FIDELITY WHERE YOU LOOK, REALISM EVERYWHERE: ATTENTION-GUIDED ADDITIVE FIDELITY ALLOCATION FOR GENERATIVE IMAGE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image codecs produce realistic reconstructions, but spatially uniform distortion pricing cannot distinguish a harmless plausible substitution from the loss of source-specific detail in a face or a word. We introduce FLARE, a one-step generative codec that adds fidelity on top of a realism floor according to a continuous spatial fidelity-need field. FLARE spatially extends rate–distortion–perception (RDP) optimization and decomposes its objective into the baseline realism-codec objective and an excess-fidelity allocation problem. A training-only auxiliary decoder makes a dedicated message latent carry the requested source evidence without imposing strong local regression on the delivered reconstruction. Field-conditioned latent gating and scaling control where and how much message rate is spent; no field is transmitted, and the one-step decoding path does not vary with the field. Because the floor objective is retained from the realism codec, FLARE at zero excess remains competitive with its backbone. Experiments show that coded rate lands where the field points, grows monotonically with its magnitude, and improves fidelity in the requested region without harming the background. For perceived quality, we instantiate the field with off-the-shelf predicted attention, motivated by fixation analysis showing that attention concentrates on faces, text, and other information-bearing content. In a subjective quality assessment with 37 observers, FLARE improves MOS by 0.29–0.57 over OneDC and AEIC-ME, and at least 35 observers rated it higher on average than either competitor. Our code is included in the supplementary material.

1 INTRODUCTION

Generative image codecs produce convincing reconstructions at rates where conventional codecs blur away fine detail (Mentzer et al., 2020; Careil et al., 2024; Xue et al., 2025; Zhang et al., 2026). Their realism, however, is distributional: a reconstruction can look natural while depicting a different face, word, or object instance. Such plausible substitutions conceal the loss of source-specific information and can be more consequential than visibly blurred but faithful detail (Fig. 1). Realism and source fidelity are therefore complementary obligations. The standard rate–distortion–perception (RDP) objective (Blau & Michaeli, 2018; 2019) underlying conventional realism codecs balances source-paired distortion against a distribution-level perception divergence. Because the perception term provides no source-paired spatial error signal, the distortion term provides the natural basis for spatial fidelity allocation. The RDP objectives, however, price this distortion uniformly across the image. A single global distortion weight cannot express that source fidelity has unequal value across the image. It may under-allocate rate where an error changes the readout a viewer takes from a region, yet over-allocate it where a comparable error causes relatively lower perceived degradation (Fig. 2). This motivates a method that allocates fidelity according to spatially varying needs arising from region-specific content, while maintaining realism across the complete image.

To realize this allocation, we introduce FLARE (*Fidelity where you Look and Realism Everywhere*), a codec built on a one-step diffusion prior. We extend RDP with a spatially varying fidelity need and decompose the resulting weighted distortion into a spatially uniform fidelity-need floor and an excess fidelity-need field. This decomposition provides a clear separation between the conventional realism-codec objective and the additional objective of spatially targeted fidelity, allowing us to focus our design on the source-specific information required by the excess field. Given a source image $\mathbf{x}$ and a continuous-domain excess fidelity-need field $\breve{\mathrm{n}}$, FLARE allocates source-specific information to the corresponding spatial locations in the encoded latent representation $\hat{\mathbf{y}}$. Specifically, $\breve{\mathrm{n}}$

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: **Qualitative comparison.** FLARE preserves information-bearing regions such as faces and text far better than competing codecs, which blur or fabricate them. White boxes mark the displayed crop regions in the left-hand GT/FLARE split. Each row reports bpp. Reconstructions are en/decoded following our subjective-study protocol (Sec. 6.3).

controls spatial gating and scaling of the latent representations, regulating the amount of source-specific information at each location. We emphasize that FLARE focuses on where and how much source fidelity is represented in the code; it does not adaptively modify the one-step generative decoding procedure, introduce separate decoding branches, or transmit an explicit spatial map or mask to the decoder. Thereby, we can retain the structural simplicity of the underlying codec. Experiments confirm that FLARE allocates rate as specified by $\breve{\pi}$ while preserving a competitive realism floor at zero excess. Although FLARE is agnostic to how $\breve{\pi}$ is generated, its perceived-quality benefit needs a field flagging where lost fidelity most impairs a viewer's readout.

Human attention is a strong candidate for this role: in human-attention data (MIT1003 (Judd et al., 2009)), attended regions are markedly enriched in faces, facial landmarks, text, and animals relative to randomly sampled regions (Sec. 5), which makes attention a practical indicator of fidelity-critical content. We therefore instantiate $\breve{\pi}$ with maps from the off-the-shelf predictor DeepGaze (Kümmerer et al., 2025). This is one instantiation; better attention or fidelity-need predictors can be substituted without changing the interface. With it, FLARE outperforms AEIC-ME (Zhang et al., 2026) and OneDC (Xue et al., 2025) in a matched-rate subjective study, raising mean opinion score and winning the large majority of pairwise preferences. The same interface accepts application-specific fields (Sec. 5), e.g. a detector marking faces, plates, or logos whose identity must be retained, without retraining the codec or decoder. Our contributions are:

- a spatial RDP formulation that decomposes its weighted distortion into a spatially uniform fidelity-need floor and an excess fidelity-need field, thereby allowing established realism-codec objectives and training principles to be retained for the floor while isolating only excess fidelity as the spatial-allocation problem to be solved;
- a training-only auxiliary-decoder supervision strategy for this isolated excess-fidelity problem that makes the latent carry source-specific information without applying a strong excess-weighted distortion loss directly to the delivered reconstruction. Controlled ablations show that this design improves fidelity while preserving the realism of decoded samples, unlike direct reconstruction supervision;
- a complete codec realization, FLARE, that uses latent gating and scaling controlled by the excess fidelity-need field to allocate fidelity while retaining a single fixed one-step decoding path whose architecture and execution do not vary with that field. Experiments verify spatially targeted rate allocation, monotonic control through the magnitude of the excess fidelity-need field, and effective distortion reduction in the requested regions, while preserving a competitive realism floor at zero excess; and
- subjective validation of a DeepGaze-based FLARE instantiation: in a matched-rate 37-observer study, it improves MOS by 0.29–0.57 over OneDC and AEIC-ME, and at least 35 of the 37 observers rated FLARE higher on average than either competitor at both rates.

2 Related Work

Rate–distortion–perception and generative compression. The perception–distortion tradeoff separates distributional plausibility from instance fidelity (Blau & Michaeli, 2018; 2019). GAN, diffu-

sion, and generative-latent codecs instantiate this tradeoff to maintain realism at extreme compression rates (Agustsson et al., 2019; Mentzer et al., 2020; Theis et al., 2022; Yang & Mandt, 2023; Careil et al., 2024; Xue et al., 2025; Zhang et al., 2025; 2026). Recent one-step codecs raise this uniform floor further: SODEC (Chen et al., 2026) through fidelity-guided decoding, DiT-IC (Shi et al., 2026) through an aligned transformer backbone, and CADC (Sheng et al., 2026) through content-adaptive quantization with an auxiliary decoder that concentrates information into the channels the diffusion decoder consumes; none exposes a spatial fidelity interface, and CADC's auxiliary decoder is content-driven and retained at inference, whereas ours supervises a field-gated message group and is discarded after training. Their distortion terms, however, are generally aggregated uniformly across image locations.

Foveated and region-adaptive coding. Foveated, ROI, and learned importance-aware codecs allocate rate non-uniformly (Geisler & Perry, 1998; Wang & Bovik, 2001; Christopoulos et al., 2000; Li et al., 2018; Prakash et al., 2017). Learned methods further use continuous importance or saliency maps both to vary local code length and to weight reconstruction error (Li et al., 2018; Patel et al., 2021). Within generative compression, selective schemes mask pixelwise distortion to preserved regions or combine ROI-restricted reconstruction losses with adversarial or perceptual training (Agustsson et al., 2019; Löhdefink et al., 2020; Ma et al., 2021). AGDL (Zhang & Wu, 2021) refines a conventional base layer with a second stream and decoding stage, guided by attention inferred from the decoded base; FLARE takes the field as input and decodes a single latent in one pass. These works establish spatially varying rate and distortion, but do not connect a generic fidelity-need field within RDP to a dedicated source-message representation.

Region-adaptive generative codecs. MRIDC (Xu et al., 2025) motivates its design with ROI-conditioned distortion plus global distributional perception, but this spatial RDP expression is a conceptual description of its goal rather than its spatial-allocation training loss. Operationally, its training-free masking module retains VQ tokens in the ROI and masks non-ROI tokens, which are predicted before iterative diffusion decoding. Relic et al. (2026) instead transmit a spatial timestep map and use it for quantization, entropy coding, and iterative denoising; their method provides arbitrary-map spatial control without an RDP formulation. Both methods realize spatial control by changing how the decoder executes—masked tokens or a transmitted timestep map steer an iterative sampler (MRIDC: 20 steps; Relic et al.: up to 50 steps). FLARE instead changes what is encoded: the excess fidelity-need field controls a dedicated, explicitly supervised latent message, and the fixed one-step decoder is left unchanged (Sec. 4.1).

3 Spatial RDP and Floor–Excess Decomposition

This section extends standard RDP with a spatial fidelity-need field to account for spatially varying demands on source fidelity, and decomposes the resulting objective into a fidelity-need-independent floor and an excess fidelity-need field. This separation retains established realism-codec objectives and training principles for the floor while isolating the excess-fidelity allocation problem.

Let $\mathbf{x} \sim p_{\mathbf{x}}$ be a source image and $\hat{\mathbf{x}}$ its reconstruction under a codec with parameters θ. The Lagrangian form of the standard RDP objective (Blau & Michaeli, 2018; 2019) is

$$\min_{\theta} \; \mathbb{E}_{\mathbf{x}}[R + \lambda_{\mathrm{D}} \Delta(\mathbf{x}, \hat{\mathbf{x}})] + \lambda_{\mathrm{P}} \rho\big(p_{\mathbf{x}}, p_{\hat{\mathbf{x}}}\big), \tag{1}$$

where $R = R_\theta(\mathbf{x})$ is the expected code length for $\mathbf{x}$, Δ is a full-reference distortion, and $\lambda_{\mathrm{D}}, \lambda_{\mathrm{P}} \geq 0$ set the rate–distortion–perception operating point. In RDP terminology, *perception* denotes distributional realism: ρ compares the distributions of source and reconstructed images, rather than measuring pairwise similarity between $\mathbf{x}$ and $\hat{\mathbf{x}}$.

3.1 Spatial Extension of the RDP with a Fidelity-Need Field

Let Ω be the continuous image domain. We separate the full-reference distortion into a non-spatial component and a pointwise-decomposable component:

$$\Delta(\mathbf{x}, \hat{\mathbf{x}}) = \Delta_{\mathrm{G}}(\mathbf{x}, \hat{\mathbf{x}}) + \int_{\Omega} \mathfrak{d}(u) \, \mathrm{d}u. \tag{2}$$

Here $\Delta_{\mathrm{G}} \geq 0$ is a paired distortion evaluated globally (or without a canonical pointwise decomposition) and $\mathfrak{d}(u) \geq 0$ a pointwise distortion; both are source-paired and distinct from the distributional

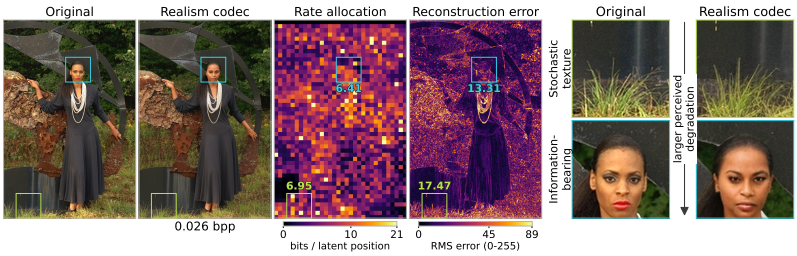


Figure 2: **Similar local rate allocation does not ensure similar perceptual fidelity.** A codec spends comparable local rate on a stochastic texture and on an information-bearing face, and the texture even incurs *higher* pixel error, yet the face suffers far larger perceived degradation—neither rate nor pixel error captures the unequal perceptual consequences of local reconstruction errors.

term ρ, and either may combine several distortions with their coefficients absorbed. We introduce a total *fidelity-need field* $\mathfrak{n} : \Omega \to \mathbb{R}_+$ that weights the pointwise distortion at each location; for brevity we write $\mathfrak{d}(u)$ for $\mathfrak{d}(u, \mathbf{x}, \hat{\mathbf{x}})$ and $\mathfrak{n}(u)$ for $\mathfrak{n}(u, \mathbf{x})$. Applying the field to the pointwise component only, with λ_{D} setting the overall distortion scale, gives

$$\min_{\theta} \mathbb{E}_{\mathbf{x}}\left[R + \lambda_{\mathrm{D}}\left(\Delta_{\mathrm{G}}(\mathbf{x}, \hat{\mathbf{x}}) + \int_{\Omega} \mathfrak{n}(u)\mathfrak{d}(u)\,\mathrm{d}u\right)\right] + \lambda_{\mathrm{P}}\rho\big(p_{\mathbf{x}}, p_{\hat{\mathbf{x}}}\big). \tag{3}$$

3.2 Floor–excess decomposition

To retain established realism-codec objectives and training principles while isolating the excess-fidelity allocation problem, we decompose the fidelity-need field into a uniform fidelity-need floor and a nonnegative excess fidelity-need field. Let $\mathfrak{n}_{\mathrm{floor}}$ denote the fixed floor; it represents the objective-level baseline fidelity pressure retained from the realism-codec objective, and we take $\mathfrak{n}_{\mathrm{floor}} = 1$ so that D^{floor} below coincides with the realism codec's own distortion (any other constant is absorbed by λ_{D}). We define

$$\breve{\mathfrak{n}}(u) := \mathfrak{n}(u) - \mathfrak{n}_{\mathrm{floor}} \geq 0, \qquad \mathfrak{n}(u) \geq \mathfrak{n}_{\mathrm{floor}}. \tag{4}$$

Here, $\breve{\mathfrak{n}}(u)$ denotes the excess fidelity need at location u above the uniform floor. Then the full weighted distortion in Eq. 3 decomposes as

$$\Delta_{\mathrm{G}} + \int_{\Omega} \mathfrak{n}(u)\mathfrak{d}(u)\,\mathrm{d}u = \underbrace{\Delta_{\mathrm{G}} + \mathfrak{n}_{\mathrm{floor}} \int_{\Omega} \mathfrak{d}(u)\,\mathrm{d}u}_{D^{\mathrm{floor}}} + \underbrace{\int_{\Omega} \breve{\mathfrak{n}}(u)\mathfrak{d}(u)\,\mathrm{d}u}_{D^{\mathrm{excess}}}, \tag{5}$$

where D^{floor} retains the global and uniform distortion terms of the realism codec, and D^{excess} isolates pointwise distortion weighted by the excess fidelity-need field—the spatial-allocation problem addressed here. The decomposition is additive at the level of the objective: raising $\breve{\mathfrak{n}}$ in one region adds fidelity there without relaxing the floor elsewhere (verified in Appendix Table 9). We likewise decompose the rate as $R = R^{\mathrm{floor}} + R^{\mathrm{excess}}$. Substituting this rate decomposition and Eq. 5 into Eq. 3 yields

$$\mathcal{L} = \underbrace{\mathbb{E}\big[R^{\mathrm{floor}} + \lambda_{\mathrm{D}} D^{\mathrm{floor}}\big] + \lambda_{\mathrm{P}}\rho(p_{\mathbf{x}}, p_{\hat{\mathbf{x}}})}_{\text{baseline realism-codec RDP objective}} + \underbrace{\mathbb{E}[R^{\mathrm{excess}} + \lambda_{\mathrm{D}} D^{\mathrm{excess}}]}_{\text{excess spatial rate–distortion objective}}. \tag{6}$$

Eq. 6 separates the complete baseline realism-codec RDP objective from the excess spatial rate-allocation problem.

4 FLaRE: Realism-Preserving Excess-Fidelity Allocation

We introduce FLaRE, a codec optimized with respect to the decomposed objective in Eq. 6. This decomposition enables a modular design in which FLaRE builds directly on a state-of-the-art realism codec (Zhang et al., 2026), AEIC-ME, as its baseline while augmenting it with an excess-fidelity allocation mechanism. Given a source image $\mathbf{x}$ and an excess fidelity-need field $\breve{\mathfrak{n}}$, the overall coding flow is summarized at the representation level as $(\mathbf{x}, \breve{\mathfrak{n}}) \to \hat{\mathbf{y}} \to \hat{\mathbf{x}}$, where $\hat{\mathbf{y}}$ denotes the compressed latent representation of $\mathbf{x}$ conditioned on the excess fidelity-need field, and $\hat{\mathbf{x}}$ denotes the reconstructed image. We first describe the coding architecture and operations, and then specify how the two objective blocks in Eq. 6 supervise this architecture during training.

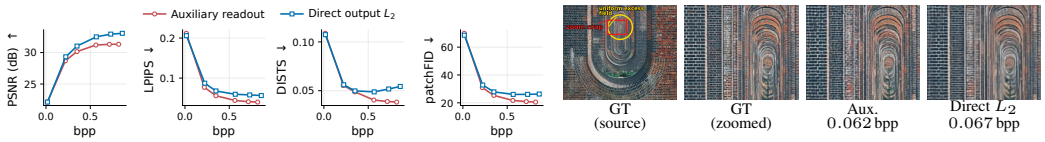


Figure 3: **Auxiliary vs. direct L_2 supervision.** *Left:* Under a uniform field $\breve{\mathfrak{n}}(u) = \eta^3$ on Kodak24, aux. supervision improves fidelity and perceptual metrics as η increases, whereas direct L_2 eventually degrades DISTS/patchFID. *Right:* Direct L_2 produces visible blur in the zoomed sample.

4.1 Codec architecture and fidelity-need-conditioned coding

We now introduce the encoder-side operations of FLARE. Let G_a denote the field-conditioned analysis transform that produces a latent $\mathbf{y}$. Subsequently, FLARE conducts field-dependent scaling, gating, quantization, and entropy coding against $\mathbf{y}$. After entropy decoding and inverse-scaling, the reconstructed latent $\hat{\mathbf{y}}$ is passed to the synthesis transform G_s to obtain the reconstruction $\hat{\mathbf{x}}$. Fig. 8 (Appendix A) shows the complete codec architecture.

