PER-IMAGE LATENT OPTIMIZATION AT EXTREME BITRATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural codecs compress with an amortized encoder: a single network, trained to perform well on average, encodes every image in one forward pass. Per-image latent rate–distortion optimization (RDO) instead refines the transmitted latent for each individual image at encoding time. It is commonly regarded as a marginal refinement, worth a few tenths of a dB at conventional rates, but that assessment was formed when bits were plentiful. Generative decoders now sustain reconstruction at bitrates orders of magnitude lower, where every bit the encoder misallocates is a visible fraction of the budget. We show that in this regime the value of per-image optimization is not a constant of the technique but grows systematically as the operating rate falls. We measure the gain as a rate-equivalent margin, the additional bits the amortized encoder requires on the same image's own rate ladder to match the optimized quality. Across three architecturally distinct codecs whose ladders span three orders of magnitude in rate, the median margin rises as rate decreases and saturates at a backbone-specific ceiling. At the lowest rates, the amortized encoder needs 1.7 to 2.2 times its bits to match per-image optimization, corresponding to median per-image BD-rate savings of 34–50%. A continuous-relaxation analysis traces this growth to two coupled trends: the per-image suboptimality of the amortized encoder widens steadily as rate falls, and the share of it that discretization-aware optimization captures becomes meaningful only at the lowest rates. The gains transfer to held-out perceptual metrics the optimizer never saw, precisely in the low-rate regime; optimizing a mismatched objective instead erases or reverses them. Practically, the gain saturates in compute: doubling the standard optimization budget adds at most a relative 12%, so a modest budget suffices whenever content is encoded once and decoded many times. Together, these results identify the amortized encoder, rather than the decoder, as a principal bottleneck of neural compression at extreme rates.

1 INTRODUCTION

Neural image compression below 0.03 bits per pixel squeezes a two-megapixel photograph into a few thousand bytes, a compression ratio of roughly $3000{:}1$. The decoder is what makes this regime viable. Diffusion- and GAN-based decoders synthesize plausible texture from a latent far too coarse to specify it (Careil et al., 2024; Theis et al., 2022), and progress at extreme rates has largely been a history of better decoders. The encoder has changed far less and remains almost universally amortized. A single feedforward network, trained to minimize an average objective over a dataset, produces each image's latent in one pass, and whatever that pass gets wrong for a particular image is sealed into the bitstream.

A codec that leaves training is not yet the codec that is deployed, and the difference has three parts. The objective differs, because deployment is scored by a criterion that the training loss only approximates. The granularity differs, because an encoder trained to be good on average is suboptimal for any single input, the amortization gap of variational inference (Cremer et al., 2018; Yang et al., 2020). The arithmetic differs, because a real file consists of hard-rounded integers priced by a true entropy coder, while training sees smooth approximations of both, the discretization gap (Yang et al., 2020). Per-image latent rate–distortion optimization (RDO) addresses all three at encoding time. With the decoder and entropy model frozen, the transmitted latent itself is optimized for the image at hand, under the deployed objective, through a discretization-aware relaxation such

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.

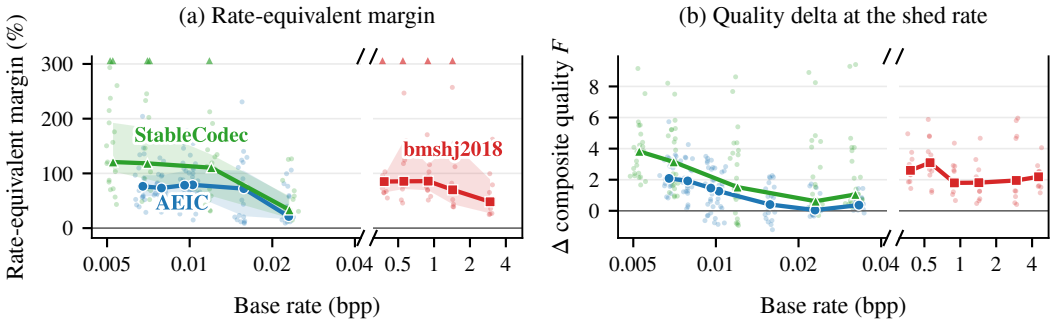


Figure 1: **The divergence of per-image optimization from amortized encoding.** (a) Median per-image rate-equivalent margin versus base rate, the extra bits the amortized encoder needs on the same image's own base ladder to match the optimized quality. Lines are medians with bootstrap 95% bands, light points are individual images, and the few points above 300% are pinned at the top edge. (b) Median composite-quality delta at the rate the optimized latent actually codes to, typically 20–40% below base (Section 3.3). The axis is broken between the extreme-rate ladders (AEIC, StableCodec) and the conventional-rate ladder (bmshj2018). Within every ladder the margin grows as rate falls and saturates at a backbone-specific ceiling.

as stochastic Gumbel annealing (Yang et al., 2020; Campos et al., 2019). Nothing about the decoder changes, only the file does.

The worth of this per-image correction has effectively been established only at conventional rates. Measured between 0.1 and 1 bits per pixel, under the codec's own training objective, per-image RDO yields a modest and essentially flat bonus of a few tenths of a dB (Campos et al., 2019; Yang et al., 2020; Xu et al., 2023), and this verdict has confined the technique to a niche. The verdict was reached, however, in a regime where the amortized encoder has bits to spare, and extreme-rate compression inverts that premise. With a few thousand bits for an entire image, misallocations that were once negligible now consume a visible fraction of the budget, and an encoding that is nearly optimal on average has no particular reason to remain nearly optimal for each image. Whether the value of per-image optimization is a constant of the technique or a function of the operating rate thus decides whether encoder-side optimization is a minor refinement or a central lever of extreme-rate compression. We are aware of no systematic measurement of this rate dependence.

This paper provides that measurement. On every codec we test, the gap between the amortized encoder and what per-image optimization achieves widens as the codec descends its own rate ladder and then saturates. Measuring the effect faithfully requires care, because optimized latents not only gain quality but also shed 20 to 40% of their bits, so quality deltas at nominally matched rates understate the effect. Our yardstick is therefore a per-image *rate-equivalent margin*, the extra bits the base codec needs, interpolated on the same image's own rate ladder, to match the optimized quality (Section 3). All rates are true entropy-coded file sizes, and all headline comparisons use matched image sets at every rung. The measurements cover three architecturally distinct codecs, introduced in Section 3.1, whose ladders span three orders of magnitude in rate, under one uniform protocol. Three results form the core of the paper.

- **The rate-equivalent margin of per-image optimization over the amortized encoder widens as rate falls, on every backbone tested (Section 4).** Within each codec's ladder the median margin grows from roughly 20–50% at the top to 72–121% at the bottom (Figure 1), where the amortized encoder needs up to 2.2× its bits to match per-image optimization. Median per-image BD-rate savings reach 34–50%. The growth is driven by quality rather than by bit shedding, the ceiling moves little when the optimization budget is doubled, and the shape replicates on a second dataset.
- **The margin's growth has a measurable mechanism (Section 5).** A continuous-relaxation bound separates how much per-image headroom the codec has from how much of it the optimizer captures. The headroom grows steadily as rate falls, the captured share becomes meaningful only at the lowest rungs, and the widening margin is the product of the two. Rounding the continuous optimum naïvely scores *below* the amortized baseline, so

discretization-aware optimization is the enabling ingredient rather than an implementation detail.

