per pixel printed in the panels. Rates differ between methods, so these examples should be read together with the quantitative curves in the main paper.

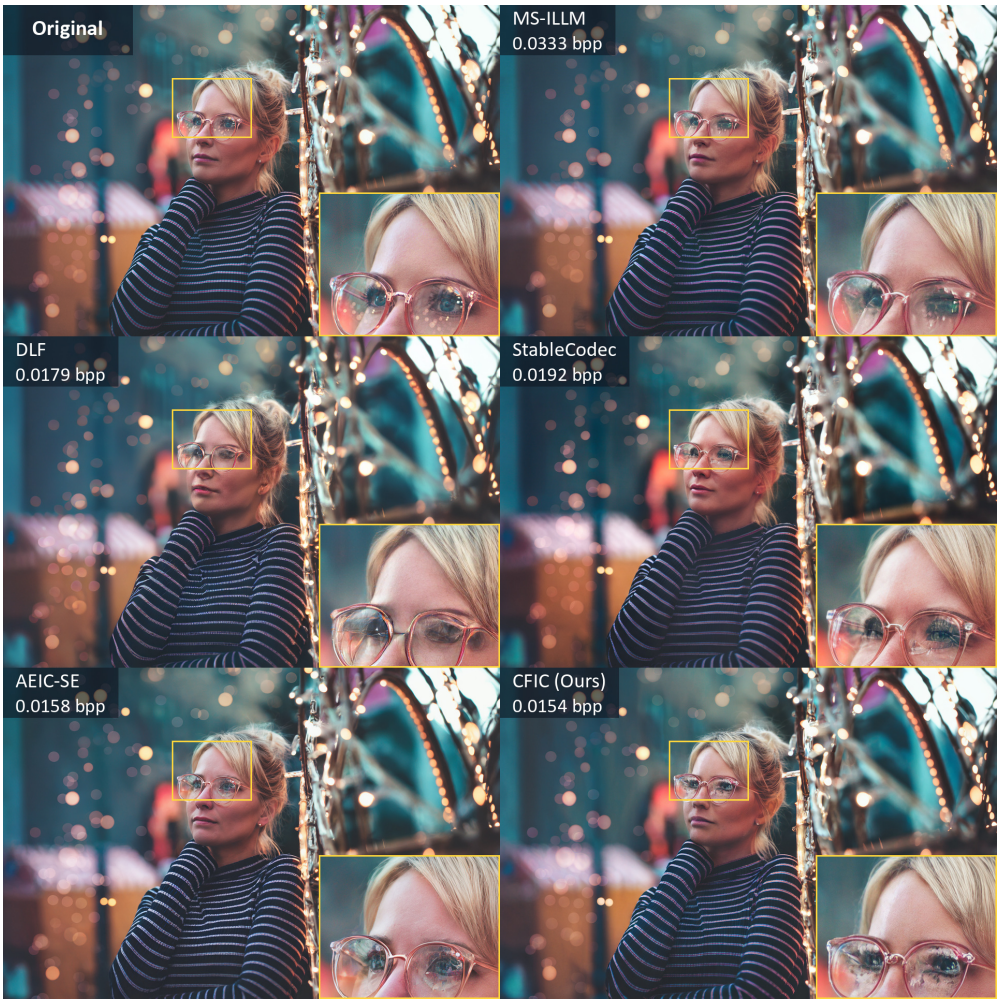


Figure D.1: High-resolution qualitative comparison on CLIC2020.



Figure D.2: High-resolution qualitative comparison on CLIC2020.



Figure D.3: High-resolution qualitative comparison on DIV2K.



Figure D.4: High-resolution qualitative comparison on DIV2K.

E LIMITATIONS AND FUTURE WORK

The sensitivity profile and compact topology have been studied for the source-conditioned backbone used here. Their transfer to other diffusion backbones, image domains, and bitrate ranges remains untested. Pruning and distillation still require the full model during training, so lower deployment cost does not imply lower training cost. The experiments focus on natural images at ultra-low bitrates and do not evaluate large distribution shifts or temporal data.

Future work includes automated topology selection, content-adaptive capacity allocation, and extensions to other generative priors and video compression. Semantic consistency and robustness outside the training distribution also warrant further study.