Latent organization. FLARE preserves the 320-channel latent interface of AEIC-ME (Zhang et al., 2026) and partitions the pre-quantization latent $\mathbf{y} \in \mathbb{R}^{H_y \times W_y \times 320}$ as

$$\mathbf{y} = \big[\ \underbrace{\mathbf{y}^{\text{base}}}_{48 \text{ channels}}\ ;\ \underbrace{\mathbf{y}^{\text{msg}}}_{272 \text{ channels}}\ \big]. \tag{7}$$

The message group $\mathbf{y}^{\text{msg}}$ supplies the field-controlled coding capacity described below. The two groups are not independently decoded: they are entropy-coded by one shared hyperprior, concatenated, and consumed by the same synthesis transform. Beyond serving as an interface, this asymmetric split is also an optimization choice: restricting the excess-fidelity role to the message group reaches a better optimum than an unsplit latent under a mathematically identical objective—attaining lower training loss while concentrating rate on the requested regions and spending fewer total bits. Appendix A.2 gives the occupancy analysis used to select the 48/272 split, Appendix H isolates this optimization benefit under the matched objective, and Appendix A.1 specifies the added modules.

Active-region construction. Let $\text{Pool}_{\max}$ denote max pooling to the latent grid, $\tau > 0$ the activation threshold, $\mathbb{I}[\cdot]$ the indicator function, Up_{NN} nearest-neighbor upsampling to the image grid, and $\odot$ elementwise multiplication. We construct

$$\mathbf{m} = \mathbb{I}[\text{Pool}_{\max}(\breve{\mathfrak{n}}) \geq \tau], \qquad \widetilde{\mathfrak{n}} = \breve{\mathfrak{n}} \odot \text{Up}_{\text{NN}}(\mathbf{m}), \tag{8}$$

where $\mathbf{m} \in \{0,1\}^{H_y \times W_y}$ is the latent-grid activation mask. The masked field $\widetilde{\mathfrak{n}}$ retains the supplied magnitude within the active region and is zero elsewhere.

Conditioned analysis and message scaling. The masked field is injected into G_a through spatially varying, multi-scale FiLM modulation:

$$\mathbf{y} = [\mathbf{y}^{\text{base}}; \mathbf{y}^{\text{msg}}] = G_a(\mathbf{x}; \widetilde{\mathfrak{n}}). \tag{9}$$

A scale network S maps the same field to a positive channelwise scale $\mathbf{s}$ on the latent grid and scales the message group as

$$\mathbf{s} = S(\widetilde{\mathfrak{n}}), \qquad \mathbf{y}^{\text{msg}}_{\text{s}} = \mathbf{y}^{\text{msg}} \odot \mathbf{s}. \tag{10}$$

All latent quantizers use unit-step rounding. Scaling before quantization and undoing it at the decoder therefore produces an elementwise effective step size of $1/\mathbf{s}$, so $\mathbf{s}$ controls the precision and source-specific information carried by the message group.

Shared hyperprior coding and hard gating. Let H_a and H_s denote the hyper-analysis and hyper-synthesis transforms, C_{QT} the four-part quadtree context model, and Q unit-step quantization. The hyper-analysis input combines the base latent, hard-gated scaled message latent, scale, and mask:

$$\hat{\mathbf{z}} = Q\big(H_a\big([\mathbf{y}^{\text{base}}; \mathbf{m} \odot \mathbf{y}^{\text{msg}}_{\text{s}}; \mathbf{s}; \mathbf{m}]\big)\big), \quad (\boldsymbol{\mu}, \boldsymbol{\sigma}) = C_{\text{QT}}(H_s(\hat{\mathbf{z}})), \tag{11}$$

where $\boldsymbol{\mu} = [\boldsymbol{\mu}^{\text{base}}; \boldsymbol{\mu}^{\text{msg}}]$ and $\boldsymbol{\sigma} = [\boldsymbol{\sigma}^{\text{base}}; \boldsymbol{\sigma}^{\text{msg}}]$. For brevity, Eq. 11 omits the dependence of C_{QT} on previously reconstructed partitions. The primary latents are quantized as

$$\hat{\mathbf{y}}^{\text{base}} = \boldsymbol{\mu}^{\text{base}} + Q\big(\mathbf{y}^{\text{base}} - \boldsymbol{\mu}^{\text{base}}\big), \quad \hat{\mathbf{y}}^{\text{msg}}_{\text{s}} = \boldsymbol{\mu}^{\text{msg}} + Q(\mathbf{m} \odot (\mathbf{y}^{\text{msg}}_{\text{s}} - \boldsymbol{\mu}^{\text{msg}})). \tag{12}$$

Thus, $\mathbf{m}$ determines where the message symbols are enabled, while $\mathbf{s}$ controls their coding precision.

Field-free fixed decoding. Let H_{inv} be the learned scale-reconstruction head (the subscript refers to inverting the encoder-side scaling). It recovers the reconstructed scale $\hat{\mathbf{s}}$ from the output of H_s, after which the message representation is de-scaled and the full latent is decoded:

$$\hat{\mathbf{s}} = H_{\text{inv}}(H_s(\hat{\mathbf{z}}))\,, \qquad \hat{\mathbf{y}}^{\text{msg}} = \hat{\mathbf{y}}_{\text{s}}^{\text{msg}}/\hat{\mathbf{s}}, \qquad \hat{\mathbf{x}} = G_s([\hat{\mathbf{y}}^{\text{base}}; \hat{\mathbf{y}}^{\text{msg}}]; \hat{\mathbf{s}}). \tag{13}$$

In FLARE, neither $\breve{\mathbf{n}}$, $\widetilde{\mathbf{n}}$, nor $\mathbf{m}$ is transmitted to the decoder. Instead, the hyperlatent $\hat{\mathbf{z}}$ absorbs the scale and mask information as in Eq. 11 at no measurable cost: with the study field the coded hyperlatent is 2–15 B per image *smaller* than with a zero field on the study set, CLIC 2020 test, and DIV2K validation (Table 12, Appendix E.1). Different fields therefore change only the coded representation; the decoding graph of Eq. 13 is fixed.

4.2 Realizing and optimizing the floor–excess objective

We realize the two blocks of Eq. 6 through the rate-accounting convention $R^{\text{floor}} = R(\hat{\mathbf{y}}^{\text{base}}) + R(\hat{\mathbf{z}})$ and $R^{\text{excess}} = R(\hat{\mathbf{y}}_{\text{s}}^{\text{msg}})$. The baseline realism-codec block is applied to the delivered reconstruction $\hat{\mathbf{x}}$, while the excess block charges the field-controlled message rate and an excess-fidelity surrogate that will be described below. At the objective level, the resulting implementation is

$$\mathcal{L}_{\text{FLARE}} = \underbrace{R(\hat{\mathbf{y}}^{\text{base}}) + R(\hat{\mathbf{z}}) + \lambda_{\text{D}} D^{\text{floor}} + \lambda_{\text{P}} \rho(p_{\mathbf{x}}, p_{\hat{\mathbf{x}}})}_{\text{baseline realism-codec block}} + \underbrace{R(\hat{\mathbf{y}}_{\text{s}}^{\text{msg}}) + \lambda_{\text{D}} \widetilde{D}^{\text{excess}}}_{\text{excess-fidelity block}}. \tag{14}$$

Individual losses used for the baseline block are detailed in the Appendix E.3.

Auxiliary-decoder supervision. Directly realizing D^{excess} would apply a strongly $\breve{\mathbf{n}}$-weighted squared error to $\hat{\mathbf{x}}$. This improves pixel fidelity but can oppose generative realism by driving conditional-mean regression and smoothing structured texture. We instead decode the message group only during training, $\hat{\mathbf{x}}^{\text{aux}} = G_s^{\text{aux}}(\hat{\mathbf{y}}^{\text{msg}})$, and use

$$\widetilde{D}^{\text{excess}} = \frac{1}{|\Omega|} \sum_{u \in \Omega} \breve{\mathbf{n}}(u)\, \|\hat{\mathbf{x}}^{\text{aux}}(u) - \mathbf{x}(u)\|_2^2\,. \tag{15}$$

This surrogate makes the message latent carry the requested source evidence without imposing the strong excess-weighted regression directly on the delivered image; Eq. 14 is thus a relaxation of Eq. 6 that constrains the representation rather than its minimizer over $\hat{\mathbf{x}}$. The auxiliary decoder is removed after training and adds neither side information nor decoding operations at inference. Pretraining additionally uses a weak field-weighted ℓ_2 on $\hat{\mathbf{x}}$ (one third of the auxiliary weight) as an initial optimization aid; it is off throughout adversarial fine-tuning (Appendix A.3). Fig. 3 isolates this choice using variants fine-tuned from the same initialization. Across a uniform excess-field sweep, auxiliary supervision improves PSNR and all three perceptual metrics monotonically with rate, while direct output supervision eventually worsens DISTS and patchFID and visibly smooths structured texture despite using more rate. Appendix E.3 gives the complete protocol.

5 Excess Fidelity-Need Fields for Perceptual Quality

To improve perceptual quality for human consumption, we present one effective way to construct the excess fidelity-need field. Motivated by the rate-allocation mismatch in Fig. 2, we assume that source fidelity matters more in (i) attended regions, consistent with foveated and saliency-aware compression (Geisler & Perry, 1998; Wang & Bovik, 2001; Li et al., 2018; Patel et al., 2021), and (ii) information-bearing regions with specific content that supports a meaningful visual readout.

Human attention can be readily estimated with predictors (Djilali et al., 2024; Kümmerer et al., 2025), while identifying and weighting information-bearing regions at inference is costly and ambiguous because their importance depends on the viewer and context. Encouragingly, our analysis of human fixations in MIT1003 (Judd et al., 2009) shows that attention preferentially localizes representative information-bearing content: relative to image area, fixation enrichment is $9.2\times$ for faces, $17.6\times$ for facial landmarks, $7.5\times$ for text, and $2.8\times$ for animals, versus $0.6\times$ for background (Fig. 4). These observations motivate predicted attention as a task-agnostic, non-exhaustive proxy for fidelity need. Appendix B.2 gives the complete analysis protocol. As one practical instantiation of this principle, we use attention maps predicted by an off-the-shelf DeepGaze MSDB (Kümmerer et al., 2025) as approximate proxies for fidelity need. We preprocess the predictions and construct the resulting excess fidelity-need fields using a fixed procedure applied to all images, without per-sample tuning or manual intervention. The preprocessing rule and all constants are given in Appendix B.1.

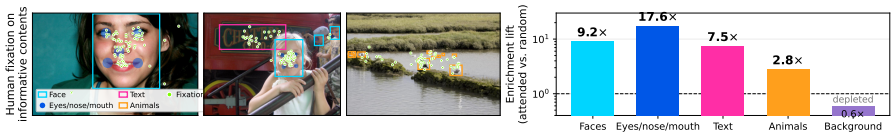


Figure 4: **Human attention is enriched on information-bearing content.** Left: MIT1003 fixations over detected faces, facial landmarks, text, and animals. Right: Enrichment lift of each class.

6 Experiments

6.1 Experimental setup

Model and operating points. FLARE is built on the one-step diffusion realism codec AEIC-ME (Zhang et al., 2026), trained on AEIC-ME's own corpus (Appendix A.3), and adds only the excess-fidelity interface of Sec. 4. Five operating points ($\lambda_{\text{rate}} \in \{2, 4, 8, 16, 32\}$) share one training recipe. At inference the excess fidelity-need field is the DeepGaze construction of Sec. 5; synthetic fields serve only as a training regularizer. **Data, rate, and metrics.** We evaluate on 40 images (24 Kodak (Eastman Kodak Company, 1999), 16 JPEG-AI test (JPEG, 2023)); objective evaluations are repeated on CLIC 2020 test (cli, 2020) (428 images) and DIV2K validation (Agustsson & Timofte, 2017) (100) in the appendix; these 2K images are coded by FLARE as single frames without tiling. Rates are payload bytes of the *real* entropy-coded bitstream plus required color-correction statistics (no container headers; Kodak untiled), with the exceptions specified in Appendix E and individual captions. Perceived quality comes from a reference-based DSIS study on a five-grade impairment scale: FLARE, AEIC-ME, and OneDC are each coded at two matched-rate points, and 37 observers rate all 240 stimuli (8,880 ratings), reported as MOS and per-observer z-scores (Appendix C). Objective quality uses PSNR, LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and patch-FID over 256×256 crops. Baselines are the realism codecs AEIC-ME and OneDC (Xue et al., 2025), the generative codecs MS-ILLM (Muckley et al., 2023) and GLC (Jia et al., 2024), and the VVC intra anchor VTM (Bross et al., 2021); multi-step diffusion codecs are omitted from the quantitative comparisons (Appendix G). MRIDC and Relic et al. are not compared directly: neither provides an implementation or weights, and their spatial control lives inside multi-step diffusion samplers with no counterpart in a one-step codec, so a reimplementation on our base would benchmark a construction of our own; Table 1 places both on the complexity axis from reported counts, the MRIDC authors' timing, and a measured lower bound for Relic et al. Appendices A.3, B, and E detail datasets, training, rate accounting, field construction, and baselines.

6.2 Codec-level validation

We test four claims from Sec. 1. **(i) Targeting** (Fig. 5a): coded rate is placed where $\breve{\mathfrak{n}}$ points—moving the field moves both the rate and the restored detail, and the concentration is position-invariant ($\approx 97\%$ of message bits inside a target covering $\approx 4.5\%$ of the image, $\approx 22\times$ its area share). **(ii) Monotonicity** (Fig. 5b): the coded message rate grows monotonically with the field magnitude, parameterized as η^3 with a scalar dial η; because the rate is strongly concave in the magnitude, the cubic form is chosen empirically so that in-region rate grows roughly linearly in η, making η a smooth rate dial. **(iii) Effective fidelity** (Fig. 5c): the redirected rate reduces distortion *inside* the requested region (e.g., +9–10 dB PSNR on sail and nose-art text) while the background is left unchanged, so fidelity is bought precisely where requested without taxing the realism floor, consistent with the additive structure of Eq. 5. **(iv) Realism-floor preservation** (Appendix Fig. 25): setting $\breve{\mathfrak{n}} \equiv 0$ makes the message rate negligible, and the resulting zero-excess operating point remains competitive with AEIC-ME in both distortion and perceptual quality, so joint training preserves the baseline objective and the floor remains a valid RDP operating point; switching the field on adds rate without degrading global or background quality, and on the 528 CLIC 2020 and DIV2K images the study field raises attended-region PSNR by 4.1–4.4 dB over the rate-matched floor (Appendix Table 9). Claims (i)–(iii) use synthetic fields with known location and magnitude (protocol in Appendix E.2). **Computational efficiency.** Table 1 compares model size and end-to-end coding time. By preserving AEIC-ME's fixed one-step decoding graph, FLARE remains close to its backbone in size and latency: it adds 0.87M parameters (0.09% of AEIC-ME's 956.58M; Table 2) and 17 ms per image, while being lighter and faster than OneDC and DiT-IC and avoiding the iterative-sampling overhead of multi-step diffusion codecs. **Limitations.** The perceived-quality instantiation pays for generating the field: DeepGaze takes $\approx$ 255 ms per Kodak image, more than the codec

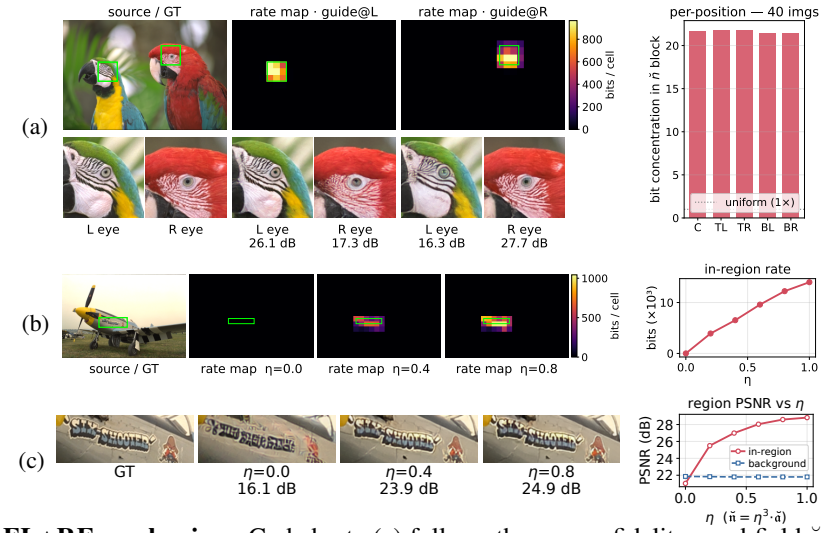


Figure 5: **FLARE mechanism.** Coded rate (a) follows the excess fidelity-need field $\breve{\mathfrak{n}}$ in location, (b) scales with its magnitude, and (c) yields effective in-region fidelity gains.

Table 1: **Per-image wall-clock time on Kodak with real entropy coding.** Measured on one NVIDIA RTX PRO 6000 Blackwell. Parameters are counted as instantiated at inference (in parentheses: after merging the adapters, where applicable). "Steps" denotes diffusion denoising steps.

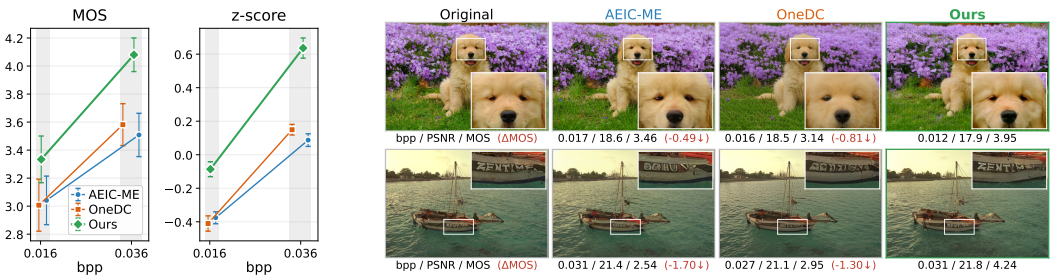
Method	Steps	Params (M)	Encode (ms)	Decode (ms)	Total (ms)
FLARE (ours)[†]	1	957.6 (926.5)	18.1	81.1	99.2
AEIC-ME (base) (Zhang et al., 2026)	1	956.6 (925.7)	13.5	68.5	82.0
OneDC (Xue et al., 2025)	1	1406.4 (1339.2)	56.0	133.8	189.8
DiT-IC (Shi et al., 2026)	1	1168.3 (1049.5)	67.2	107.1	174.4
PerCo (Careil et al., 2024)	20	1295.0	278	861	1139
DiffEIC (Li et al., 2024)	50	1379.5	116	3107	3223
DiffC (Vonderfecht & Liu, 2025)	77/86	1066.2	2640	2100	4740
MRIDC (Xu et al., 2025)[‡] (mask 30%)	20	1551.6	83	1350	1433
Relic et al. (2026)[§]	5–50	953.7	—	>127–960	>127–960
VTM (intra, CPU) (Bross et al., 2021)	—	—	4402.6	96.6	4499.2

[†]Excludes field generation: DeepGaze takes ≈ 255 ms per Kodak image (Sec. 6.2). [‡,§]No public implementation; numbers as reported by the authors. [‡]Authors' timing on an NVIDIA A800. [§]Lower bound: $N \times$ the SD v2.1 UNet (18.5 ms) plus one VAE decode (34.9 ms).

itself, so online field generation is a meaningful end-to-end overhead. Table 1 reports codec-side cost only because the predictor is an independently executed, off-the-shelf front end behind a fixed field interface; the field can be precomputed or cached, and cheaper or more reliable predictors can be substituted without retraining FLARE or modifying its decoder.