- **The quality gains of per-image optimization survive held-out metrics they were never optimized for (Section 6).** Judged by metrics excluded from the objective, including a distributional patch-FID, quality improves precisely in the low-rate regime where the margin is large and fades or reverses at high rates. Two no-reference aesthetics metrics dissent, and we report their dissent as a measure of the gain's metric-specific component. The gains are also objective-specific, since optimizing PSNR instead of the deployed perceptual objective at the same extreme rate reduces composite quality by nearly four points. The deployed objective, not a proxy, must be the one optimized.

The compute-exchange probes of Section 4.3 add a practical corollary. The gain saturates in compute: doubling the standard encoding budget adds at most a relative 12%, while a quarter of the budget still captures roughly a quarter of the gain. Beyond the numbers, the results reframe per-image RDO. Its value is a property not of the optimizer but of the distance between a codec's amortized encoder and its per-image frontier, and below roughly 0.01 bits per pixel that distance makes it one of the largest levers available without touching the decoder.

2 Related work

Learned image compression. End-to-end learned codecs pair an analysis and synthesis transform with a learned entropy model, trained on a rate–distortion Lagrangian (Ballé et al., 2017; Theis et al., 2017). The scale hyperprior (Ballé et al., 2018) and its context-model extensions (Minnen et al., 2018; Cheng et al., 2020; He et al., 2022) remain the backbone of most systems, including the CompressAI implementations (Bégaint et al., 2020) we use. This line optimizes the *expected* objective over a training set, so the encoder is amortized in exactly the sense our measurements interrogate, and we ask how far it sits from its own per-image optimum as a function of rate.

Extreme-rate generative compression. Below roughly 0.05 bpp, reconstruction becomes generation. Diffusion- and GAN-decoder codecs synthesize texture conditioned on a compact latent, including text- or latent-conditioned diffusion codecs (Careil et al., 2024; Lei et al., 2023), direct diffusion-based coding (Theis et al., 2022), and one-step-diffusion decoders that make such systems practical (Sauer et al., 2024). Two of our three backbones belong to this family: AEIC, the reference codec of an ultra-low-bitrate coding challenge (Ling et al., 2026), and StableCodec (Zhang et al., 2025), both introduced in Section 3.1. Prior work in this regime concentrates on decoder-side realism and rate control, while the encoder remains amortized throughout. Our contribution is orthogonal, quantifying on frozen instances of such codecs how much achievable per-image quality the amortized encoder leaves on the table, and showing that this quantity grows as rate falls.

Per-image and inference-time optimization. Content-adaptive encoding refines latents or encoder parameters per image at encoding time (Campos et al., 2019; Lu et al., 2020; Zou et al., 2022). Yang et al. (2020) formalized the discretization obstacle and proposed stochastic Gumbel annealing (SGA), decomposing the encoder's suboptimality into amortization and discretization gaps at conventional rates, with follow-ups refining the annealing and extending to video (Xu et al., 2023). Adjacent strategies adapt more than the latent. Instance-adaptive fine-tuning transmits a model update (van Rozendaal et al., 2021), implicit neural codecs make the entire representation per-image (Dupont et al., 2021), and GAN-based perceptual codecs (Mentzer et al., 2020) shape the operating points from which such adaptation starts. These works report gains at a few operating points, typically between 0.1 and 1 bpp, and treat the gain as roughly constant. We complement this line with a rate-equivalence methodology that remains meaningful when optimization sheds a large fraction of the bits, a systematic sweep over each codec's full ladder that reveals a rate dependence invisible at isolated points, and a mechanism measurement showing that the total per-image slack itself grows as rate falls, a property of the codecs rather than of SGA.

Amortization gaps in inference. The gap between amortized and instance-optimal inference is well studied for VAEs (Cremer et al., 2018; Marino et al., 2018), where per-sample refinement of the posterior closes part of the ELBO gap. Neural codecs are VAEs operating under a discretization

constraint at test time (Ballé et al., 2017), and this changes the picture qualitatively. We find the continuous, VAE-style refinement bound far from attainable, since its rounded projection scores below the amortized baseline. The practically relevant gap is the discretization-aware one, and its rate dependence has not previously been mapped.

3 MEASUREMENT SETUP

3.1 CODECS AND RATE LADDERS

We measure three frozen codecs chosen to span architecture families. **AEIC** is a perceptually fine-tuned one-step-diffusion codec, released as the reference codec of the LoViF ultra-low-bitrate challenge (Ling et al., 2026); we use it as a frozen public artifact. An analysis transform produces a latent that is entropy-coded with a masked-group autoregressive model, and a single-step diffusion decoder (Sauer et al., 2024) synthesizes the reconstruction. We use seven public rate-ladder checkpoints spanning 0.007–0.033 bpp, indexed A1–A7 from high to low rate. **bmshj2018** is the scale-hyperprior codec of Ballé et al. (2018) in the CompressAI implementation (Bégaint et al., 2020). Qualities $q \in \{1, 2, 3, 4, 6, 8\}$ span 0.4–4.5 bpp. The model is trained for MSE, which makes it a deliberately mismatched, distortion-oriented backbone. **StableCodec** (Zhang et al., 2025) couples a VAE latent with an ELIC-style (He et al., 2022) four-group masked autoregressive latent codec and a one-step diffusion decoder. Five fine-tuned checkpoints, indexed S1–S5 from high to low rate, span 0.005–0.032 bpp.

Evaluation images are a fixed 24-image subset of the challenge's publicly available validation set at 2K resolution, plus Kodak-24 (Eastman Kodak Company, 1993) for cross-dataset replication. Quality is the challenge composite

$$F(\hat{\mathbf{x}}, \mathbf{x}) \;=\; \mathrm{PSNR} + 10 \cdot \text{MS-SSIM} + 20\,(1 - \mathrm{LPIPS}) + 25\,(1 - \mathrm{DISTS}), \tag{1}$$

a weighted blend of distortion and perceptual terms (Zhang et al., 2018; Ding et al., 2022; Wang et al., 2003) reflecting the perception–distortion tradeoff (Blau & Michaeli, 2018) and representative of how extreme-rate codecs are scored in practice. All rates are *true entropy-coded bitstream sizes*, range-coded and including the hyper-latent and header, never proxy likelihoods.