6.3 Subjective-quality evaluation

At both matched-rate operating points, FLARE achieves the highest MOS and z-score against AEIC-ME and OneDC (Fig. 6a), improving MOS by 0.29–0.57; for every competitor and rate point, at least 35 of the 37 observers gave FLARE a higher mean rating over the 40 images than the competitor (35, 36, 36, and 36 observers). Here, the z-score normalizes the variance of each observer's ratings. The examples in Fig. 6b illustrate the corresponding source-fidelity failures: AEIC-ME and OneDC remain globally plausible but deform the dog's eyes and turn the boat lettering into incorrect text, whereas FLARE preserves both. **Attribution to excess-fidelity allocation.** The subjective gain could in principle stem from FLARE's realism floor rather than from where it places fidelity. We test this on the study data itself by splitting the 40 images into terciles of the *excess-rate share* actually spent by the study stimulus (the fraction of FLARE's bits attributable to the excess field, measured on the same checkpoint with the field set to zero) and recomputing the matched-rate MOS gain per tercile (Fig. 7). On images where little excess was spent, FLARE is on par with or slightly below the baselines; the gain grows monotonically with excess share and reaches +0.9–+1.3 MOS in the top tercile. A per-image regression that controls for the rate difference to the baseline attributes the gain to excess share (+2.2 MOS per unit share, $p < 10^{-5}$), whereas FLARE's zero-excess floor is



(a) (b)

Figure 6: **Subjective quality evaluation results.** (a): MOS and z-scores at the two matched rates (shaded), with 95% confidence intervals. (b): Reconstructions (bpp / PSNR / MOS and MOS gap).

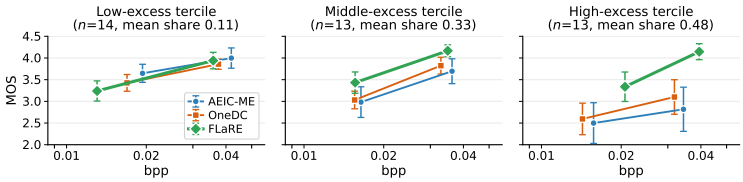


Figure 7: **Rate vs. MOS by excess-rate share.** Images are split into terciles of the share of FLARE's bits spent on the excess field; each panel plots the two study points of every codec for one tercile (95% confidence intervals). Appendix D tabulates the matched-rate MOS gains.

objectively on par with the baselines at matched rate, on the study set as well as on CLIC 2020 and DIV2K. The gain therefore follows the allocation, not the floor; Appendix D gives the protocol, the floor comparison, and field controls: on the same weights, random or centered fields of the same active area and magnitude leave the attended region at its field-off quality (Table 11).

6.4 Application-specific fidelity allocation

The same encoder-side interface also accepts a *task-specific* excess fidelity-need field from an application detector (Sec. 5), directing rate to regions a downstream recognizer must read correctly. On three forced-recognition tasks—license-plate OCR (CCPD (Xu et al., 2018)), face identity (ArcFace (Deng et al., 2019)), and logo-brand classification (LogoDet-3K (Wang et al., 2022))—a simple binary region-of-interest field raises recognition far above the realism floor ($\eta{=}0$) and above rate-matched neural and traditional codecs, most of all where the attended region is *small*: near 0.01–0.03 bpp FLARE reads the plate string and preserves face identity while AEIC-ME, OneDC, MS-ILLM, and VTM fail on the same crops (Appendix Fig. 22). Protocol/curves are in Appendix F.

7 Conclusion

Generative image codecs attain realism at low rates by optimizing an RDP objective whose distortion is priced uniformly across the image, so they cannot tell a harmless plausible substitution in a texture from the loss of source-specific detail in a face or a word and may fabricate message-carrying content. We addressed this by extending RDP with a spatial fidelity-need field and decomposing its weighted distortion into a uniform floor and a nonnegative excess: the floor keeps the established realism-codec objective intact, and only the excess remains as a spatial-allocation problem. FLARE realizes this decomposition with a dedicated message latent whose rate is gated and scaled by the field, supervised through a training-only auxiliary decoder that imposes no strong local regression on the delivered image, and decoded by the unchanged one-step generator without any transmitted map. Coded rate thus lands where the field points and grows monotonically with its magnitude, the requested regions gain fidelity while the realism floor around them is left intact, and with zero excess the codec stays competitive with its backbone. Driven by predicted attention, FLARE was preferred over state-of-the-art realism codecs by nearly all observers in a matched-rate study; driven by a task detector, it raised downstream recognition without retraining.

AI USE STATEMENT

We used generative AI tools (large language model assistants) in this work. The table lists every task for which the ICLR 2027 AI policy requires disclosure.

Task with required disclosure	Used	Scope
Generate synthetic data sets	No	—
Help develop theoretical models or conceptual frameworks	No	Sec. 3 is the authors' own
Formulate mathematical claims	No	—
Provide critical ingredients for proving mathematical claims	No	—
Assist in the writing of proofs	No	—
Propose or refine hypotheses	No	the research hypotheses and the attribution analysis of Appendix D were designed by the authors
Design or provide feedback on research methodology or experiments	Yes (limited)	feedback on the analysis protocols of Appendices D, E.1, E.2
Implement methods	No	the instantiation of the method (latent split, gating and scaling, auxiliary decoder, training recipe) was designed and implemented by the authors; AI assisted only with code edits under their direction (second table)
Assist with translation	No	—
Clean and reformat dataset	No (N/A)	—
Support qualitative and thematic data analysis	No (N/A)	—
Interpret results	Yes (limited)	assisted the authors in reading intermediate results; all conclusions were drawn and verified by the authors

The second table covers the tasks with recommended disclosure.

Task with recommended disclosure	Used	Scope
Formulate questions for surveys or interviews	No	—
Create or modify scientific figures or images	Yes	plotting scripts for result figures, under the authors' specification
Suggest experimental parameters	No	—
Create or edit software code	Yes	evaluation and analysis scripts; edits to training code under the authors' direction
Creation of artifacts	No	—
Draft parts of a research paper	Yes	drafts of text passages, later revised by the authors
Transcribe recordings of research material	No (N/A)	—
Summarize or analyse existing literature	Yes (limited)	summaries of the two closest related methods for the availability audit
Discover research topics or identify gaps	No	—
Brainstorming	Yes	discussion of analysis options with the authors
Sourcing/searching for information	Yes	verifying availability, step counts, and settings of related methods
Edit a research paper to improve readability	Yes	condensing and copy-editing
Identify relevant literature	No	—
Format references	No	—
Suggest a structure for a research paper	Yes (limited)	placement of individual paragraphs and figures
Propose a title or keywords for a research paper	No	—

We have reviewed all AI-assisted work: all code was reviewed and tested by the authors, every reported number was produced by the described pipelines on the stated data, and all AI-drafted text was checked against the experiments and the cited sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The subjective study (Sec. 6.1, Appendix C) involved 37 observers rating compressed natural images against their originals; the task carries no risk beyond ordinary screen viewing. Participants were informed of the study purpose, procedure, duration, and voluntary nature of participation before taking part, and provided informed consent electronically. Ratings are stored under random participant identifiers without personal information. The stimuli are drawn from the public Kodak and JPEG-AI test sets; the application-specific experiments use the public CCPD, COCO, and LogoDet-3K benchmarks under their respective licenses. FLARE preserves source detail where a supplied field requests it; with a face or license-plate detector as the field source this can retain identifying content at low rates, which is the intended benefit for authorized applications but could also serve surveillance. The codec itself does not identify anyone, and the same interface can add no fidelity beyond the floor in regions that should not be preserved. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The architecture, latent split, gating, scale transport, and auxiliary decoder are specified layer by layer in Appendix A.3, together with the full training configuration (data, losses, weights, schedules, and operating points) and the verified statement that the delivered-image fidelity term is used only in pretraining. The field construction is given in Appendix B with all constants, and the synthetic-field protocol in Appendix E.2. Every rate is measured from the real entropy-coded bitstream under the accounting in Appendix E, which also lists the baseline checkpoints and settings and the availability audit for methods we could not run. The subjective protocol, interface, observer set, and analysis are documented in Appendix C, and the attribution analysis with its per-image regression in Appendix D. The supplementary material contains the training and evaluation code, the field construction, the real-bitstream coding pipeline, and the training configurations of every released model. The implementation extends the public AEIC-ME code base (Zhang et al., 2026) (MIT license), from which the backbone transforms, context model, one-step decoder, and training loop are inherited; the checkpoints of the five operating points (5×564 MB, SHA-256 digests included) are hosted in an anonymized OSF project linked from its README (`https://osf.io/s7xmg/?view_only=5e709f442f58444cbe5796029135fa99`). All of it will be released publicly upon publication.

REFERENCES

Workshop and challenge on learned image compression (CLIC). https://compression.cc, 2020.

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *ICCV*, 2019.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *ICML*, 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 (TCSVT)*, 31(10):3736–3764, 2021.

Marlène 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 (ICLR)*, 2024.

Zheng Chen, Mingde Zhou, Jinpei Guo, Jiale Yuan, Yifei Ji, and Yulun Zhang. Steering one-step diffusion model with fidelity-rich decoder for fast image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

Charilaos Christopoulos, Athanassios Skodras, and Touradj Ebrahimi. The JPEG2000 still image coding system: an overview. *IEEE Transactions on Consumer Electronics*, 2000.

Ze Cui, Jing Wang, Shangyin Gao, Bo Bai, Tiansheng Guo, and Yihui Feng. Asymmetric gained deep image compression with continuous rate adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

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 (TPAMI)*, 44(5):2567–2581, 2022.

Yasser Abdelaziz Dahou Djilali, Kevin McGuinness, and Noel O'Connor. Learning saliency from fixations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 383–393, 2024.

Eastman Kodak Company. Kodak lossless true color image suite. https://r0k.us/graphics/kodak/, 1999.

Wilson S. Geisler and Jeffrey S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. In *Human Vision and Electronic Imaging III, Proc. SPIE*, volume 3299, pp. 294–305, 1998. doi: 10.1117/12.320120.

ITU-R. Methodologies for the subjective assessment of the quality of television pictures. Technical Report Recommendation ITU-R BT.500-15, International Telecommunication Union, 2023.

ITU-T. Subjective video quality assessment methods for multimedia applications. Technical Report Recommendation ITU-T P.910, International Telecommunication Union, 2023.

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

JPEG. JPEG AI common training and test conditions. Technical report, ISO/IEC JTC 1/SC 29/WG 1, 2023.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009.

Matthias Kümmerer, Harneet Singh Khanuja, and Matthias Bethge. Modeling saliency dataset bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22077–22088, 2025.

Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. LSDIR: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1775–1787, 2023.

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

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

Jonas Löhdefink, Andreas Bär, Nico M. Schmidt, Fabian Hüger, Peter Schlicht, and Tim Fingscheidt. Focussing learned image compression to semantic classes for V2X applications. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1641–1648, 2020. doi: 10.1109/IV47402.2020.9304779.

Yi Ma, Yongqi Zhai, Chunhui Yang, Jiayu Yang, Ruofan Wang, Jing Zhou, Kai Li, Ying Chen, and Ronggang Wang. Variable rate ROI image compression optimized for visual quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1936–1940, 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/CLIC/html/Ma_Variable_Rate_ROI_Image_Compression_Optimized_for_Visual_Quality_CVPRW_2021_paper.html.

Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *NeurIPS*, 2020.

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 *ICML*, 2023.

Yash Patel, Srikar Appalaraju, and R. Manmatha. Saliency driven perceptual image compression. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 227–236, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Patel_Saliency_Driven_Perceptual_Image_Compression_WACV_2021_paper.html.

Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Semantic perceptual image compression using deep convolution networks. In *Data Compression Conference (DCC)*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Lucas Relic, Roberto Azevedo, Yang Zhang, Stephan Mandt, Markus Gross, and Christopher Schroers. Region-adaptive generative compression with spatially varying diffusion models, 2026. URL https://arxiv.org/abs/2604.01122.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024. doi: 10.1007/978-3-031-73016-0_6. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/11557_ECCV_2024_paper.php.

Xihua Sheng, Lingyu Zhu, Tianyu Zhang, Dong Liu, Shiqi Wang, and Jing Wang. CADC: Content adaptive diffusion-based generative image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Junqi Shi, Ming Lu, Xingchen Li, Anle Ke, Ruiqi Zhang, and Zhan Ma. DiT-IC: Aligned diffusion transformer for efficient image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Lucas Theis, Tim Salimans, Matthew D. Hoffman, and Fabian Mentzer. Lossy compression with gaussian diffusion, 2022.

Jeremy Vonderfecht and Feng Liu. Lossy compression with pretrained diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. In *ACM TOMM*, 2022.

Zhou Wang and Alan C. Bovik. Embedded foveation image coding. *IEEE Transactions on Image Processing*, 10(10):1397–1410, 2001. doi: 10.1109/83.951527.

Jinchang Xu, Shaokang Wang, Jintao Chen, Zhe Li, Peidong Jia, Fei Zhao, Guoqing Xiang, Zhijian Hao, Shanghang Zhang, and Xiaodong Xie. Decouple distortion from perception: Region adaptive diffusion for extreme-low bitrate perception image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18051–18061, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xu_Decouple_Distortion_from_Perception_Region_Adaptive_Diffusion_for_Extreme-low_Bitrate_CVPR_2025_paper.html.

Zhenbo Xu, Wei Yang, Ajin Meng, Nanxue Lu, Huan Huang, Changchun Ying, and Liusheng Huang. Towards end-to-end license plate detection and recognition: A large dataset and baseline. In *ECCV*, 2018.

Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. One-step diffusion-based image compression with semantic distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ruihan Yang and Stephan Mandt. Lossy image compression with conditional diffusion models, 2023. NeurIPS 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. StableCodec: Taming one-step diffusion for extreme image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Tianyu Zhang, Dong Liu, and Chang Wen Chen. Ultra-low bitrate perceptual image compression with shallow encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xi Zhang and Xiaolin Wu. Attention-guided image compression by deep reconstruction of compressive sensed saliency skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13354–13364, 2021.

A Architecture, Entropy Model, and Training

A.1 FLaRE-specific module specification

This subsection specifies the FLaRE model evaluated throughout the experiments. We describe only components introduced for FLaRE. The AEIC-ME analysis/synthesis transforms, quadtree context model, hyperprior backbone, SD-Turbo denoiser, LoRA adaptation, and VAE renderer follow their cited sources and are not repeated. Let B denote batch size. For an input of spatial size $H \times W$, let $h = H/32$ and $w = W/32$. Our implementation uses a 320-channel latent split into 48 base channels and 272 message channels, a scale-network width of 32, an auxiliary-decoder base width of 256, max pooling to the latent grid for the activation mask, an activation threshold $\tau = 10^{-3}$ during training (inference uses the operating-point-dependent τ_q of Eq. 20), and 128-pixel symmetric field padding. Let $\ell \in \{1, \dots, 5\}$ index the analysis-transform layers, and let γ_ℓ and β_ℓ denote their FiLM scale and bias projections.

Table 2: FLaRE-specific modules in our implementation. Parameter counts include biases. Counts for modified inherited modules report only the additional parameters relative to the AEIC-ME codec. The deployed additions total 866,248 parameters; the 7,758,764-parameter auxiliary decoder is discarded after training.

FLaRE component	role	tensor interface		added params.	deployed
Multi-scale field FiLM	conditions the analysis transform at five resolutions	$\breve{\mathfrak{n}} \to \{\gamma_\ell, \beta_\ell\}_{\ell=1}^{5}$		556,480	yes
Scale network S	predicts a channelwise message scale on the latent grid	$B\times1\times H\times W$ $B\times272\times h\times w$	$\to$	46,288	yes
Hyper-analysis extension	supplies scaled message content, scale, and gate state to the shared hyperprior	$B\times593\times h\times w$		98,280	yes
Activation mask	zeroes the field and the coded message residual outside the active region (Eqs. 8 and 12)	field and message residuals		0	yes
Scale-reconstruction head	reconstructs the message scale $\hat{\mathbf{s}}$ from the decoded hyperlatent	$B\times320\times h\times w$ $B\times272\times h\times w$	$\to$	88,880	yes
Synthesis FiLM head	conditions the first synthesis feature on the reconstructed scale	$B\times272\times h\times w$ $(\gamma, \beta)_{1024}$	$\to$	76,320	yes
Auxiliary decoder	produces the training-only auxiliary reconstruction	$B\times272\times h\times w$ $B\times3\times H\times W$	$\to$	7,758,764	no

Fidelity-need-conditioned analysis. The masked excess fidelity-need field $\widetilde{\mathfrak{n}}$ of Eq. 8 modulates the analysis transform after each of its five downsampling stages. For the encoder feature $\mathbf{h}_\ell$ at the ℓth analysis-transform layer, the corresponding head applies

$$\mathbf{h}'_\ell = \mathbf{h}_\ell \odot \big(1 + \gamma_\ell(\widetilde{\mathfrak{n}})\big) + \beta_\ell(\widetilde{\mathfrak{n}}). \tag{16}$$

The final convolutions producing γ_ℓ and β_ℓ are initialized to zero, so all five heads initially preserve the inherited encoder. Table 3 gives the exact heads.

Table 3: Multi-scale FiLM heads in the FLaRE analysis transform. "Resize" denotes bilinear resizing of the single-channel field to the target feature resolution.

scale	field projection	modulation heads	target feature	params.
$H/2$	3×3 conv., stride 2, $1\to32$; GELU	two 3×3 convs., $32\to64$	64 ch.	37,312
$H/4$	resize; 3×3 conv., $1\to32$; GELU	two 3×3 convs., $32\to128$	128 ch.	74,304
$H/8$	resize; 3×3 conv., $1\to32$; GELU	two 3×3 convs., $32\to192$	192 ch.	111,296
$H/16$	resize; 3×3 conv., $1\to32$; GELU	two 3×3 convs., $32\to256$	256 ch.	148,288
$H/32$	resize; 3×3 conv., $1\to32$; GELU	two 3×3 convs., $32\to320$	320 ch.	185,280

Scale network and message gating. The activation mask $\mathbf{m}$ and masked field $\widetilde{\mathfrak{n}}$ are formed as in Eq. 8; both the analysis transform and the scale network S consume $\widetilde{\mathfrak{n}}$. The scale network applies the sequence in Table 4. Symmetric padding supplies real field context at every boundary; the four-cell crop corresponds to $128/32$ latent cells on each side.

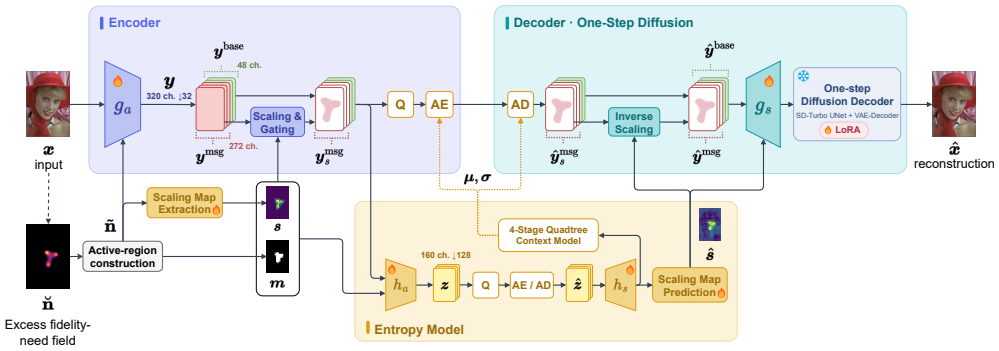


Figure 8: **FLARE architecture overview.** The excess fidelity-need field $\breve{\mathbf{n}}$ yields a scale map $\mathbf{s}$ and activation mask $\mathbf{m}$ that scale and gate the 272-channel message latent $\mathbf{y}^{\mathrm{msg}}$, while the 48-channel base latent is coded unchanged. A saliency-aware entropy model—a shared hyperprior with a four-stage autoregressive spatial–channel context—codes the concatenated latent, and the fixed one-step diffusion decoder (SD-Turbo (Sauer et al., 2024) UNet with our LoRA plus the VAE decoder) renders the reconstruction $\hat{\mathbf{x}}$. Flame/snowflake icons mark trained/frozen modules. The analysis/synthesis transforms and one-step decoding graph follow the AEIC-ME backbone; only the message path is conditioned on the field.

Table 4: Scale network S used in our implementation. The five stride-2 convolutions directly aggregate the full-resolution field instead of first averaging it onto the latent grid.

stage	operator	output	params.
input	symmetric (reflect) padding of 128 pixels on all four sides	$B\times1\times(H+256)\times(W+256)$	0
1	3×3 conv., stride 2, pad 1, $1\to32$; GELU	$B\times32\times\frac{H+256}{2}\times\frac{W+256}{2}$	320
2–5	four 3×3 convs., stride 2, pad 1, $32\to32$; GELU	$B\times32\times\frac{H+256}{32}\times\frac{W+256}{32}$	36,992
head	1×1 conv., $32\to272$; softplus	$B\times272\times\frac{H+256}{32}\times\frac{W+256}{32}$	8,976
output	center-crop four latent cells per side	$\mathbf{s}\in\mathrm{R}_{+}^{B\times272\times h\times w}$	0

The hyper-analysis transform receives $[\mathbf{y}^{\mathrm{base}}, \mathbf{y}_{\mathrm{s}}^{\mathrm{msg}}\odot\mathbf{m}, \mathbf{s}, \mathbf{m}]$, totaling $48+272+272+1=593$ channels. The same $\mathbf{m}$ masks the entropy-coded message residual rather than its predicted mean (Eq. 12), so where $\mathbf{m}=0$ the coded message equals its predicted mean and its residual rate approaches zero without suppressing the always-coded base group.

Scale recovery and fixed decoding. The decoder never receives $\breve{\mathbf{n}}$. Instead, the shared hyperprior reconstructs an scale field from the decoded hyperlatent, de-scales the message representation, and uses the same field to modulate the first feature of the inherited synthesis transform. The exact additions are listed in Table 5. Fig. 9 visualizes, for two study images, the field, the activation mask, the encoder-side scale map $\mathbf{s}$, the decoder-side estimate $\hat{\mathbf{s}}$ recovered from the hyperlatent, and the message bits per latent cell. $\hat{\mathbf{s}}$ follows the spatial pattern of $\mathbf{s}$ but differs from it in absolute scale: the training objective constrains only the reconstruction, not the agreement of the two maps, so the optimization does not force them onto a common scale, and the synthesis transform is trained against $\hat{\mathbf{s}}$ as recovered. This has no effect on the rate: the coded message symbols are formed on the encoder side from $\mathbf{s}\odot\mathbf{y}^{\mathrm{msg}}$, so $\mathbf{s}$ alone sets the effective quantization step and hence the message bits, which are confined to the active cells.

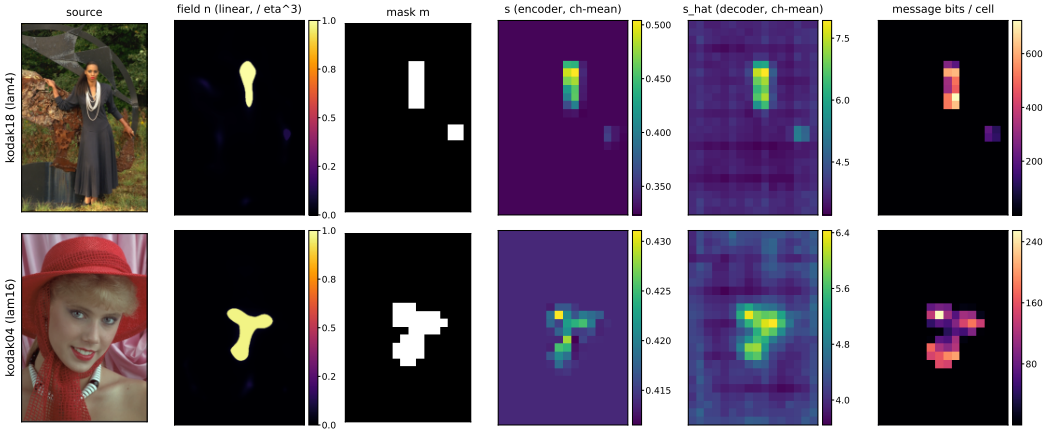


Figure 9: **Scale maps on two study images** (Kodak 18 at $\lambda_{\text{rate}} = 4$, Kodak 04 at $\lambda_{\text{rate}} = 16$). Columns: source; excess fidelity-need field $\breve{\mathbf{n}}$ normalized by η_q^3 (linear scale); activation mask $\mathbf{m}$; encoder-side scale map $\mathbf{s}$ and decoder-side estimate $\hat{\mathbf{s}}$, each averaged over the 272 message channels; and message bits per latent cell (sum of the coded symbols' code lengths). Color scales are per panel. $\hat{\mathbf{s}}$ reproduces the spatial pattern of $\mathbf{s}$ at a different absolute scale, which the objective does not constrain; the rate is determined by $\mathbf{s}$ on the encoder side.

Table 5: Scale transport through the shared hyperprior and the fixed synthesis transform. The last FiLM convolution is zero-initialized, preserving the inherited decoder at initialization. The reconstructed $\hat{\mathbf{s}}$ is an internal decoded feature, not a transmitted spatial map.

operation	exact structure	output	added params.
Hyper-analysis input	concatenate $[\mathbf{y}^{\text{base}}, \mathbf{y}_{\text{s}}^{\text{msg}} \odot \mathbf{m}, \mathbf{s}, \mathbf{m}]$; expand the inherited first 3×3 conv. from 320 to 593 inputs	$B\times593\times h\times w$	98,280
Scale-reconstruction head	1×1 conv. 320 → 32; GELU; 3×3 conv. 32 → 272; exponential	$\hat{\mathbf{s}} \in \mathrm{R}_{+}^{B\times272\times h\times w}$	88,880
Message recovery	$\hat{\mathbf{y}}^{\text{msg}} = \hat{\mathbf{y}}_{\text{s}}^{\text{msg}} / \max(\hat{\mathbf{s}}, 10^{-3})$	$B\times272\times h\times w$	0
Synthesis FiLM	1×1 conv. 272 → 32; GELU; 1×1 conv. 32 → 2048; split into $(\delta\gamma, \beta)$ of 1024 channels each; apply $\mathbf{f}_0' = (1+\delta\gamma)\odot \mathbf{f}_0 + \beta$	$B\times1024\times h\times w$	76,320

Training-only auxiliary decoder. The auxiliary decoder receives the post-quantization, de-scaled message group $\hat{\mathbf{y}}^{\mathrm{msg}}$. Its input is not hard-masked; spatial selectivity is imposed by the field-weighted reconstruction objective in Eq. 15. Its output uses a final $\tanh$ and lies in $[-1, 1]$. Table 6 gives the complete sequence. Adapter, StarBlock, and Upsample are inherited AEIC-ME blocks (Zhang et al., 2026); we report their placement and channel dimensions without reproducing their internal definitions.

Table 6: Complete training-only auxiliary decoder used in our experiments. Each Upsample is the inherited 3×3 convolution followed by $2\times$ PixelShuffle. The module has 7,758,764 parameters and is absent from the deployed codec.

stage	operator	output	params.
0	input message latent	$B\times272\times h\times w$	0
1	Adapter $272\to256$	$B\times256\times h\times w$	216,664
2	StarBlock$(256)\times2$	$B\times256\times h\times w$	792,192
3	Upsample $256\to256$	$B\times256\times2h\times2w$	2,360,320
4	StarBlock$(256)\times2$	$B\times256\times2h\times2w$	792,192
5	Upsample $256\to192$	$B\times192\times4h\times4w$	1,770,240
6	StarBlock$(192)\times2$	$B\times192\times4h\times4w$	446,688
7	Upsample $192\to128$	$B\times128\times8h\times8w$	885,248
8	StarBlock(128)	$B\times128\times8h\times8w$	99,744
9	Upsample $128\to64$	$B\times64\times16h\times16w$	295,168
10	StarBlock(64)	$B\times64\times16h\times16w$	25,296
11	Upsample $64\to32$	$B\times32\times32h\times32w$	73,856
12	Adapter $32\to3$; $\tanh$	$B\times3\times H\times W$	1,156

A.2 Latent occupancy and channel split

The 48/272 split in Sec. 4.1 is based on the occupancy of the released, unmodified AEIC-ME realism codec (Zhang et al., 2026). Before adding the fidelity-need input, scaling network, or message objective, we encode Kodak (Eastman Kodak Company, 1999) and sum the Gaussian entropy model's $-\log_2$ likelihood over spatial locations to measure the mean rate of each latent channel.

Table 7: Per-channel latent occupancy of the unmodified AEIC-ME realism codec on Kodak. Only 18–25 of 320 channels carry 99% of the latent rate at the evaluated operating points, and the 48 highest-rate channels carry at least 99.9%.

base point	latent rate (bpp)	channels for 99%	top-48 share	near-inactive
$\lambda2$	0.032	25	100.0%	$295/320$ (92%)
$\lambda8$	0.014	24	100.0%	$296/320$ (92%)
$\lambda16$	0.007	23	100.0%	$297/320$ (93%)
$\lambda32$	0.003	18	99.9%	$302/320$ (94%)

Fig. 10 shows the same result as sorted per-channel rates and cumulative rate shares. Between 295 and 302 channels are near-inactive across the four operating points, so assigning 48 channels to the base role leaves substantial margin over the observed occupancy while providing 272 dimensions for the message role. This diagnostic determines only the group sizes; FLARE jointly learns the two roles rather than freezing a subset of base-codec channels.

A.3 Training configuration

Data and fields. Pretraining uses the training corpus of AEIC-ME (Zhang et al., 2026): the 2K-resolution sets Flickr2K (Lim et al., 2017), DIV2K (Agustsson & Timofte, 2017) (train), and CLIC (cli, 2020) (2020 train and validation, 2021 test), $4{,}119$ images after discarding those smaller than 512 pixels on either side, plus the first $10{,}000$ images of LSDIR (Li et al., 2023), for $14{,}119$ images in total. All fine-tuning stages use the $4{,}119$-image 2K set only. Every training image is paired with a precomputed DeepGaze MSDB attention map (long edge 1024 pixels, stored at 8-bit precision). Training samples are 512×512 random crops with random horizontal and vertical flips applied jointly to the image and its map; a crop is redrawn up to eight times until at least 0.5% of its pixels have attention above 0.5, and the map passes through a smoothstep gate that zeroes values below 0.1 and leaves values above 0.3 unchanged. On 60% of steps the excess fidelity-need field is

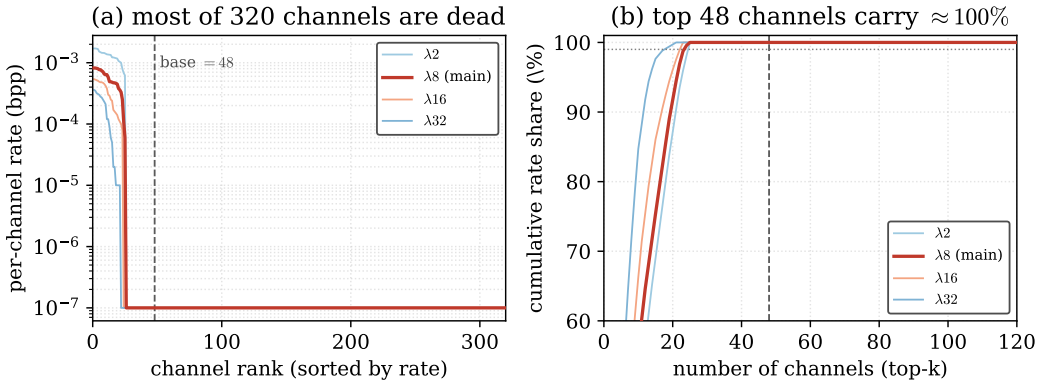


Figure 10: **Latent occupancy of the unmodified realism codec. (a)** Sorted per-channel rate on Kodak: the rate collapses after approximately 25 channels, well before the 48-channel base allocation (dashed). **(b)** Cumulative rate share: the 48 highest-rate channels account for at least 99.9% at every operating point.

a synthetic field (random blocks and shapes with random strength); on the remaining 40% it is a dilated DeepGaze field. The allocation mechanism therefore sees both structured and attention-derived inputs and is not tied to a single field generator. Sec. 5 and Appendix B describe the inference-time field construction.

Initialization. The analysis transform and the auxiliary decoder are initialized from a dense-supervised stage that involves no diffusion decoder: a gained variable-rate autoencoder (Cui et al., 2021) with the same 48/272 channel split is trained for 1M steps on 256×256 crops of the same corpus (batch size 128 on two GPUs) with an MSE objective, the auxiliary decoder reconstructing the image from the quantized latent. This choice matches the mechanism of FLARE: the field-driven scaling of the message latent sets the effective quantization step size of its elements, and a gained autoencoder likewise controls rate by rescaling the representation before quantization, so its analysis transform is already trained to produce a latent whose step size, not its structure, varies with the rate. Only its analysis transform and auxiliary decoder are transferred; its gain vectors and entropy model are discarded, and the synthesis transform, hyperprior, context model, scale network, and all field-related heads start from their own initialization, with the SD-Turbo denoiser and VAE renderer from their public weights. All pretrained variants used in the ablations (Appendices E.3 and H) share this initialization.

Stages and operating points. The codec is then pretrained for 370k steps (batch size 8, no discriminator) and adversarially fine-tuned (batch size 32). The base operating point $\lambda_{\text{rate}} = 8$ is fine-tuned from the pretrained codec for 50k steps; each remaining operating point is warm-started from the preceding one and fine-tuned for a further 20k steps, changing only λ_{rate} ($8 \to 16 \to 32$ and $8 \to 4 \to 2$). All five operating points share the recipe below.

Objective. Let R^{base}, R^{z}, and R^{msg} denote the per-sample rates of the base latent, hyperlatent, and scaled message latent. The fine-tuning loss is

$$\lambda_{\text{rate}}\big(R^{\text{base}}+R^{\text{z}}\big)+2\,\ell_2+1\,\text{DISTS}+0.1\,\text{CLIP}+0.1\,\mathcal{L}_{\text{adv}} + R^{\text{msg}}+A\,\text{MSE}_{\text{aux}}, \qquad A = 0.03{\cdot}255^2, \tag{17}$$

where CLIP is the CLIP-feature loss (Radford et al., 2021) inherited from AEIC-ME and MSE_{aux} is the field-weighted auxiliary-decoder error of Eq. 15; crops whose active field fraction is below 0.5% are excluded from this term. The first group of terms realizes the baseline realism-codec block and the last two the excess-fidelity block of Eq. 14. Note that λ_{rate} scales only the floor rate: operating points move the floor, whereas the message rate–distortion trade-off is fixed and the rate spent in requested regions is set by the field magnitude. Eq. 6 uses a single multiplier because it describes one operating point. Across operating points we deliberately hold the weight of R^{msg} relative to MSE_{aux} fixed so that a given field value requests the same additional fidelity at every operating point, which is what makes the field a rate-independent control; the operating points thus trace a two-parameter family (λ^{floor} varied, λ^{excess} fixed) rather than a single-λ curve, and the price of a

message bit relative to a base bit changes by $16\times$ from $\lambda_{\text{rate}} = 2$ to 32 by design. Pretraining uses the same objective without the adversarial term and with LPIPS in place of DISTS.