3.2 PER-IMAGE LATENT RDO PROTOCOL

For each image and rung, the encoder output $\mathbf{y}_0 = g_a(\mathbf{x})$ is refined by optimizing an additive perturbation $\boldsymbol{\delta}$ with the decoder, entropy model, and hyper-latent frozen. Discretization is handled by stochastic Gumbel annealing (Yang et al., 2020). With residual $r = \mathbf{y}_0 + \boldsymbol{\delta} - \boldsymbol{\mu}$, temperature-$\tau$ stochastic rounding $s_\tau(r)$ interpolates between the floor and ceiling grid points with Gumbel-perturbed logits, and τ anneals from 0.5 toward 0.2. The objective is the deployed score of Eq. (1), made differentiable with the implementations of Chen & Mo (2022), plus a two-tier rate penalty

$$\mathcal{L}(\boldsymbol{\delta}) = -F\big(g_s(s_\tau(r) + \boldsymbol{\mu}), \mathbf{x}\big) + \lambda(R_{\mathrm{hard}}) \cdot R_{\mathrm{soft}}, \qquad \lambda(R) = \begin{cases} \lambda_0 & R \le R_{\mathrm{base}} \\ 8\lambda_0 & \text{otherwise,} \end{cases} \tag{2}$$

which drives every optimized latent toward or below its own image's base rate. The standard budget is 600 annealed steps plus 150 straight-through (Bengio et al., 2013) polish steps, repeated for two re-annealing rounds. Candidate states are entropy-coded with the real range coder at regular intervals, and the best true-coded state within a 2% tolerance of the base rate (at most $1.02\times$) is kept. The identical protocol, hyperparameters included, is applied to all three codecs and all rungs, at a cost of 0.5–3 GPU-hours per image at 2K resolution (Figure 3 prices this compute exchange, analyzed in Section 4.3).

3.3 THE RATE-EQUIVALENT MARGIN

Optimized latents do not merely gain quality. Because Eq. (2) keeps rate at or below base, they also shed 20–40% of their bits, and at several mid-ladder rungs the quality delta is near zero while the bitstream shrinks by a third. Quality deltas at nominally matched rates therefore understate the effect. Our primary yardstick removes this ambiguity. For image i at rung k, let $(R_j^{(i)}, F_j^{(i)})_j$ be the

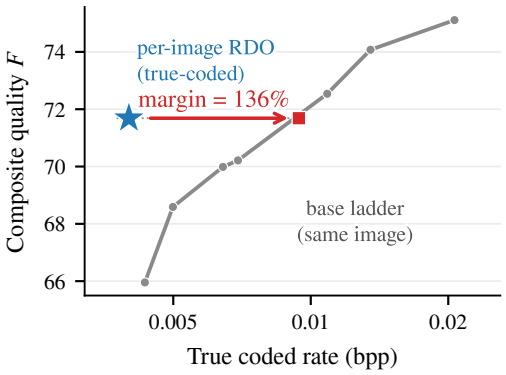


Figure 2: **Margin construction** on one image (AEIC, lowest rung). The horizontal distance from the optimized point (star) to the same image's base ladder (gray), relative to the optimized rate, is the rate-equivalent margin.

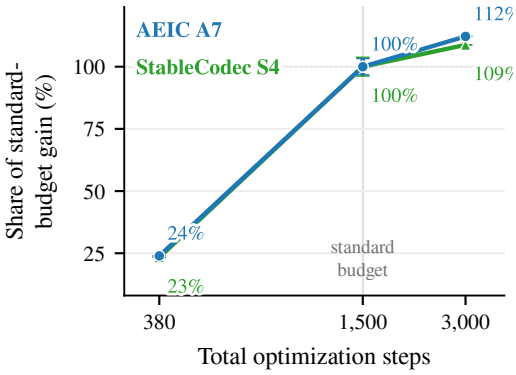


Figure 3: **Compute exchange** on the two probed rungs (AEIC A7, StableCodec S4): median gain as a fraction of the standard-budget gain, versus total optimization steps. Error bars give the restart spread at the standard budget.

image's *own* base ladder over all rungs, and let $(R^{(i)}_{\text{sga}}, F^{(i)}_{\text{sga}})$ be the optimized operating point. The **rate-equivalent margin** is

$$M^{(i)}_k = \frac{R_{\text{eq}} - R^{(i)}_{\text{sga}}}{R^{(i)}_{\text{sga}}}, \qquad \text{where } F_{\text{base}}(R_{\text{eq}}) = F^{(i)}_{\text{sga}}, \tag{3}$$

and $F_{\text{base}}(\cdot)$ interpolates the base ladder piecewise-linearly, taking the smallest rate achieving the target quality if the ladder is locally non-monotone. The margin is the fraction of extra bits the amortized encoder needs to match the optimized quality of the same image; Figure 2 illustrates the construction on a real image, where the optimized point also codes below the entire base ladder. Because the denominator is the optimized rate, rate shedding alone produces a nonzero margin even at zero quality gain, so we also report per-image BD-rate (Bjøntegaard, 2001) between the two ladders, which integrates over the shared quality range and is shed-insensitive (Section 4.5). Margins requiring extrapolation beyond the ladder are flagged and excluded from primary statistics. Base-ladder cells exhibiting corrupted evaluations, 4 of roughly 700 image-by-rung cells (Appendix B), are excluded and disclosed.

Statistical protocol. The measured grids are AEIC 7×24, bmshj2018 6×12 on a fixed half of the image set to limit compute, and StableCodec 5×24, all complete. All headline tables further restrict every rung of a backbone to the identical, matched image subset, eliminating composition drift, with matched counts of 22, 11, and 24 after intersecting rungs and excluding the corrupted cells above. Margins are heavy-tailed to the right, since single steep-ladder images push means to several times the median, so we report *medians* with bootstrap 95% confidence intervals over 1,000 resamples of images, retaining means in released summaries. Monotone rate dependence is tested with Page's trend test over the matched image-by-rung matrix.

4 The Divergence Law

4.1 Within-Ladder Growth on Three Backbones

Figure 1 and Table 1 give the central result. Within every backbone's own rate ladder, the median rate-equivalent margin rises as rate falls and then saturates. We refer to this empirical regularity as the *divergence law*. On AEIC it grows from 21% at the highest interpolable rung to a plateau of 72–79% below 0.016 bpp. On bmshj2018 it grows from 48% to 85–86% below 1 bpp, margins that Section 6.2 traces to objective retargeting. On StableCodec it grows from 33% to 110–121% at the lowest rungs, where the amortized encoder needs $2.2\times$ its bits to match per-image optimization. All medians use interpolated margins only (extrapolated cases dropped, counts in Table 1), and top-ladder rungs where optimized quality exceeds the ladder ceiling for most images show quality deltas only. Page's trend test rejects rate independence separately on each backbone, with one-sided

Table 1: The divergence law on matched image sets. Margin medians (Eq. (3)) with bootstrap 95% CIs over *interpolated* per-image margins (n_{int} of n matched). Pareto-3 is the fraction of images improving PSNR, LPIPS, and DISTS at no more bits. Rungs where margins are extrapolated for almost all images show ΔF only. Bold marks each backbone's largest margin median and largest Pareto-3 fraction.