Delivered-image fidelity term used only in pretraining. Pretraining additionally includes a field-weighted squared error on the delivered reconstruction, $0.01 \cdot 255^2 \cdot \frac{1}{|\Omega|} \sum_u \breve{\mathfrak{n}}(u) \|\hat{\mathbf{x}}(u) - \mathbf{x}(u)\|_2^2$, as an initial optimization aid that lets the message group influence the output early in training. Its weight is zero throughout adversarial fine-tuning, so no field-weighted regression on the delivered image is present in the final training stage of any released operating point; the excess-fidelity supervision of the released models comes only from the auxiliary decoder (Sec. 4.2).

Optimization. We use Adam with gradient clipping at 0.1 on four GPUs. Pretraining uses a learning rate of 10^{-4}, reduced to 5×10^{-5} for the final 20k steps. Fine-tuning uses 5×10^{-5} for the codec and 2×10^{-5} for the discriminator, decayed stepwise within each operating point to 10^{-6} ($\lambda_{\text{rate}} = 8$) or 5×10^{-6} (all others). Every forward pass, in training and evaluation, pads the field symmetrically by 128 pixels before the scale network and applies the activation mask of Eq. 8 to the field, so that scale and rate maps match between training and inference.

B Perceptual Fidelity-Need Construction and Lift Analysis

B.1 Attention-map preprocessing and field construction

DeepGaze output and corpus calibration. DeepGaze MSDB (Kümmerer et al., 2025) is run in its generalization mode (`dataset=None`) with the MIT1003 center bias (Judd et al., 2009). Our fixed inference configuration uses 15 pixels per degree and resizes the input long side to 1024 pixels, with both dimensions rounded to multiples of 16. Exponentiating the returned log-density gives p_i^{DG} on the prediction grid V_i, which is bilinearly resized to the coded image resolution as p_i. Our preprocessing then calibrates its magnitude across images. Rather than applying per-image min–max normalization, which would assign the same unit peak to both strongly concentrated and diffuse predictions, we use the fixed maximum $A_{\max} = 5.6736 \times 10^{-4}$ measured over the 14,119-image training corpus under the same inference configuration:

$$\mathfrak{a}_i^{(0)}(u) = \text{clip}\left(\frac{p_i(u)}{A_{\max}}, 0, 1\right). \tag{18}$$

Concentration modulation and peak safeguard. We summarize the spatial concentration of the prediction by its effective-support fraction: with p_i^{DG} normalized to unit sum over the grid V_i, the participation ratio $1/\sum_u p_i^{\text{DG}}(u)^2$ divided by $|V_i|$,

$$D_i = \left(|V_i| \sum_{u \in V_i} p_i^{\text{DG}}(u)^2\right)^{-1} \in (0, 1], \qquad \kappa_i = \text{clip}\left[\left(\frac{\bar{D}}{D_i}\right)^2, 0.25, 4\right], \tag{19}$$

where $\bar{D} = 0.099562$ is the training-corpus geometric mean; $D_i = 1$ for a uniform map and small for a concentrated one. Thus concentrated predictions are amplified and diffuse predictions are attenuated. We index the operating points from the low-rate model ($q = 0$ for $\lambda_{\text{rate}} = 16$, incrementing q whenever λ_{rate} is halved) and set

$$\eta_q = \eta_{\text{base}} \Gamma^q, \qquad \tau_q = \eta_q^3 S_q, \tag{20}$$

with $\eta_{\text{base}} = 0.205$ and $\Gamma = 1.725$. Here S_q is the selection level in the normalized attention domain rather than a second field-magnitude parameter; it decreases slowly with the operating point (0.1162 at $q = 0$ to 0.1002 at $q = 2$) so that, when fewer bits are available, the excess is concentrated on fewer latent cells, and the support widens slightly as the rate budget grows. Because S_q enters the peak safeguard, the component completion, and the top-1 floor below, the processed map $\breve{\mathfrak{a}}_i$ itself differs slightly between operating points, not only its magnitude η_q^3. Let r_i be the largest latent-cell maximum in $\mathfrak{a}_i^{(0)}$. Multiple sharp peaks can produce a large D_i and excessive attenuation, so for $r_i \geq 0.075$ we apply the peak safeguard

$$\kappa_i \leftarrow \max\left(\kappa_i, \frac{S_q}{0.5 r_i}\right). \tag{21}$$

This limits the effective selection level to one half of the image's own peak. The modulated map is $\mathfrak{a}_i^{(1)} = \text{clip}(\kappa_i \mathfrak{a}_i^{(0)}, 0, 1)$.

Component completion and nonempty support. Thresholding a smooth attention peak can select only part of a coherent region. We therefore form 8-connected components of $\{u : \mathfrak{a}_i^{(1)}(u) \geq 0.5S_q\}$. A component is filled with its maximum value if it contains a seed at or above S_q and covers at most 35% of the image; the area constraint prevents diffuse low-contrast components from flooding the field. This is an intensity-based connected-component operation and uses no semantic detector or annotation. If max-pooling the result would leave every latent cell inactive, we sharpen and rescale the map as $\breve{\mathfrak{n}} \leftarrow \mathrm{clip}\big[1.02\,\tau_q\,(\breve{\mathfrak{n}}/p_{\max})^e, 0, 1\big]$ with $e = \mathrm{clip}[\ln 1.02/\ln(p_{\max}/p_2), 1, 40]$, where $p_{\max}$ and p_2 are the largest and second-largest max-pooled cell values, so that exactly the strongest latent cell crosses the threshold. Fig. 11 shows the complete procedure on three study images.

Excess field and subjective-study operating points. Writing the processed attention map as $\breve{\mathfrak{a}}_i$, the codec input is $\breve{\mathfrak{n}}_i = \eta_q^3 \breve{\mathfrak{a}}_i$. We pad this field symmetrically, consistently with the coded image, and max-pool it to the latent grid. The activation mask of Eq. 8 retains a latent cell when its pooled value is at least τ_q; the same pooled field drives the message gate. Although the message gate is implemented as a soft ramp, the shared activation mask makes it zero outside the retained cells and saturated inside them for these inference runs. Table 8 reports the two configurations used to generate the final subjective stimuli. The parameters were fixed before the study and applied to all 40 images without per-sample tuning or manual intervention.

B.2 Attention–information lift protocol

We use all 1,003 natural images and their measured human fixation maps from MIT1003 (Judd et al., 2009). Connected components in each binary fixation map are represented by their centroids. Representative information-bearing regions are detected independently of the compression model: InsightFace supplies face boxes and five facial landmarks, EasyOCR supplies text boxes at confidence at least 0.6, and YOLOv8-OIV7 supplies animal and general-object boxes. Each facial landmark is expanded to a disk whose radius is 6% of the corresponding face size, and object boxes covering more than half the image are excluded. The complement of the union of the face, text, landmark, and general-object masks defines background.

For a content class c, its fixation share is the fraction of all fixation points that fall inside its mask, and its area prevalence is the fraction of image pixels covered by that mask, pooled over the dataset. We report enrichment lift as the ratio of fixation share to area prevalence, equivalent to comparison with uniformly sampled image locations. The measured lifts are $9.16\times$ for faces, $17.63\times$ for facial landmarks, $7.47\times$ for text, and $2.82\times$ for animals; background is depleted to $0.59\times$. These detector-defined categories are representative rather than exhaustive, so the analysis supports enrichment of human attention on information-bearing content but does not measure coverage of every content type or claim that attention and information importance are equivalent.

C Subjective Study and Additional Perceptual Results

The stimulus set consists of 40 images: 24 from the conventional Kodak benchmark (Eastman Kodak Company, 1999) and 16 from the modern JPEG-AI test set (JPEG, 2023), the latter resized so that the longer side is 1024 px (bicubic, downscaling only). Each image is coded by three methods—ours, OneDC, and AEIC-ME—at two operating points, giving $3 \times 2 \times 40 = 240$ stimuli. The operating points are chosen so that the three methods sit at comparable rate: ≈ 0.016 bpp for the lower setting (0.0156–0.0173 bpp across the three methods) and ≈ 0.036 bpp for the higher one (0.0341–0.0376 bpp). Rates are totals over the 40-image set divided by its total pixel count, including the real entropy-coded payload and required color-correction statistics; ours lies between the two baselines at both points.

We assess perceived quality with a Double-Stimulus Impairment Scale (DSIS) test following ITU-R BT.500 (ITU-R, 2023) and ITU-T P.910 (ITU-T, 2023). On each trial the original image and one reconstruction are shown side by side (the original on a fixed side and explicitly labeled as the reference), and the participant rates the reconstruction on the standard five-grade impairment scale (5: imperceptible, …, 1: very annoying), presented in the participants' native language. 37 observers (researchers and graduate students in image and audio processing; no screening or exclusion was applied) each rated all 240 stimuli in a single self-paced session; this fully crossed design yields one rating per observer and stimulus. Trial order was randomized independently for every participant

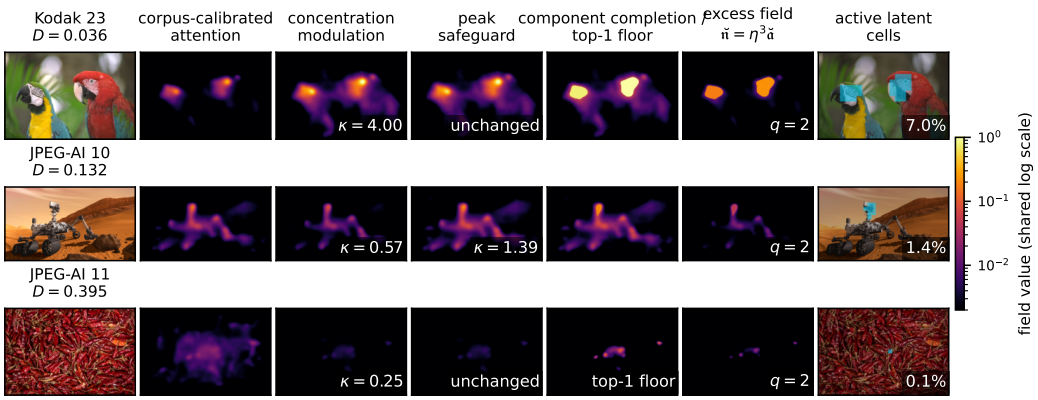


Figure 11: **Attention-map preprocessing used for the subjective-study stimuli.** The columns show corpus calibration, concentration modulation, the peak safeguard, component completion (including the top-1 floor when needed), the resulting excess fidelity-need field, and the active latent cells. The concentrated prediction in the first row is amplified; the multi-peak prediction in the second row is restored by the peak safeguard; and the diffuse prediction in the third row is reduced to one active cell. Cyan denotes active support. All rows use the high-rate study setting ($\lambda_{\text{rate}} = 4$, $q = 2$).

Table 8: Attention-derived field settings used in the subjective study. Active support is the test-set mean fraction of the image covered by selected latent cells.

Point	λ_{rate}	q	S_q	Active support
Low	16	0	0.1162	3.953%
High	4	2	0.1002	4.330%

from a per-participant seed. Before the scored trials participants were shown the five-grade scale with its verbal labels and the response controls. No practice trials were administered, by design: practice stimuli showing, e.g., a strongly distorted face or text would have cued observers to inspect such regions, biasing where they attend during the scored trials.

The test was delivered in the browser and run on each participant's own display. To exclude resampling artifacts each stimulus is rendered at native resolution, with browser zoom disabled and continuously monitored, a minimum display resolution of 2560×1600 *physical* pixels, and full-screen presentation; participants whose display did not satisfy these constraints were blocked from starting, and the pair is verified to fit on screen before each trial is timed. Because the design is within-subject and each impairment judgment is made against the reference displayed on the *same* screen, per-display differences in calibration, luminance, and ambient conditions act identically on the reference and the reconstruction and therefore contribute noise rather than bias to the between-method comparison.

Test interface. Fig. 12 shows the rating screen. The reference is displayed on the left and the reconstruction on the right at native resolution, the five-grade impairment scale is selected by button or number key and confirmed with the space bar, and a strip at the top lists the ratings given so far. The interface was presented to participants in their native language; the labels are shown here in English.

Additional rate-matched examples. Figs. 13–18 show six further study images on which the three codecs are matched in rate per image (FLARE within $\pm 20\%$ of both baselines, or cheaper) and the MOS gap to both baselines is large, complementing the two examples of Fig. 6b. Each page shows, at native resolution, the source and the three study stimuli (left column), each codec's spatial rate allocation (middle column: code length of the coded latents per cell, as bit density on a shared colour scale), and, for FLARE, the excess fidelity need field that drove the allocation (right). The baselines remain globally plausible but alter the face (Kodak 18, Kodak 04), smear the cap lettering

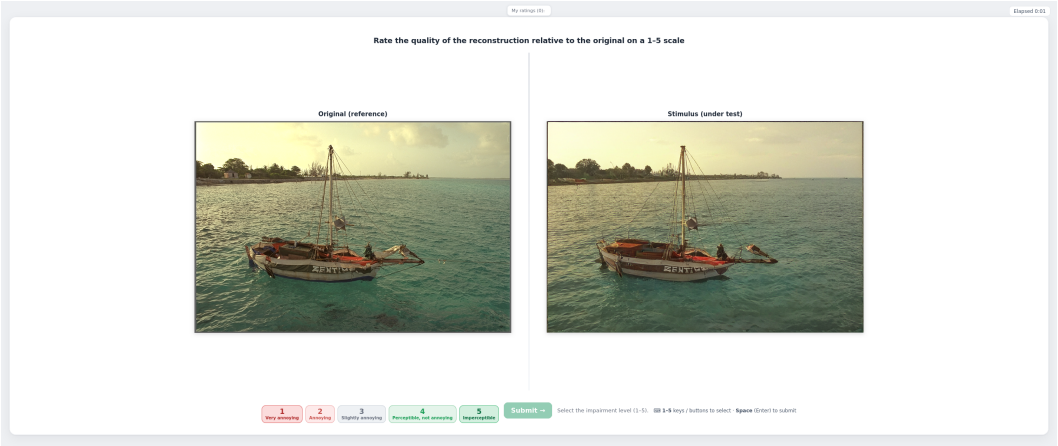


Figure 12: **DSIS rating screen used in the subjective study** Left: reference; right: reconstruction under test; bottom: the five-grade impairment scale and submit control.

(Kodak 03), dissolve the lighthouse railing (Kodak 19), alter the eye and ear of the profile portrait (JPEG-AI 16), and deform the dog's eyes and muzzle (JPEG-AI 13, the image of Fig. 6b shown here in full), whereas FLARE preserves these structures at the same rate.

D Attribution of the Subjective Gain to Excess-Fidelity Allocation

The study of Sec. 6.3 compares FLARE driven by the attention-derived excess field against two realism codecs that receive no field. Its MOS gain can therefore be read in two ways: (H1) observers reward the fidelity that the field places in attended regions, or (H2) FLARE's realism floor is itself preferred over the baselines, independently of where excess is spent. Because a subjective test of every ablation condition is not feasible within the fully crossed design (each added condition adds 80 trials per observer), we separate the two hypotheses using the study data and rate-matched within-codec controls. All reconstructions below are produced by the bit-exact stimulus pipeline (same checkpoints, same activation threshold, same real rANS coder); the regenerated study stimuli match the originals pixel for pixel.

Excess-rate share. For each study stimulus we re-encode the image on the same checkpoint with the field set to zero and define the excess-rate share as $(R_{\text{field}} - R_{\text{zero}})/R_{\text{field}}$, the fraction of the stimulus's bits attributable to the excess field. It ranges from 0.02 to 0.70 (mean 0.37) at the low-rate point and from 0.01 to 0.50 (mean 0.23) at the high-rate point, so the study already spans images on which FLARE behaves almost like its floor and images on which excess dominates.

The floor is not preferred objectively. Under H2 the zero-excess floor should already beat the baselines. At matched rate it does not (Table 9): the floor's global LPIPS and DISTS are within noise of AEIC-ME and OneDC, and its LPIPS inside the attended region is likewise on par. The study stimuli differ from the floor almost exclusively in the attended region, where LPIPS drops from 0.221 to 0.155 (low rate) and from 0.156 to 0.117 (high rate). Their global LPIPS is slightly higher than the rate-matched floor's only because the floor spends the same total budget uniformly; the field itself does not degrade the background. Switching the field on in the *same* checkpoint (the "same λ" rows) adds whole-image rate ($0.010 \to 0.016$ and $0.028 \to 0.037$ bpp) while global PSNR, LPIPS, and DISTS all improve slightly and background LPIPS is unchanged ($0.251 \to 0.245$ and $0.158 \to 0.159$). The same picture holds at scale on CLIC 2020 test (428 images) and DIV2K validation (100 images), neither of which overlaps the training corpus: at matched rate the floor's global PSNR, LPIPS, and DISTS track AEIC-ME closely (OneDC has lower global LPIPS but higher DISTS throughout), whereas the study field raises attended-region PSNR by 4.1–4.4 dB over the floor and lowers attended LPIPS from 0.217 to 0.127 and from 0.138 to 0.092 on CLIC, and from 0.258 to 0.149 and from 0.158 to 0.110 on DIV2K; relative to the same checkpoint with the field off,

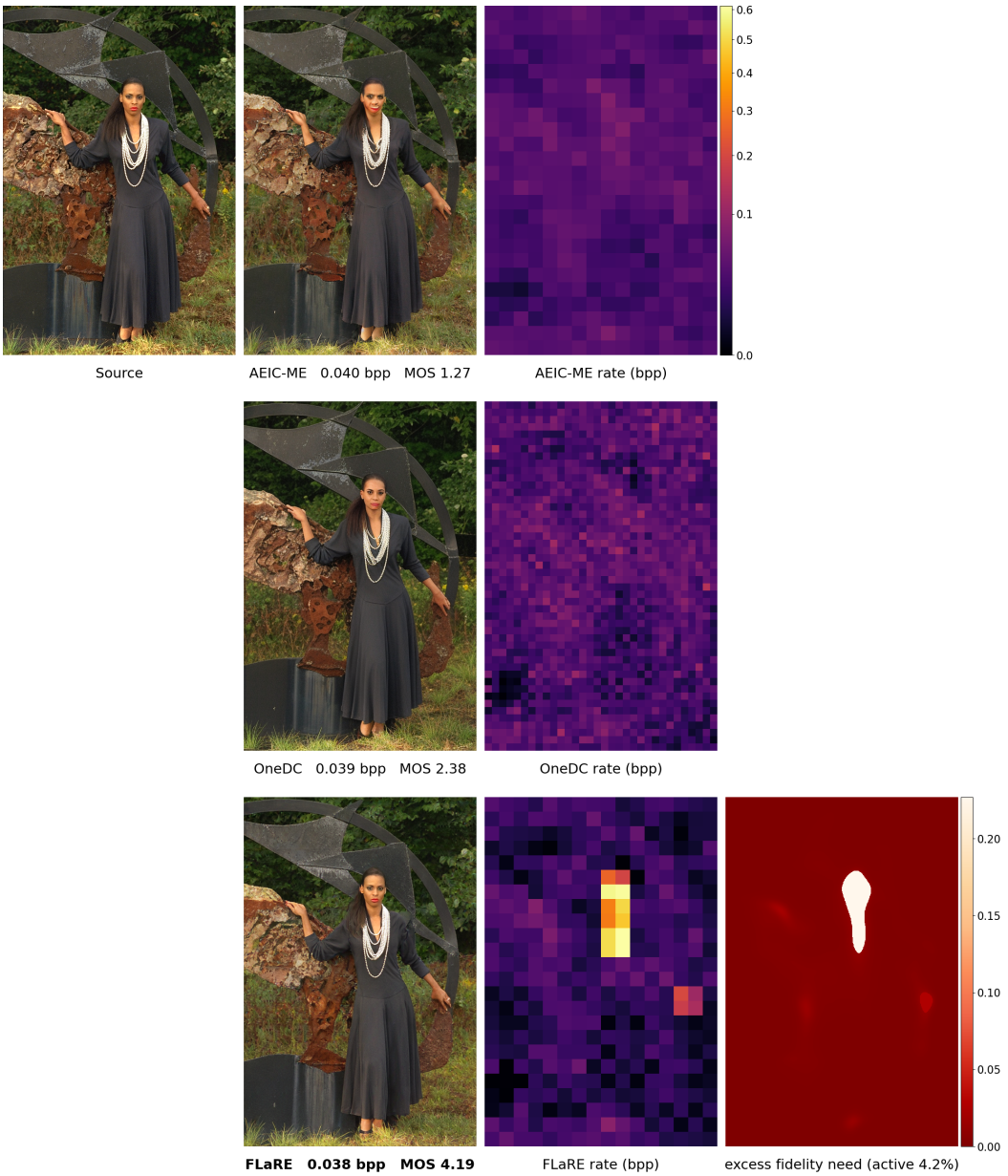


Figure 13: **Kodak 18, high-rate point** (native resolution). Left column: source and the three study stimuli, AEIC-ME (0.040 bpp, MOS 1.27), OneDC (0.039 bpp, MOS 2.38), FLARE (0.038 bpp, MOS 4.19). Middle column: each codec's spatial rate allocation (code length of the coded latents per cell, as bit density on a shared square-root colour scale). Right: the excess fidelity need field $\breve{n}$ that drove FLARE (active region 4.2% of the area).

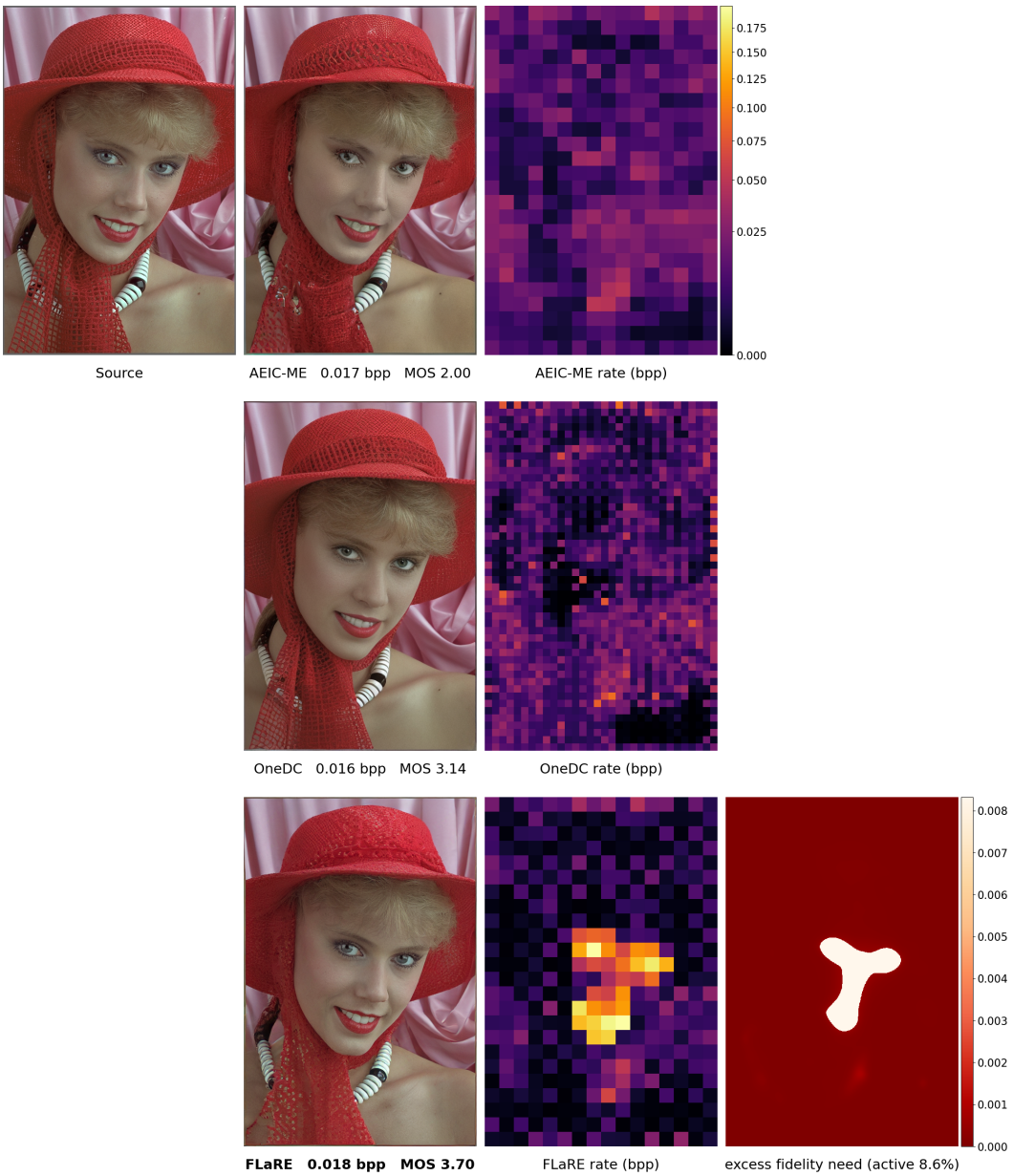


Figure 14: **Kodak 04, low-rate point** (native resolution). Left column: source and the three study stimuli, AEIC-ME (0.017 bpp, MOS 2.00), OneDC (0.016 bpp, MOS 3.14), FLARE (0.018 bpp, MOS 3.70). Middle column: each codec's spatial rate allocation (code length of the coded latents per cell, as bit density on a shared square-root colour scale). Right: the excess fidelity need field $\breve{n}$ that drove FLARE (active region 8.6% of the area).

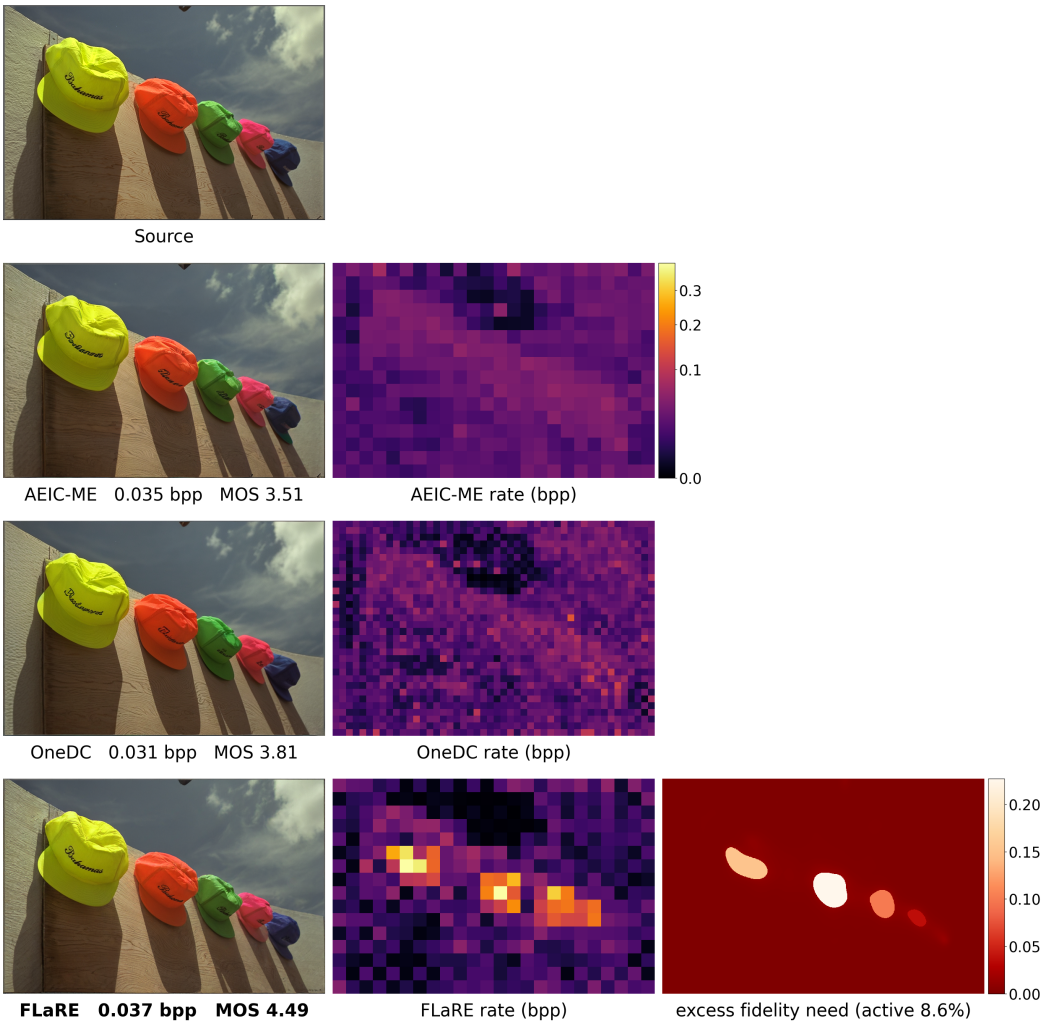


Figure 15: **Kodak 03, high-rate point** (native resolution). Left column: source and the three study stimuli, AEIC-ME (0.035 bpp, MOS 3.51), OneDC (0.031 bpp, MOS 3.81), FLARE (0.037 bpp, MOS 4.49). Middle column: each codec's spatial rate allocation (code length of the coded latents per cell, as bit density on a shared square-root colour scale). Right: the excess fidelity need field $\breve{n}$ that drove FLARE (active region 8.6% of the area).

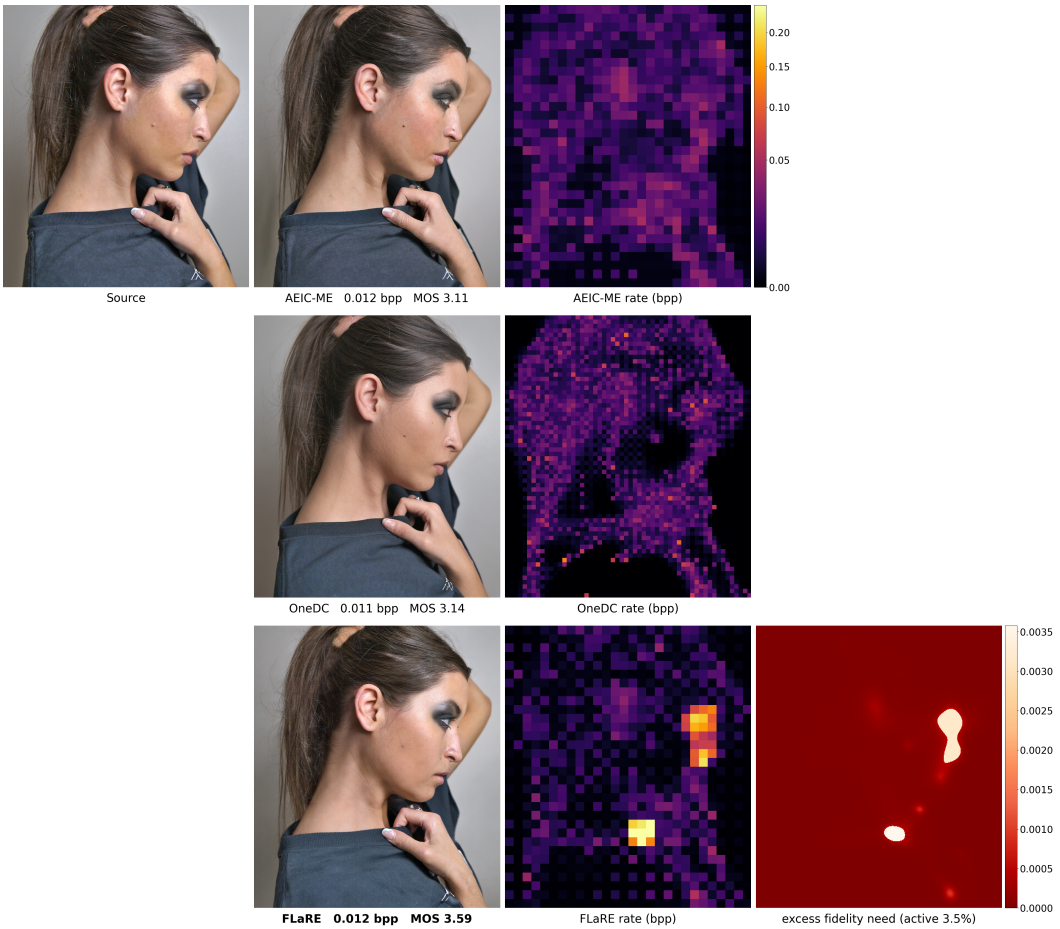


Figure 16: **JPEG-AI 16, low-rate point** (native resolution). Left column: source and the three study stimuli, AEIC-ME (0.012 bpp, MOS 3.11), OneDC (0.011 bpp, MOS 3.14), FLARE (0.012 bpp, MOS 3.59). Middle column: each codec's spatial rate allocation (code length of the coded latents per cell, as bit density on a shared square-root colour scale). Right: the excess fidelity need field $\breve{n}$ that drove FLARE (active region 3.5% of the area).

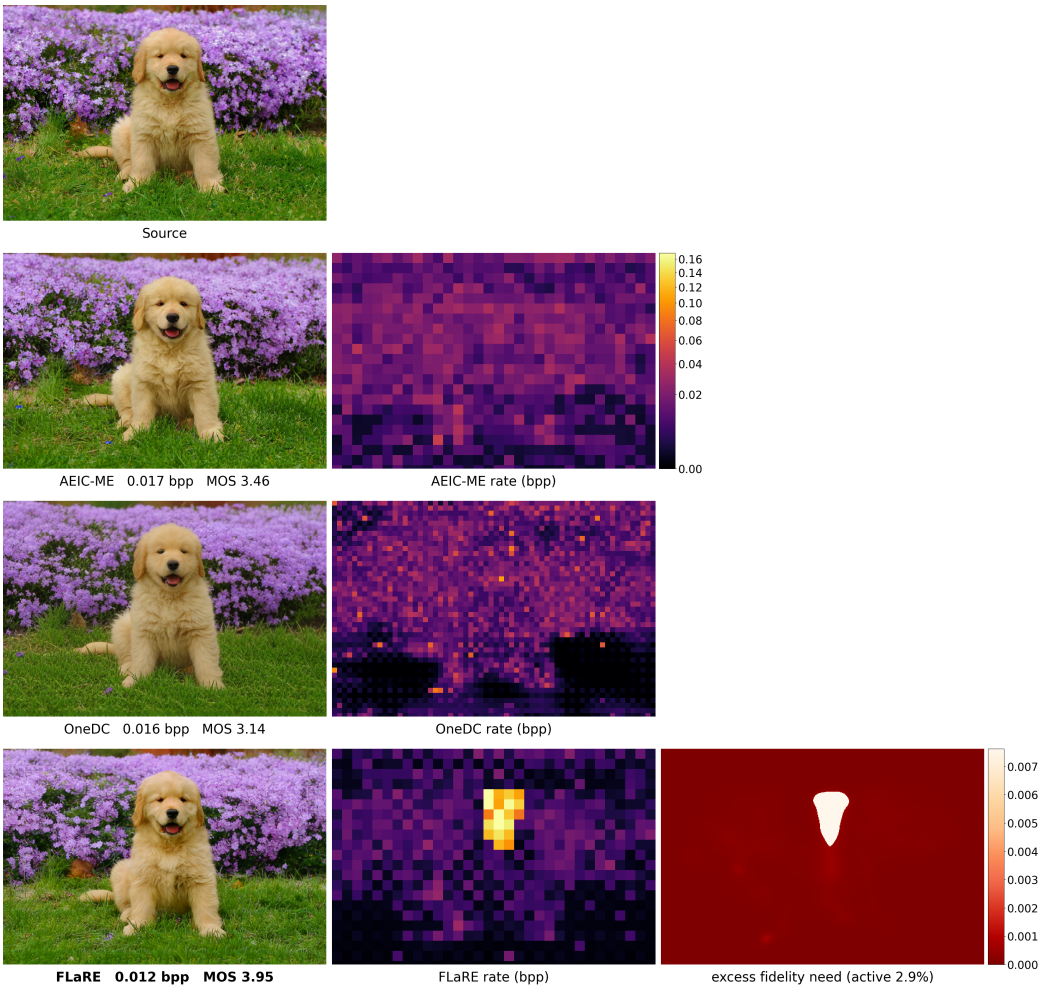


Figure 17: **JPEG-AI 13, low-rate point** (native resolution). Left column: source and the three study stimuli, AEIC-ME (0.017 bpp, MOS 3.46), OneDC (0.016 bpp, MOS 3.14), FLARE (0.012 bpp, MOS 3.95). Middle column: each codec's spatial rate allocation (code length of the coded latents per cell, as bit density on a shared square-root colour scale). Right: the excess fidelity need field $\breve{n}$ that drove FLARE (active region 2.9% of the area).