Backbone (n)	Rung	Base bpp	Margin med. [CI]	n_{int}	ΔF med.	Pareto-3
AEIC (22)	A1	0.0332	—	1	+0.37	0.45
	A2	0.0230	21% [16, 57]	19	+0.05	0.09
	A3	0.0157	72% [22, 105]	22	+0.41	0.41
	A4	0.0103	**79%** [61, 108]	22	+1.26	0.59
	A5	0.0096	78% [62, 105]	22	+1.46	0.68
	A6	0.0079	73% [59, 97]	22	+1.92	0.91
	A7	0.0068	76% [58, 97]	22	+2.08	**1.00**
bmshj2018 (11)	q8	4.53	—	0	+2.19	**0.91**
	q6	2.95	48% [27, 89]	10	+1.95	0.18
	q4	1.43	70% [44, 156]	11	+1.81	0.11
	q3	0.89	**86%** [69, 116]	11	+1.80	0.11
	q2	0.56	86% [71, 159]	11	+3.10	0.00
	q1	0.38	85% [65, 96]	11	+2.60	0.11
StableCodec (24)	S1	0.0322	—	0	+1.05	0.54
	S2	0.0231	33% [18, 59]	15	+0.62	0.38
	S3	0.0120	110% [59, 134]	16	+1.52	0.67
	S4	0.0070	118% [89, 183]	18	+3.15	0.92
	S5	0.0053	**121%** [103, 192]	21	+3.82	**0.96**

$p \leq 4.3 \times 10^{-6}$ throughout. The per-image distributions shift as a whole rather than through outliers, and at the lowest rungs the great majority of images sit above a 50% margin. The quality gains arrive at *shed* rates (Figure 1b). On AEIC and StableCodec the median quality delta grows monotonically at the low end, and at their lowest rungs nearly every image improves PSNR, LPIPS, and DISTS simultaneously at no more bits (the Pareto-3 column of Table 1), while bmshj2018 instead trades PSNR for perceptual quality (Section 5.2). Figure 8 in the appendix shows qualitative examples at the lowest AEIC rung.

4.2 The growth is not a bit-shedding artifact

Because the margin's denominator is the optimized rate, pure bit shedding already produces a nonzero margin. If quality were unchanged, an optimized latent coding to a fraction of its base rate would show a *shed floor* of $R_{\text{base}}/R_{\text{sga}} - 1$. We therefore split every per-image margin into this floor plus a *quality-driven excess*. The floor is roughly flat in rate on all three backbones, with no monotone trend. The excess is what diverges. From the high end to the low end of each ladder, its median rises from 0% to 61% on AEIC, from 30% to 57% on bmshj2018, and from 18% to 115% on StableCodec. At the high-rate end the margin is almost entirely shed floor, while at the lowest rungs it is dominated by genuine quality headroom. BD-rate, which integrates over the shared quality range and is shed-insensitive by construction, corroborates this (Section 4.5).

4.3 The ceiling is a frontier property

Margin growth saturates near the bottom of each ladder. Saturation could reflect the codec's per-image frontier or merely an under-converged optimizer. Three probes on six matched images, run at the lowest AEIC rung (A7) and the second-lowest StableCodec rung (S4), separate these possibilities (Figure 3). Doubling the budget to $1{,}200{+}300$ steps per round improves the median gain by a relative 9–12%. Independent restarts reproduce the standard-budget result within $\sigma \approx 0.1$ points, an order of magnitude below the effect. Cutting the budget by a factor of four loses about three quarters of the gain. The ceiling therefore reflects the codec rather than the optimization, to first order, and the quoted ceilings are roughly 10% conservative.

4.4 Margins are image properties that track frontier distance

The per-image distributions (Figure 6, appendix) are wide and heavy-tailed, but they are not noise. Per-image margin *ranks* are strongly consistent across the three lowest rungs of every backbone,

with mean Spearman correlations between 0.83 and 0.98, so the margin is a stable property of the image and codec pair. Which images benefit differs by backbone. The hardest images benefit most on bmshj2018, the easiest on StableCodec, and AEIC shows no preference. Amortization errors are structured, and their structure is consistent with an architectural signature. At the same time, the law is a *within-ladder* law, and absolute bpp does not predict the margin across backbones. bmshj2018 at 0.9 bpp and AEIC at 0.010 bpp both sit near an 80% margin, and StableCodec exceeds both at every shared rate. The Kodak-24 replication shows the same dissociation. StableCodec reproduces the rising shape, bmshj2018's two interpolable rungs sit at challenge-comparable levels, and StableCodec's absolute levels drop on Kodak's smaller, cleaner images (Appendix C). Both observations fit one description: **the margin measures the distance between the amortized encoder and the codec's per-image frontier**, grown by descending the ladder and modulated by how hard the images stress the backbone. Section 5 measures this distance directly.

4.5 BD-RATE CORROBORATION

Integrated over the overlapping quality range of each image's two ladders, per-image BD-rate medians are $\mathbf{-34.0}\%$ [CI $-47.9, -27.2$] for AEIC, $\mathbf{-39.5}\%$ [$-53.3, -33.8$] for bmshj2018, and $\mathbf{-50.0}\%$ [$-71.2, -38.4$] for StableCodec. Splitting each ladder in half confirms the law in this shed-insensitive currency as well. On every backbone the low-rate half saves more bits than the high-rate half, most sharply on StableCodec (-48.8 versus -32.0%).

5 MECHANISM AND ANATOMY OF THE GAIN

5.1 GROWING SLACK AND LATE-ONSET CAPTURE

Two quantities could drive the growth of the margin: how much per-image headroom the codec has at each rung, and how much of that headroom discretization-aware optimization actually captures. We measure both on AEIC, the backbone with the densest ladder.

A continuous-relaxation bound on per-image headroom. For four AEIC rungs and six images we re-run the protocol of Section 3.2 with rounding removed. The perturbed latent is decoded continuously, with the soft likelihood rate. The resulting operating point is an idealized, uncodable frontier. No coder realizes a continuous latent, and at these rungs the soft rate is optimistic by a factor of about 2.5. Its quality nevertheless upper-bounds any per-image latent refinement through the frozen decoder. We call its gain over the amortized baseline the image's **slack**.

Slack grows and capture arrives late. Figure 4 shows both quantities with their per-image points. Median slack rises strictly as rate falls, from $+5.7$ composite points at the highest measured rung to $+13.7$ at the lowest. The amortized encoder's per-image suboptimality is therefore not constant. *The frontier itself retreats from the encoder as the ladder descends.* Meanwhile, the share of slack that true-coded SGA captures stays at 1–2$\%$ over the two highest rungs and then climbs to 14% at the lowest (Figure 4b). The divergence law of Section 4 is the product of these two effects, and neither alone accounts for it. We decompose one backbone, while the law itself is replicated on three.

Discretization locks most of the slack. Rounding the continuous optimum and entropy-coding it scores *below* the amortized baseline, shown as the red $\times$ in Figure 4a. The continuous solution drifts far off-grid, and naïve projection destroys it. Two consequences follow. First, discretization-aware optimization through annealed stochastic rounding is the enabling ingredient of per-image RDO, not a refinement, consistent with Yang et al. (2020) but measured here at rates where the effect is an order of magnitude larger. Second, the 86% of slack that remains uncaptured even at the lowest rung is locked behind quantization. Any method that closes part of this gap, whether through finer grids, learned per-image quantization, or distilling optimized latents into the encoder, has measured headroom of over ten composite points per image at extreme rates.

5.2 THREE ROUTES TO ONE LAW

Decomposing the median quality delta into the composite's four terms (Figure 7, appendix) shows that the three backbones achieve the same law by different means. AEIC's gains are *balanced*: its

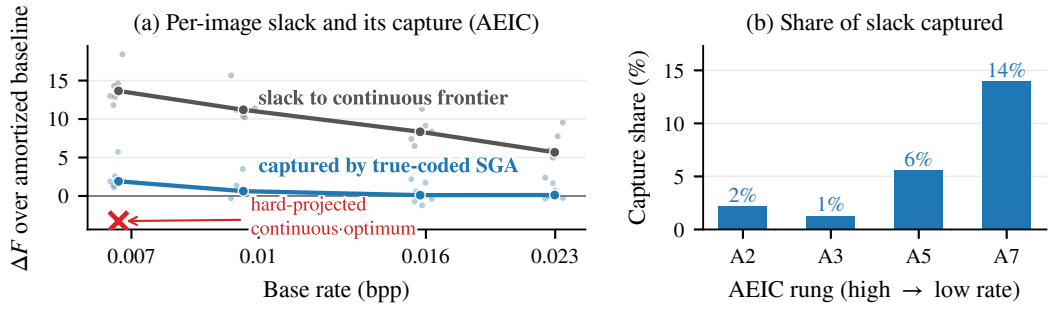


Figure 4: **Mechanism (AEIC, 4 rungs × 6 images).** (a) Per-image slack to the continuous frontier (gray) and the part captured by true-coded SGA (blue); the red × is the hard-projected continuous optimum, which lands *below* the baseline. (b) The captured share of slack rises from 1–2% to 14% as rate falls.

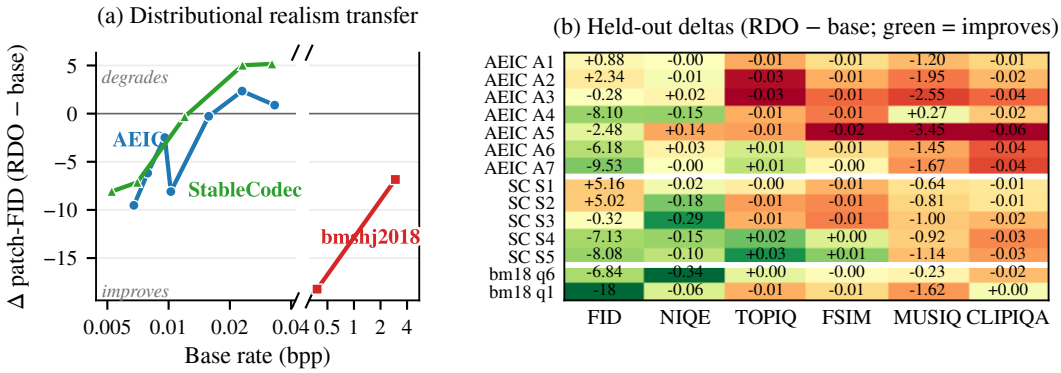


Figure 5: **Held-out transfer.** (a) Change in patch-FID versus ground truth. Realism gains concentrate in the low-rate regime and fade or reverse at high rates. (b) All held-out metrics at all measured rungs, color-coded from improvement (green) to degradation (red), dissenters included.

PSNR, LPIPS, and DISTS contributions all grow as rate falls, which is why its Pareto-3 fraction reaches 1.0 at the lowest rung. bmshj2018 *trades*: optimization consistently sacrifices PSNR for large perceptual-term gains, so on this MSE-trained codec the composite objective reallocates bits from distortion to perception. Its top rung is the exception that proves the rate-pressure reading: at q8 (4.5 bpp; Pareto-3 0.91 in Table 1) bits are abundant enough that optimization improves every term at once while shedding only 7% of them, and the PSNR-sacrificing trade emerges only from q6 downward. StableCodec is *distortion-led*: PSNR is the single largest contributor at every rung, with the perceptual terms joining at the lowest rungs. That one rate-dependence law emerges from three different gain compositions supports interpreting it as a property of amortization under rate pressure rather than of any single metric channel, and it previews the control of Section 6.2, since what per-image optimization amplifies is whatever the objective values.

6 Validity of the Gains

Optimizing differentiable perceptual metrics invites two objections, that the gains are adversarial artifacts of the optimized metrics and that they are generic effects of any per-image optimization rather than of the deployed objective. We test both.

6.1 Transfer to Never-Optimized Metrics

We score every base and optimized reconstruction with metrics excluded from the objective, namely the no-reference NIQE (Mittal et al., 2013), MUSIQ (Ke et al., 2021), and CLIP-IQA (Wang et al., 2023), the full-reference FSIM (Zhang et al., 2011) and TOPIQ (Chen et al., 2024), and a patch-FID (Heusel et al., 2017) against ground truth computed per rung and arm (base or optimized) over all

Table 2: Objective control at the AEIC 0.008 bpp rung (matched n=6, objective swapped). Mean deltas versus base, medians in parentheses for the composite.

Objective	ΔF	ΔPSNR	ΔLPIPS	ΔDISTS
Composite (deployed)	$+\mathbf{1.61}$ ($+1.55$)	$+0.47$	-0.024	-0.032
PSNR only	-3.91 (-4.08)	$+0.12$	$+0.081$	$+0.079$
Perceptual only	-0.19 (-0.20)	-0.73	-0.014	-0.032

non-overlapping 256^2 crops, roughly 800 per arm. The battery covers every AEIC and StableCodec rung and the two bmshj2018 rungs with saved reconstructions (q1, q6). **Patch-FID improves where the margin is large and worsens where it is small** (Figure 5). The realism gains concentrate at the low-rate rungs, reaching improvements of up to 18 FID points, and reverse at the high-rate rungs of AEIC and StableCodec, while bmshj2018 improves at both measured rungs, far more strongly at the lower. The law thus reproduces in a distributional metric the optimizer never saw. Across the full battery (Figure 5b), NIQE improves at every StableCodec rung and both measured bmshj2018 rungs while staying neutral on AEIC, TOPIQ flips positive at the lowest rungs of AEIC and StableCodec, and FSIM does so on StableCodec.

The dissenters quantify the metric-specific component. MUSIQ decreases consistently, and CLIP-IQA decreases on AEIC and StableCodec, so part of the composite-score gain is stylization that these aesthetics predictors penalize. We therefore scope all claims to the low-rate regime, where patch-FID improves on all three backbones, NIQE on two, TOPIQ on the two diffusion codecs, and FSIM on StableCodec, reporting the dissent rather than selecting agreeable metrics.