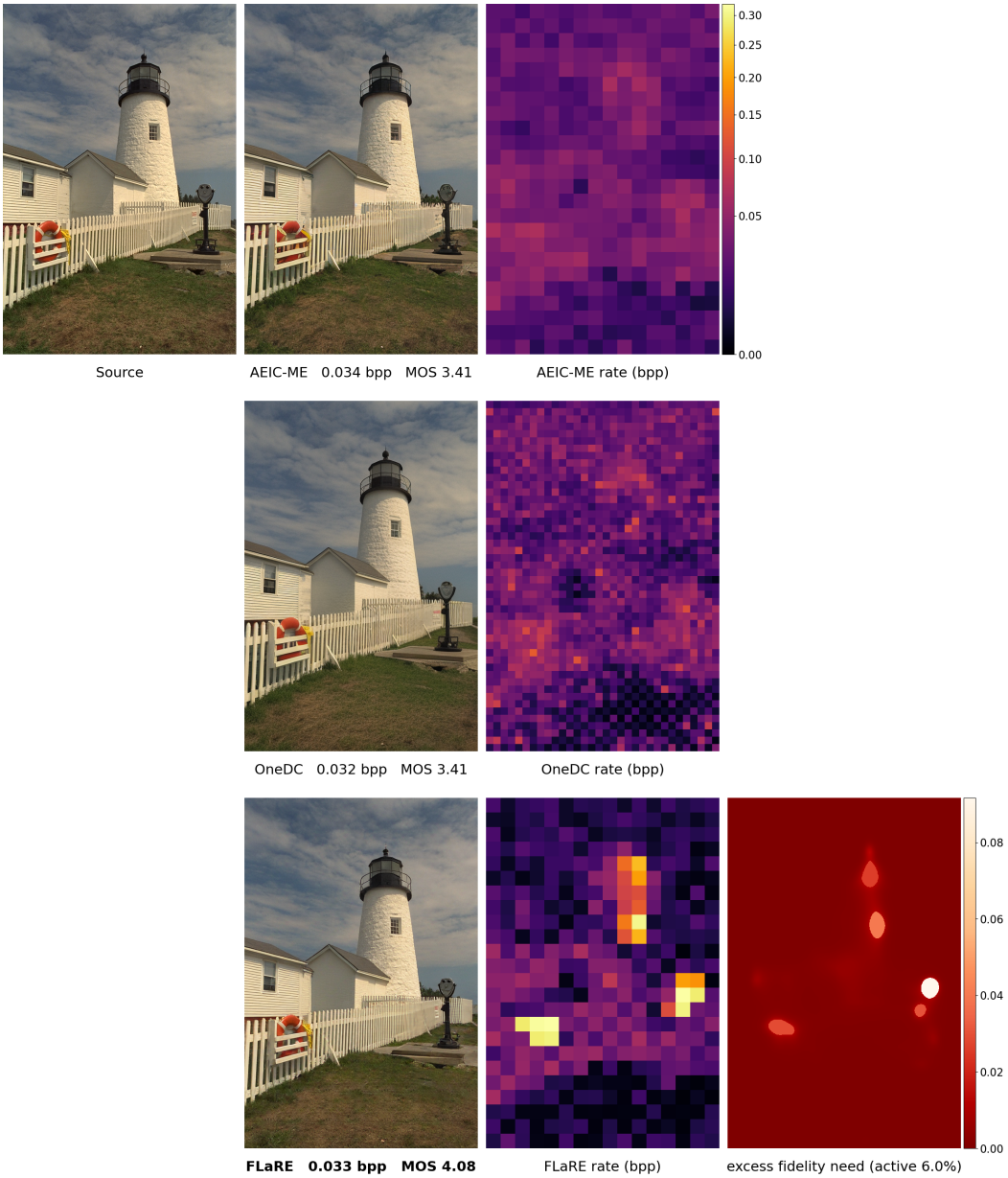


Figure 18: **Kodak 19, high-rate point** (native resolution). Left column: source and the three study stimuli, AEIC-ME (0.034 bpp, MOS 3.41), OneDC (0.032 bpp, MOS 3.41), FLARE (0.033 bpp, MOS 4.08). Middle column: each codec's spatial rate allocation (code length of the coded latents per cell, as bit density on a shared square-root colour scale). Right: the excess fidelity need field $\breve{n}$ that drove FLARE (active region 6.0% of the area).

global and background LPIPS again do not degrade on either set. On these 2K images FLARE and AEIC-ME are decoded with the same tiled UNet/VAE protocol of the AEIC-ME evaluation code.

Table 9: Objective quality at matched rate on the 40-image study set, CLIC 2020 test (428 images), and DIV2K validation (100 images); the latter two contain no training images. bpp is always the whole-image rate (payload bits over all pixels), including in rows whose quality columns are restricted to a region. The "zero field, same bpp" row is the zero-excess floor of FLARE log-rate interpolated between adjacent operating points to the study field's rate. On the study set the baselines are the study operating points; on CLIC and DIV2K they are likewise interpolated between their adjacent operating points (AEIC-ME ft16/8/4/2, OneDC $\lambda \in \{12.2, 7.4, 4.6, 2.9\}$); on these 2K images FLARE and AEIC-ME use the same tiled decoding. "Att." restricts the metric to the active cells of the study field and "Bg." to their complement: PSNR from the squared error over those pixels, LPIPS (AlexNet) as the mean of the full-image spatial LPIPS map over them, with no masking of the inputs. *The study checkpoint with the field switched off, at its own (lower) rate; this row is not rate-matched and isolates what the field adds to the identical model.

Codec	bpp	PSNR	LPIPS	DISTS	Att. PSNR	Att. LPIPS	Bg. PSNR	Bg. LPIPS
Study set, low-rate point (≈ 0.016 bpp)								
AEIC-ME	0.017	21.44	0.188	0.098	18.50	0.213	21.73	0.187
OneDC	0.016	21.01	0.173	0.110	18.00	0.205	21.32	0.172
FLARE, zero field, same λ^*	0.010	20.59	0.252	0.122	17.79	0.280	20.84	0.251
FLARE, zero field, same bpp	0.016	21.62	0.204	0.102	18.88	0.221	21.86	0.204
FLARE, study field	0.016	20.90	0.240	0.111	**22.18**	**0.155**	20.87	0.245
Study set, high-rate point (≈ 0.037 bpp)								
AEIC-ME	0.038	22.69	0.138	0.080	19.62	0.159	23.04	0.137
OneDC	0.034	22.44	0.126	0.087	19.50	0.145	22.79	0.125
FLARE, zero field, same λ^*	0.028	22.82	0.159	0.085	20.03	0.172	23.10	0.158
FLARE, zero field, same bpp	0.037	23.31	0.143	0.078	20.51	0.156	23.60	0.143
FLARE, study field	0.037	23.08	0.156	0.080	**24.56**	**0.117**	23.03	0.159
CLIC 2020, low-rate point (≈ 0.010 bpp)								
AEIC-ME	0.010	24.08	0.187	0.082	20.41	0.222	24.32	0.187
OneDC	0.010	23.67	0.151	0.085	19.82	0.201	23.96	0.150
FLARE, zero field, same λ^*	0.007	23.76	0.209	0.088	20.12	0.253	24.00	0.208
FLARE, zero field, same bpp	0.010	24.58	0.183	0.078	21.00	0.217	24.82	0.182
FLARE, study field	0.010	24.02	0.203	0.082	**25.26**	**0.127**	24.02	0.205
CLIC 2020, high-rate point (≈ 0.025 bpp)								
AEIC-ME	0.025	25.90	0.127	0.060	22.36	0.151	26.19	0.126
OneDC	0.025	25.85	0.107	0.063	22.21	0.135	26.17	0.106
FLARE, zero field, same λ^*	0.021	26.39	0.132	0.058	23.05	0.149	26.65	0.132
FLARE, zero field, same bpp	0.025	26.79	0.122	0.054	23.46	0.138	27.06	0.122
FLARE, study field	0.025	26.61	0.130	0.054	**27.85**	**0.092**	26.61	0.132
DIV2K, low-rate point (≈ 0.011 bpp)								
AEIC-ME	0.011	21.31	0.228	0.090	19.19	0.259	21.41	0.227
OneDC	0.011	20.98	0.189	0.100	18.80	0.233	21.09	0.188
FLARE, zero field, same λ^*	0.009	21.23	0.245	0.095	19.08	0.289	21.32	0.244
FLARE, zero field, same bpp	0.011	21.78	0.223	0.087	19.71	0.258	21.88	0.223
FLARE, study field	0.011	21.35	0.241	0.088	**23.80**	**0.149**	21.32	0.243
DIV2K, high-rate point (≈ 0.031 bpp)								
AEIC-ME	0.031	23.45	0.146	0.065	21.27	0.166	23.56	0.145
OneDC	0.031	23.31	0.128	0.071	21.17	0.150	23.43	0.127
FLARE, zero field, same λ^*	0.027	23.79	0.155	0.063	21.70	0.167	23.90	0.155
FLARE, zero field, same bpp	0.031	24.13	0.146	0.060	22.04	0.158	24.24	0.145
FLARE, study field	0.031	23.90	0.153	0.060	**26.32**	**0.110**	23.87	0.154

Dose response of the MOS gain. Under H1 the MOS gain should vanish where little excess is spent and grow with the excess share; under H2 it should be roughly constant. We split the images into terciles of excess share within each rate point and, for every tercile, form each codec's two-point rate–MOS curve (aggregate bpp over the tercile's images, mean MOS) and read the baseline's MOS at FLARE's rate by log-rate interpolation. Fig. 7 (main text) and Table 10 show the result: in the low-excess tercile the three codecs are within -0.23 to $+0.11$ MOS of each other, FLARE sitting slightly below AEIC-ME at the low-rate point, in the middle tercile FLARE leads by $+0.3$–$+0.5$, and in the high-excess tercile by $+0.5$–$+1.3$. Rate-saving figures are not reported for the upper terciles because they would require extrapolating the baselines' two-point curves.

Per-image regression. Excess share is correlated with the per-image rate difference between FLARE and the baseline (Spearman 0.67), so we regress the per-image MOS difference (FLARE

Table 10: **Matched-rate MOS gain of FLARE over each baseline by excess-rate share.** Images are split into terciles (within each rate point) of the share of FLARE's bits spent on the excess field. Gains are MOS differences at FLARE's aggregate rate, with the baseline's MOS interpolated between its two study points on a log-rate axis. The "All" row instead reports the unadjusted MOS differences between the study operating points.

			Low rate (≈ 0.016 bpp)		High rate (≈ 0.037 bpp)	
Excess-share tercile	Images	Mean share	vs. AEIC-ME	vs. OneDC	vs. AEIC-ME	vs. OneDC
Low	14	0.11	-0.23	-0.05	$+0.01$	$+0.11$
Middle	13	0.33	$+0.50$	$+0.40$	$+0.50$	$+0.28$
High	13	0.48	$+0.73$	$+0.51$	$+1.27$	$+0.91$
All	40	0.30	$+0.29$	$+0.33$	$+0.57$	$+0.50$

minus baseline; 160 pairs) on excess share, the log-rate difference, a baseline indicator, and a rate-point indicator; because each image contributes four pairs, standard errors are clustered by image (40 clusters). Excess share receives a coefficient of $+2.24$ MOS per unit share (SE 0.46, $p < 10^{-4}$), the log-rate difference $+0.49$ (SE 0.28, $p = 0.09$), and the baseline indicator is not significant. The MOS gain is thus explained by how much excess was spent, not by the rate difference alone.

Field-placement controls. Finally, holding the checkpoint, threshold, and field magnitude (η_q^3) fixed, so that only the placement of the field changes, we replace the study field by (i) a zero field, (ii) random rectangles and (iii) a centered block on the latent grid, both with the same number of active cells as the study field ($\approx 4\%$ of the image), and (iv) a uniform field. Table 11 reports the whole-image rate and the LPIPS inside and outside the attended region on the same weights. Random and centered placement spend about the same excess rate as the study field yet leave the attended region at its field-off quality (the centered block overlaps it only partially), and the background stays within 0.01 LPIPS of the zero field in every case: message bits improve only the cells they land in. Raising the floor instead to the study rate on other weights gives an attended-region LPIPS of 0.221 (low) and 0.156 (high), still far from the study field. A uniform field cannot be rate-matched on the same weights: because the activation threshold opens every cell, its lowest achievable rate is 0.136 bpp (low) and 0.169 bpp (high), four to ten times the study rates.

Table 11: **Field-placement controls on the same weights.** Study checkpoints and thresholds fixed; only the field changes. bpp is the whole-image rate over the 40-image set; "Att." and "Bg." restrict LPIPS to the study field's active region and to its complement. Random and centered fields have the same number of active cells and the same magnitude as the study field.

	Low-rate point ($\lambda_{\text{rate}} = 16$)			High-rate point ($\lambda_{\text{rate}} = 4$)		
Field	bpp	Att. LPIPS↓	Bg. LPIPS↓	bpp	Att. LPIPS↓	Bg. LPIPS↓
Zero (field off)	0.010	0.280	0.251	0.028	0.174	0.159
Random rectangles	0.014	0.277	0.243	0.039	0.178	0.157
Centered block	0.015	0.261	0.246	0.040	0.168	0.158
Study field	0.016	**0.155**	0.245	0.037	**0.117**	0.159

Caveats. Terciles contain 13–14 images, so per-tercile MOS means carry roughly ± 0.07 uncertainty against differences of 0.5–1.3. The analysis is observational within the study rather than a new subjective test; images on which much excess is spent are those with faces and text, which is the mechanism under study rather than a confound, and the regression controls for the accompanying rate difference. Together with the objective parity of the floor, the results support H1: the subjective gain follows the selective allocation of fidelity, not the baseline performance of the codec.

Study images and their excess fields. Fig. 19 shows all 40 study images with the excess fidelity-need field used at each rate point, grouped by tercile of the mean excess share over the two points (the terciles of Table 10 are formed within each rate point, so a few images change group between the two). Images in the high-excess tercile contain faces, text, or compact objects on which the field concentrates; images in the low-excess tercile have diffuse attention, and the field retains at most a few latent cells, so FLARE codes them almost as its floor does.

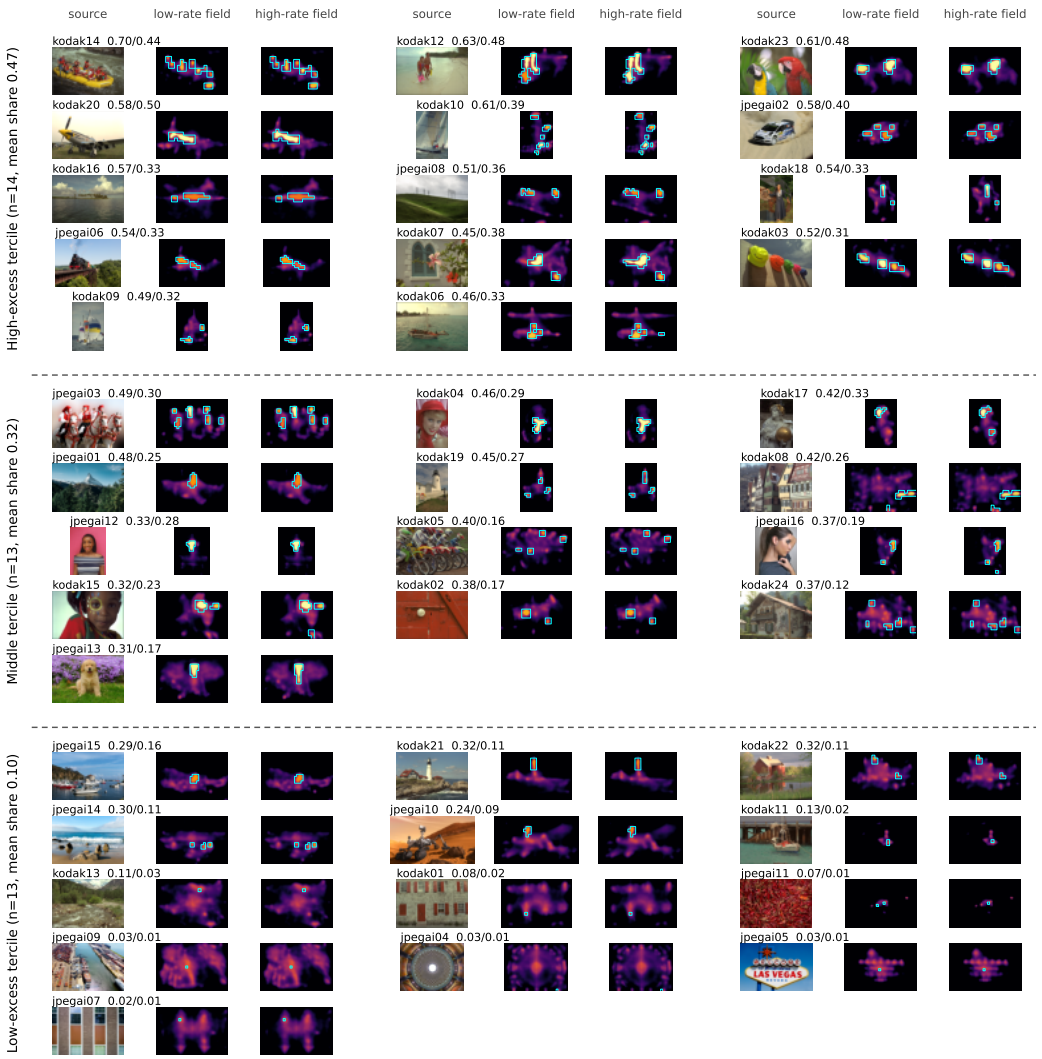


Figure 19: **Study images and their excess fidelity-need fields, grouped by excess-rate share.** For each image: the source, the excess field at the low-rate point, and at the high-rate point (both normalized by η^3, log color scale; cyan outlines mark the active latent cells). The label gives the excess-rate share at the low / high rate point; dashed rules separate the terciles of mean share.

E REPRODUCIBILITY AND ABLATIONS

Rate accounting. For the study and its field controls, we count the actual entropy-coded payload, excluding file-container headers, plus any required color-correction statistics. The AdaIN color correction used by FLARE and AEIC-ME when $WH > 768^2$ requires the source RGB mean and standard deviation, each quantized to 16 bits: 96 bits (12 bytes) per corrected image. These statistics are included in rates used for comparisons and rate-matched interpolation, but not in tables explicitly decomposing latent payload. None of the Kodak images requires this correction. Subjective rates are total bits over the 40-image set divided by its total pixel count; the objective RD curves average per-image bpp. Container conventions of the existing baseline runs are retained; Fig. 25 specifies its full-container accounting and the remaining entropy-estimated exception (GLC on Kodak). The task evaluation retains likelihood-estimated rates for FLARE and measured rates for AEIC-ME; the latter include color statistics where applied.