6.2 The objective matters

At a fixed AEIC rung (0.008 bpp) and matched images, we re-run the identical protocol under the deployed composite of Eq. (1), PSNR alone, and the perceptual terms alone. The composite objective is Pareto-positive on every metric while shedding bits (Table 2). PSNR-only optimization is severely counterproductive, losing nearly four composite points while gaining a mean of only 0.12 dB of PSNR. Perceptual-only optimization trades away 0.7 dB of PSNR for a small net loss. At extreme rates, per-image optimization is therefore not generically helpful but amplifies whatever objective it is given, so the deployed objective, and not a proxy, must be optimized.

Re-running the protocol on bmshj2018 with its *own training objective* (MSE) yields essentially nothing, with a median PSNR change of 0.00 dB and PSNR-ladder margins of 0.1–0.2% at every interpolable rung (Appendix C). The amortized encoder is near-optimal for the objective it was trained on, consistent with the small amortization gaps of Yang et al. (2020), so bmshj2018's large composite margins in Table 1 almost entirely measure *objective-retargeting capacity*, how far a frozen MSE codec can be steered toward a different deployed objective per image. That capacity too grows as rate falls, while the pure amortization reading rests on the two metric-aligned codecs.

7 Conclusion

We measured the value of per-image latent optimization as a function of operating rate, on three architecturally distinct codecs under one protocol, and found that it grows strongly as rate falls. As a codec descends its ladder, the median rate-equivalent margin grows and saturates at a backbone-specific ceiling, until the amortized encoder needs 1.7–$2.2\times$ its bits to match per-image optimization, a median BD-rate saving of 34–50%. The growth is driven by measured slack to the per-image frontier compounded by late-onset capture, and it transfers to held-out realism metrics in the regime claimed. Below roughly 0.01 bpp, encode-side optimization buys what would otherwise require 1.7–$2.2\times$ the bitrate. The exchange saturates in compute: doubling past the standard budget adds at most a relative 12%, while a quarter of the budget still captures roughly a quarter of the gain (Figure 3), a favorable trade whenever content is encoded once and decoded many times. With 86% of the slack still locked behind discretization at the lowest rung, the next gains in extreme-rate compression lie on the encoder side, through distilling optimized latents into the encoder, learned per-image quantization, or rate-adaptive encoder capacity, and the law gives these efforts a target and a yardstick.

REPRODUCIBILITY STATEMENT

All measurements use public codecs and datasets. The optimization protocol, margin definition, and statistics are fully specified in Section 3, with hyperparameters in Appendix A. We will release the measurement code, per-image result tables, and analysis scripts. Two reproducibility caveats are documented in Appendix B. First, range-coded bitstreams whose probability tables are computed on GPU are bound to the GPU architecture and framework stack that produced them, and cross-arch decoding yields corrupted reconstructions, so all rate measurements pin both. Second, four of roughly 700 base-ladder evaluation cells were detected as corrupted decodes by a monotonicity audit and excluded. The detector and the excluded cells are released with the data.

AI USE STATEMENT

In this work, we used generative AI tools (large language model coding assistants) to help develop the measurement and analysis code, to help draft and edit the manuscript text, and to assist with literature search. We did not use generative AI tools to generate research ideas, to produce or select experimental results, or to create figures beyond plotting scripts that render the measured data. All AI-assisted work was reviewed by the authors. AI-assisted code was tested against reference implementations and audited end to end, every number reported in the paper was regenerated from the released per-image result files by the analysis scripts, and all AI-assisted text was checked by the authors for factual accuracy against the underlying measurement artifacts. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 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 *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rkcQFMZRb.

Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. CompressAI: a PyTorch library and evaluation platform for end-to-end compression research. *CoRR*, abs/2011.03029, 2020. URL https://arxiv.org/abs/2011.03029.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013. URL http://arxiv.org/abs/1308.3432.

Gisle Bjøntegaard. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, ITU-T SG16/Q6, Video Coding Experts Group (VCEG), Austin, TX, USA, April 2001.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 6228–6237. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00652. URL https://doi.org/10.1109/CVPR.2018.00652.

Joaquim Campos, Simon Meierhans, Abdelaziz Djelouah, and Christopher Schroers. Content adaptive optimization for neural image compression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/CLIC_2019/Campos_Content_Adaptive_Optimization_for_Neural_Image_Compression_CVPRW_2019_paper.html.

Marlène 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ktdETU9JBg.

Chaofeng Chen and Jiadi Mo. IQA-PyTorch: PyTorch toolbox for image quality assessment. https://github.com/chaofengc/IQA-PyTorch, 2022.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Trans. Image Process.*, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466. URL https://doi.org/10.1109/TIP.2024.3378466.

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 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 7936–7945. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00796. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Cheng_Learned_Image_Compression_With_Discretized_Gaussian_Mixture_Likelihoods_and_Attention_CVPR_2020_paper.html.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1078–1086. PMLR, 2018. URL https://proceedings.mlr.press/v80/cremer18a.html.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810. URL https://doi.org/10.1109/TPAMI.2020.3045810.

Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. COIN: COmpression with implicit neural representations. *CoRR*, abs/2103.03123, 2021. URL https://arxiv.org/abs/2103.03123.

Eastman Kodak Company. Kodak lossless true color image suite (PhotoCD PCD0992). https://r0k.us/graphics/kodak/, 1993.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5708–5717. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00563. URL https://doi.org/10.1109/CVPR52688.2022.00563.

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. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 5128–5137. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00510. URL https://doi.org/10.1109/ICCV48922.2021.00510.

Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text + sketch: Image compression at ultra low rates. *CoRR*, abs/2307.01944, 2023. doi: 10.48550/ARXIV.2307.01944. URL https://doi.org/10.48550/arXiv.2307.01944.

Xiaoyue Ling, Chuqin Zhou, Xin Li, Guo Lu, Junxi Zhang, Qizhi Huang, Yifei Wang, Wenzhuo Ma, Zhenzhong Chen, Cheng-Yeh Chen, Fanqing Zhou, Hengyuan Na, Pukun Zhao, Tiancheng Si, Wenlong Xu, Wang lei, Xudong Wang, Xu, Zhimeng Huang, Jiaqi Zhang, Siwei Ma, Chuanmin Jia, Zetian Song, Xiaohan Lu, Qizhe Wang, Md. Mehedi Hasan, Shah Mohammad Abdul Mannan, Dipit Saha, Ahnaf Tahmid, Mohammad Raihan Rashid, Xudong Zhang, Chuangchuang Zeng, Shien Chen, Wang Luo, Yinlin Zhu, Di Wu, Tiange Zhang, Yilin Guo, Qingtong Liu, Junjun Si, Shuhong Liao, Xiandong MENG, Qirui Chen, and Shangquan Sun. LoViF 2026 challenge on ultra-low bitrate image compression: Methods and results. In *The Second Workshop on Low-Level Vision Frontiers with Generative AI, Preference Optimization, Agentic Systems and World Models@ECCV2026*, 2026. URL https://openreview.net/forum?id=CSFBAUIvM4.

Guo Lu, Chunlei Cai, Xiaoyun Zhang, Li Chen, Wanli Ouyang, Dong Xu, and Zhiyong Gao. Content adaptive and error propagation aware deep video compression. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12347 of *Lecture Notes in Computer Science*, pp. 456–472. Springer, 2020. doi: 10.1007/978-3-030-58536-5_27. URL https://doi.org/10.1007/978-3-030-58536-5_27.

Joe Marino, Yisong Yue, and Stephan Mandt. Iterative amortized inference. In *International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3403–3412. PMLR, 2018. URL https://proceedings.mlr.press/v80/marino18a.html.

Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8a50bae297807da9e97722a0b3fd8f27-Abstract.html.

David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 10794–10803, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/53edebc543333dfbf7c5933af792c9c4-Abstract.html.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726. URL https://doi.org/10.1109/LSP.2012.2227726.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI*, volume 15144 of *Lecture Notes in Computer Science*, pp. 87–103. Springer, 2024. doi: 10.1007/978-3-031-73016-0_6. URL https://doi.org/10.1007/978-3-031-73016-0_6.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rJiNwv9gg.

Lucas Theis, Tim Salimans, Matthew D. Hoffman, and Fabian Mentzer. Lossy compression with Gaussian diffusion. *CoRR*, abs/2206.08889, 2022. doi: 10.48550/ARXIV.2206.08889. URL https://doi.org/10.48550/arXiv.2206.08889.

Ties van Rozendaal, Iris A. M. Huijben, and Taco Cohen. Overfitting for fun and profit: Instance-adaptive data compression. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=oFp8Mx_V5FL.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 2555–2563. AAAI Press, 2023. doi: 10.1609/AAAI.V37I2.25353. URL https://doi.org/10.1609/aaai.v37i2.25353.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pp. 1398–1402. IEEE, 2003. doi: 10.1109/ACSSC.2003.1292216. URL https://doi.org/10.1109/ACSSC.2003.1292216.

Tongda Xu, Han Gao, Chenjian Gao, Yuanyuan Wang, Dailan He, Jinyong Pi, Jixiang Luo, Ziyu Zhu, Mao Ye, Hongwei Qin, Yan Wang, Jingjing Liu, and Ya-Qin Zhang. Bit allocation using optimization. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38377–38399. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23c.html.

Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/066f182b787111ed4cb65ed437f0855b-Abstract.html.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730. URL https://doi.org/10.1109/TIP.2011.2109730.

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 Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. StableCodec: Taming one-step diffusion for extreme image compression. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 17379–17389. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01614. URL https://doi.org/10.1109/ICCV51701.2025.01614.

Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang. The devil is in the details: Window-based attention for image compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 17471–17480. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01697. URL https://doi.org/10.1109/CVPR52688.2022.01697.

A PROTOCOL DETAILS

SGA details. With residual r and temperature τ, stochastic rounding draws $w = \mathrm{softmax}\big((\ell + g)/\tau\big)$ over the two neighboring grid points, where $\ell = \big(\log(1 - \mathrm{frac}(r)), \log \mathrm{frac}(r)\big)$ (log arguments clamped at 10^{-6}) and g are Gumbel samples. The soft value is the w-weighted combination. Temperature anneals as $\tau_t = \max(0.5\, e^{-0.004t}, 0.2)$ over 600 steps, followed by 150 straight-through polish steps at $0.2\times$ learning rate. The delta is optimized with Adam (lr 5×10^{-3}, gradient-norm clip 0.3), and the schedule repeats for two re-annealing rounds with a fresh optimizer. Soft rates use the Gaussian conditional likelihood with scales clamped at 0.11 and tail mass 10^{-9}. True rates re-encode the rounded latent with the codec's range coder every 100 steps, and a running

calibration factor maps soft to true rate for the two-tier penalty in the main objective. Rate-infeasible final states fall back to the last true-coded candidate. The identical configuration, including λ_0, is used for all backbones, rungs, datasets, and controls. Only the step-budget probes of Section 4.3 vary the schedule length.

Continuous-relaxation runs. The mechanism bound of Section 5 removes rounding everywhere (soft value $= r$), keeps the identical objective, schedule, and rate pressure with the soft rate, and reports the continuous reconstruction's quality with the soft rate estimate. The hard projection additionally rounds the final continuous latent and entropy-codes it with the range coder.

Rate accounting. All reported rates are complete bitstream sizes, namely the range-coded latent plus hyper-latent strings plus container header, divided by the ground-truth pixel count. StableCodec images are replicate-padded to 256-multiples, since its latent grid is $1/64$ of the image and the four-group mask requires divisibility by 4. Padding bits are charged to the bitstream while quality is scored on the unpadded region, identically for base and optimized arms.

B Data-integrity audits

Because the campaign spans roughly 150 GPU-jobs, we ran continuous audits. Three findings materially affect measurement hygiene and are disclosed here.

Corrupted base cells. A monotonicity detector flags any base-ladder cell whose composite score falls more than 2 points below a *cheaper* rung of the same image, which is impossible for a healthy ladder. It found 4 of roughly 700 cells, all consistent with entropy-decode corruption. They are excluded from ladders and deltas, and no excluded cell is load-bearing for any conclusion.

Cross-architecture reproducibility. Bitstreams range-coded with GPU-computed probability tables decode correctly only on the producing combination of GPU architecture and framework stack. We verified value-level reproducibility across architectures by optimizing the same image independently on two GPU families, with PSNR within 0.015 dB and rate within 0.8%, while keeping each measurement's encode and decode pinned to one architecture.

Result-file consolidation. Early sweep shards wrote a shared per-directory result file, which under concurrency loses sibling entries to last-writer-wins races. All grids were rebuilt with per-job files and merged loaders, missing entries were recomputed from saved reconstructions and bitstreams (validated bit-exact against a pre-loss snapshot on the one overlapping image), and completeness was re-verified cell by cell (7×24, 6×12, 5×24 complete).

C Additional results

Kodak-24 replication. On bmshj2018 with in-run base ladders and all 24 images matched, margin medians are 74.9% at q1 (0.41 bpp) and 92.4% at q2 (0.61 bpp), with ΔF medians of $+2.8$ and $+3.0$. On StableCodec (S2, S3, S4, all 24 images matched), margins rise $8.9\% \rightarrow 43.1\% \rightarrow 66.5\%$ as bpp falls from 0.026 to 0.0086. The shapes replicate the challenge-set law. The level drop discussed in Section 4.4 is a StableCodec effect, while bmshj2018's Kodak levels straddle their challenge-set counterparts (74.9% versus 85% at q1, 92.4% versus 86% at q2).

Full held-out battery. Signed deltas (optimized minus base) for all rungs and metrics, including the dissenting MUSIQ and CLIP-IQA columns, are released as `heldout_metrics.json`. Section 6.1 quotes the patch-FID, NIQE, TOPIQ, and FSIM patterns without selection.

Strict versus inclusive margins. Headline medians drop per-image extrapolated margins. Including them as looser, extrapolation-based estimates raises low-rung StableCodec medians (S4: 169% versus 118%) and barely moves AEIC or bmshj2018. All conclusions are unchanged under either convention, and both sets ship with the released tables.

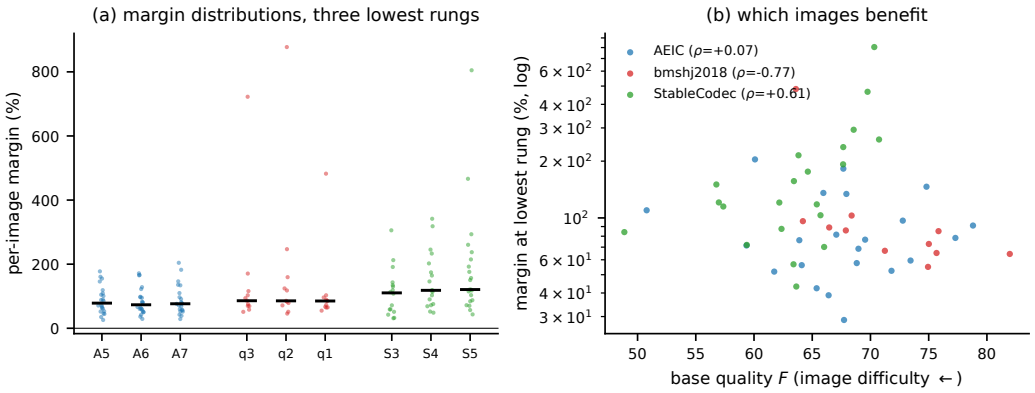


Figure 6: **Per-image anatomy of the margin.** (a) Full margin distributions at the three lowest rungs of each backbone (black bar = median). The distributions are wide, heavy-tailed, and stable, with per-image ranks correlating across rungs at $\bar{\rho} = 0.83$–0.98. (b) Margin at the lowest rung versus the image's base quality. Harder images benefit most on bmshj2018, easier ones on StableCodec, and AEIC shows no preference.

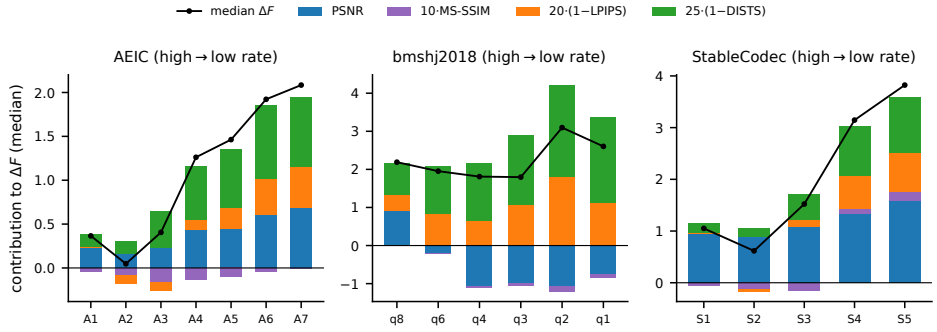


Figure 7: **Gain routes.** Median contribution of each composite term to ΔF per rung (bars; black line = median ΔF), ordered high to low rate. AEIC gains are balanced across terms, bmshj2018 trades PSNR (negative bars) for perceptual terms, and StableCodec is distortion-led, with PSNR the largest contributor. Three routes, one law.

MSE-objective control on bmshj2018. Re-running the full protocol with the MSE objective, which is the backbone's training objective, on six images at q1, q2, q4, and q6 gives a median ΔPSNR of 0.00 dB at every rung and PSNR-ladder rate-equivalent margins of 0.1%, 0.2%, and 0.1% at q6, q4, and q2 (q1 is extrapolation-limited, $n{=}1$). Per-image optimization of the training objective finds nothing to fix. The composite-objective margins of Table 1 on this backbone measure retargeting capacity.

Restart variance. Two independent restarts of the full protocol at the lowest well-populated rung of each backbone reproduce the median gain within ± 0.15 (StableCodec: $+2.40/+2.59$ versus $+2.54$; AEIC: $+1.78/+1.81$ versus $+1.90$). Run-to-run optimization noise is thus an order of magnitude smaller than the gains it is used to certify.

Qualitative comparison. Figure 8 shows base versus per-image-optimized reconstructions at the lowest AEIC rung (0.0068 bpp). Optimization recovers texture and structure that the amortized encoding blurs or hallucinates differently, at equal or fewer bits.

Rate shedding. Across rungs and backbones the optimized latents code to 60–80$\%$ of their base rate under the two-tier penalty, and per-rung shed fractions are included in the released tables. This is the phenomenon that motivates the rate-equivalent margin over quality-at-matched-rate deltas (Section 3.3).

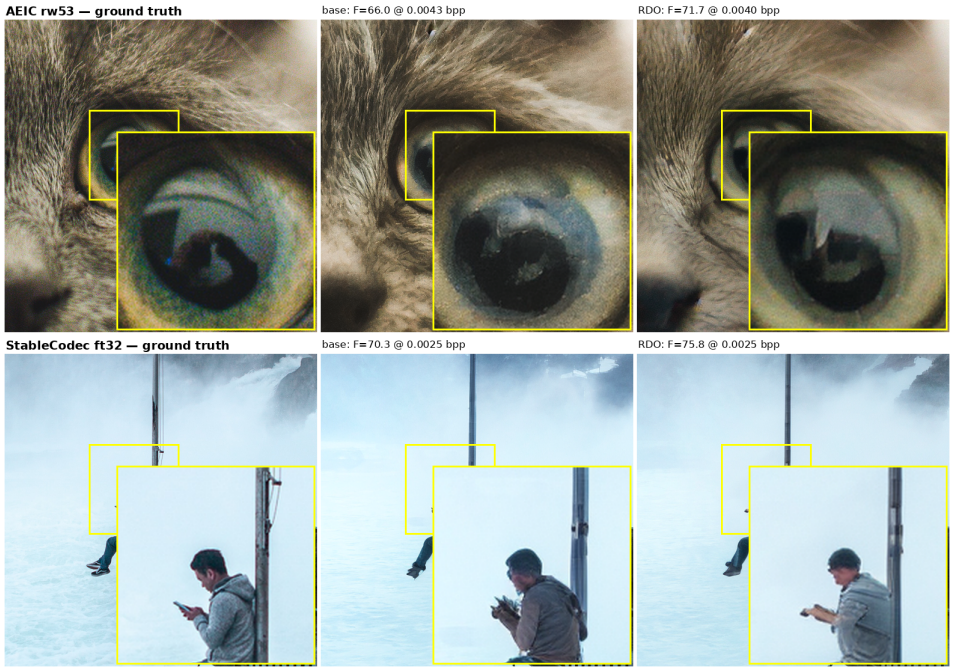


Figure 8: Ground truth, base encoder, and per-image RDO reconstructions (AEIC A7, 0.0068 bpp base rate; 460^2 crops; RDO uses equal or fewer bits on both images).