Baseline configurations. AEIC-ME (Zhang et al., 2026) and OneDC (Xue et al., 2025) are run from their public checkpoints; because they are the closest realism codecs to our base model, their operating points are selected to bracket FLARE's rate at each matched-rate point. MS-ILLM (Muckley et al., 2023), GLC (Jia et al., 2024), and the VTM (Bross et al., 2021) intra anchor are evaluated from their released implementations. All baselines are scored on the identical image set with the same real-bitstream rate accounting and the same objective-metric implementations used for FLARE.

Model-availability audit. We considered direct comparisons with MRIDC and Relic et al., the most closely related fidelity-allocation methods, and verified their availability in September 2026. MRIDC's official repository contains a README and result images only; its task list marks inference code, training code, and pretrained models as unreleased, and its last commit dates from March 2025. Relic et al. report that they could compare against MRIDC only by parameter count, and their own paper provides neither an implementation nor reconstructions. A faithful reimplementation is not a matter of porting a loss. MRIDC couples a pretrained VQ encoder, a second lightweight VQ encoder, a bidirectional masked-latent predictor fine-tuned from a pretrained model, and a ControlNet-guided Stable Diffusion v2.1 (Rombach et al., 2022) decoder run for 20 denoising steps, trained jointly on VQ, prediction, and guidance losses with the remaining parts frozen. Relic et al. fine-tune Stable Diffusion v2.1 into a spatially varying diffusion-dequantization model (200k steps at batch 32), train a timestep-conditioned entropy model for a further 375k steps, losslessly transmit the timestep map, and decode with $\max(\mathbf{t})$ DDIM iterations (5 to 50 in their examples). Each of these components is the method's defining mechanism, and none has a counterpart in a one-step codec: reproducing them means training the full pipelines from scratch with no reference reconstructions to validate against, and grafting only the spatial-control element onto our base codec would benchmark a construction of our own rather than the published method. We therefore compare against methods for which faithful, matched-rate evaluation is possible, and place MRIDC and Relic et al. on the complexity axis using their reported step counts and parameter counts, together with the MRIDC authors' own timing and, for Relic et al., a lower bound on decode time obtained by running the Stable Diffusion v2.1 UNet they fine-tune for the reported number of steps on our hardware (Table 1).

E.1 RATE COST OF THE IMPLICIT FIELD TRANSPORT

The decoder receives no field; the scale and mask enter the hyper-analysis input of Eq. 11, and the decoder recovers the scale from the hyperlatent $\hat{\mathbf{z}}$. This subsection measures what that implicit transport costs. We code each of the 40 study images twice with the study checkpoints and activation threshold of each operating point (Table 8) and the real rANS pipeline that produced the stimuli: once with the study field and once with a zero field, and repeat the measurement on CLIC 2020 test (428 images) and DIV2K validation (100 images) with the DeepGaze fields of Table 9. The coded hyperlatent string and the four quadtree strings are recorded separately. Because the base and message channels share the quadtree strings, their split is taken from the codec's own likelihoods, which reproduce the real string lengths to within 2.5%. Table 12 reports per-image means.

Hyperlatent. On the study set the coded hyperlatent is 3.3 B (low rate) and 1.7 B (high rate) per image *smaller* with the study field than with a zero field. The per-image difference ranges from -20 to $+8$ B and is positive for only 5 (low) and 7 (high) of the 40 images, which is within the 4-byte granularity of the range coder. The 2K images behave the same way: the hyperlatent is 14.8 and 9.8 B smaller on CLIC and 12.1 and 6.8 B smaller on DIV2K, the difference never exceeds $+8$ B

Table 12: **Coded bytes per image with the study field versus a zero field** (per-image means over the 40-image study set, CLIC 2020 test, and DIV2K validation; real rANS bytes for the total payload and the hyperlatent, likelihood estimates for the base and message groups). "Excess" is the total-payload difference. The last two columns give what sending the field explicitly would cost instead: the entropy of the binary activation mask at latent resolution (2–4% active cells) and its `zlib` size.

Point	Field	Total	Hyperlatent	Base	Message	Excess	Mask entropy	Mask zlib
Study (40)								
Low (λ16)	zero	643.7	204.2	412.8	1.6	—	—	—
	study	1051.5	200.9	374.3	455.2	407.8	13.3	24.3
	Δ	+407.8	−3.3	−38.4	+453.6			
High (λ4)	zero	1855.9	303.2	1521.0	0.3	—	—	—
	study	2386.3	301.5	1361.2	691.1	530.4	14.1	24.7
	Δ	+530.4	−1.7	−159.9	+690.7			
CLIC2020 test (428)								
Low (λ16)	zero	2616.9	989.4	1588.1	8.2	—	—	—
	study	3553.0	974.5	1489.4	1062.0	936.1	47.6	37.8
	Δ	+936.1	−14.8	−98.8	+1053.7			
High (λ4)	zero	7240.9	1485.3	5693.5	4.0	—	—	—
	study	8690.0	1475.6	5349.7	1803.2	1449.1	51.2	39.6
	Δ	+1449.1	−9.8	−343.9	+1799.3			
DIV2K val (100)								
Low (λ16)	zero	3171.3	1025.7	2101.9	8.9	—	—	—
	study	4035.1	1013.6	2019.6	969.9	863.8	40.8	33.2
	Δ	+863.8	−12.1	−82.3	+961.0			
High (λ4)	zero	9533.2	1528.2	7932.0	3.5	—	—	—
	study	11083.0	1521.4	7626.6	1862.7	1549.8	44.3	35.1
	Δ	+1549.8	−6.8	−305.4	+1859.2			

on any image, and it is positive for 22–36 of the 428 CLIC images and 9–14 of the 100 DIV2K images. The hyperlatent is 19% (low) and 13% (high) of the total payload with the study field on the study set and 14–27% on the 2K sets. The implicit transport of the scale and mask therefore has no measurable rate cost on any of the three sets.

Where the excess goes. The excess payload (408 and 530 B per image) is carried by the message group (+454 and +691 B), while the base group becomes smaller with the field on (−38 and −160 B, i.e., 9% and 30% of the excess; 10–24% on CLIC and DIV2K). At the level of the coded representation the field therefore does not sit on top of a byte-identical floor code: part of the base rate in the active cells is re-coded through the shared hyperprior and quadtree context once the message carries the content. This coupling stays inside the coded representation; the additivity of Sec. 3.1 is stated at the level of the objective, and the delivered quality outside the field is unchanged when the field is switched on (Table 9, "same λ" rows). We report the displacement as an observation and do not attribute it to a specific mechanism.

Comparison with explicit transport. Losslessly sending even the binary activation mask at latent resolution would cost at least 13–14 B per image on the study set and 41–51 B on the 2K sets (its entropy bound; 24–40 B with `zlib`), and the 8-bit field values on the active cells a further 15–18 B, i.e., 3–5% of the excess and 0.4–1.3% of the total payload, whereas the measured cost of the implicit transport is zero or negative.

E.2 Synthetic-field mechanism-validation protocol

To decouple the allocation mechanism from the DeepGaze surrogate, we drive the trained FLARE model with synthetic excess fidelity-need fields $\breve{\mathfrak{n}} = \eta^3 \breve{\mathfrak{a}}$, whose target region and strength are known exactly, with $\breve{\mathfrak{a}}$ denoting a hand-placed indicator. We measure per-latent coded rate from the entropy model on the 40-image set at a single fixed checkpoint. This protocol isolates where and how much rate the model assigns from any property of the attention predictor.

At scale. Fig. 20 repeats the three tests of Fig. 5 on CLIC 2020 test (428 images) and DIV2K validation (100 images), neither of which overlaps the training corpus, with the same checkpoint and a square block of side $0.25 \min(H, W)$ (4.9% and 4.8% of the image). At $\eta = 1$ the block receives 96.9% (CLIC) and 96.3% (DIV2K) of the message bits, a concentration of 19.9× and 20.2× its area share (median; the ceiling is $1/\text{area} \approx 20.5\times$), with at least 90% of the bits inside the block on 95% and 96% of the images and no dependence on the block's position (19.5–21.0×

across the five placements). In-block message bits are strictly non-decreasing in η on every image of both sets. From $\eta = 0$ to 1, in-block PSNR rises by 9.3 dB (CLIC) and 9.8 dB (DIV2K) at the median, while background PSNR changes by $+0.03$ dB on both sets.

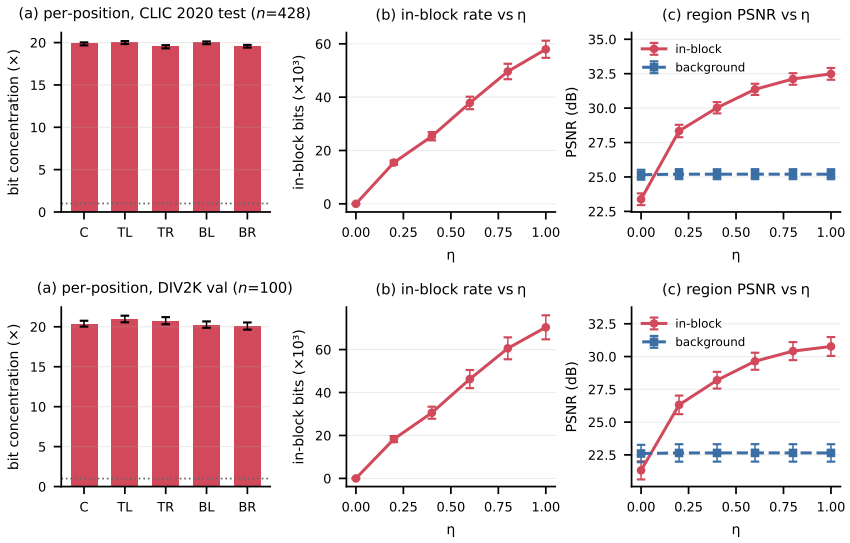


Figure 20: **Mechanism tests of Fig. 5 on CLIC 2020 test (top) and DIV2K validation (bottom).** (a) Message-bit concentration inside a block placed at five positions ($\eta = 1$; dotted line: uniform allocation). (b) In-block message bits versus η. (c) PSNR inside the block and in the background versus η. Means with 95% confidence intervals over images.

E.3 Auxiliary-decoder versus direct-output supervision

In FLARE, the pointwise component $\mathfrak{d}$ is squared error. We assign the AEIC-ME backbone's scalar full-reference regularizers—LPIPS and CLIP in pretraining, and DISTS and CLIP in adversarial finetuning—to Δ_G; the adversarial loss realizes the separate distributional perception term. This is an implementation choice rather than a unique decomposition. Other distortion measures may admit alternative spatial decompositions or motivate different pointwise components; we leave their systematic design to future work.

For Fig. 3, we fine-tune variants with auxiliary-decoder supervision and direct-output supervision from the same initialization and evaluate them on Kodak24 under the uniform excess fidelity-need field $\breve{\mathfrak{n}}(u) = \eta^3$. Thus, every location requests the same additional fidelity, and η parameterizes the supplied field rather than the codec. At $\eta = 1$, direct output supervision obtains 33.06 dB rather than 31.32 dB, but is worse in LPIPS (0.056 vs. 0.040), DISTS (0.054 vs. 0.038), and patchFID (26.3 vs. 20.5), despite using slightly more rate (0.861 vs. 0.821 bpp). Across the sweep, the auxiliary-decoder variant improves all four metrics monotonically, whereas DISTS and patchFID under direct output supervision reach their best values near $\eta = 0.5$ and then worsen.

F Application-Specific Evaluation

We evaluate whether a *task-specific* excess fidelity-need field lets FLARE preserve the source content a downstream recognizer must read, under a matched-rate comparison against general codecs. The purpose is to show that the field interface accommodates such extended scenarios without retraining, not to establish end-to-end codec superiority in them. The experiments are scoped to small target regions, where a codec that prices distortion uniformly must raise the rate of the whole image to give the region enough bits, so the rate saved by allocating fidelity to the region alone is largest; the competitors are not given a region input (see *Competitors* below).

Tasks, data, and ground truth. Three forced-recognition tasks, 100 images each: (i) **license plate** on CCPD (Xu et al., 2018), ground-truth plate string and box decoded from the CCPD filename; (ii) **face identity** on COCO images containing detected faces (no face-size filter, so many small

faces); (iii) **logo brand** on LogoDet-3K (Wang et al., 2022) with box and brand annotations. Because the benefit depends on region size (Sec. 6.4), for plate and logo we additionally report the 100 *smallest*-ROI images (plate longest side 102–139 px; logo primary-mark ≈20 px).

Excess fidelity-need field. The field is a binary region-of-interest indicator (union of task boxes $= 1$, background $= 0$) scaled by the dial, $\breve{\mathfrak{n}} = \eta^3 \cdot \mathbf{1}[\text{ROI}]$, with $\eta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$; η=0 is the realism floor (no excess). We run three operating points $\lambda_{\text{rate}} \in \{16, 8, 4\}$ giving an $\eta \times \lambda$ grid. The inference regime matches the rest of the paper (symmetric field padding, activation mask, and message gate as in Appendix A.3); rate accounting follows Appendix E.

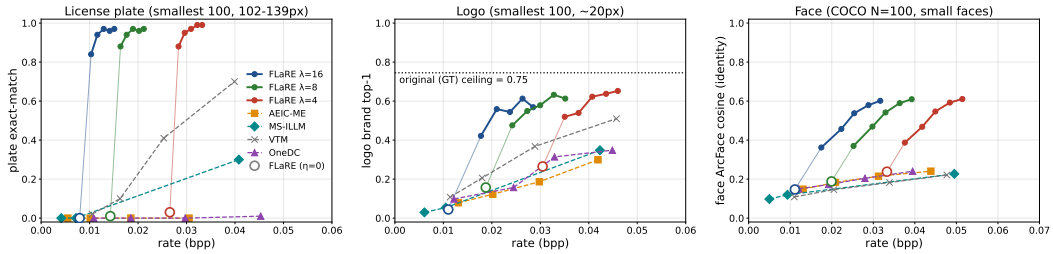
Recognizers (forced readout). Applied to each reconstruction on the GT box: HyperLPR3 for plates (alphanumeric character accuracy and exact-string match vs. the CCPD string; the original scores 1.00 exact); InsightFace ArcFace for faces (cosine of the recon embedding, landmark-aligned on the GT face, to the original embedding—a forced identity-preservation measure); and a ResNet-50 brand classifier (2486 classes, 88.9% clean top-1) for logos (forced top-1/top-5 on the GT crop).

Competitors. AEIC-ME, OneDC, MS-ILLM, and VTM compress the *whole* image (no ROI field) and are swept to bracket FLARE's per-image rate; the same recognizer scores every reconstruction. This experiment characterizes the interface rather than the codec alone: FLARE receives the task region at the encoder (dataset boxes for plates and logos, an off-the-shelf face detector run on the source for faces), whereas the competitors have no region input. It shows what a task field buys at these rates, not that FLARE would outperform a codec equipped with its own ROI mechanism, such as VTM with block-level delta-QP.

Results. Fig. 21 reports rate vs. recognition. On the small-ROI plate and logo sets and on faces, FLARE's realism floor (η=0, open markers) sits down among the baselines, and turning on the excess field lifts recognition steeply: near 0.01 bpp FLARE reaches 0.84 plate exact-match, whereas every rate-matched baseline (VTM, OneDC, AEIC-ME, and MS-ILLM at its very-low-rate operating points) stays at ≤ 0.02, and at 0.02–0.05 bpp FLARE preserves 2–3× the face identity of any baseline. On the smallest logos FLARE leads by 2–4× top-1 at matched rate.

G Rate–Distortion Curves and BD-rate on the Subjective-Study Set

To complement the matched-rate subjective study (Appendix C), we report full rate–distortion (RD) behavior on the exact 40-image set used for the human study (Kodak24 (Eastman Kodak Company, 1999) plus 16 native-resolution JPEG-AI images), evaluated with objective metrics across the whole rate range. FLARE is instantiated at five operating points from the same DeepGaze-based excess fidelity-need field construction as the subjective stimuli, with $\lambda \in \{2, 4, 8, 16, 32\}$; because the lowest-rate operating point ($\lambda = 32$, ≈ 0.0136 bpp) sits below the useful quality range of the perceptual metrics and destabilizes the polynomial BD-rate fit, we exclude it from the curves and the BD-rate integration, leaving four FLARE points spanning 0.0166–0.0584 bpp (including color statistics). Rates include the *real* entropy-coded payload and required color-correction statistics, retaining the original container conventions of the baseline runs; Kodak images are coded without tiling. Competitors are the neural realism codecs AEIC-ME and OneDC, the generative codecs MS-ILLM (Muckley et al., 2023) and GLC (Jia et al., 2024), and the VVC intra anchor (VTM) (Bross et al., 2021). Multi-step diffusion codecs (PerCo, DiffEIC, DiffC) are not included in this 40-image comparison: OneDC and AEIC-ME report the strongest results among generative codecs at these rates, including over multi-step samplers, and StableCodec, the predecessor of AEIC-ME by the same authors, is superseded by it over 0.005–0.035 bpp; the multi-step codecs appear on the complexity axis of Table 1 and in the qualitative comparison of Fig. 1. Among the recent one-step codecs of Sec. 2, DiT-IC (Shi et al., 2026) is the only one with released weights and is compared on Kodak, DIV2K, and CLIC in Fig. 25; CADC (Sheng et al., 2026) has no public implementation, and the SODEC repository (Chen et al., 2026) contains neither code nor checkpoints at the time of writing, so neither can be run at our operating points. AGDL (Zhang & Wu, 2021) is likewise not compared: its repository contains no code or weights, and it refines a JPEG base layer at 0.16–0.24 bpp toward ROI PSNR, an order of magnitude above our rates and without a realism objective. We report PSNR and, for perceptual fidelity, LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022), and a patch-FID ("FID") computed over 256×256 crops. On this 40-image set the patch-FID is estimated from a few hundred crops, far fewer than the 2048-dimensional Inception feature space, so its absolute



(a) Plate (smallest 100) (b) Logo (smallest 100) (c) Face (COCO)

Figure 21: **Rate vs. downstream recognition** for FLARE (three λ operating points, η sweep; open marker = realism floor η=0) against rate-matched AEIC-ME, OneDC, MS-ILLM, and VTM. FLARE reaches recognition at far lower rate on the small-region tasks.

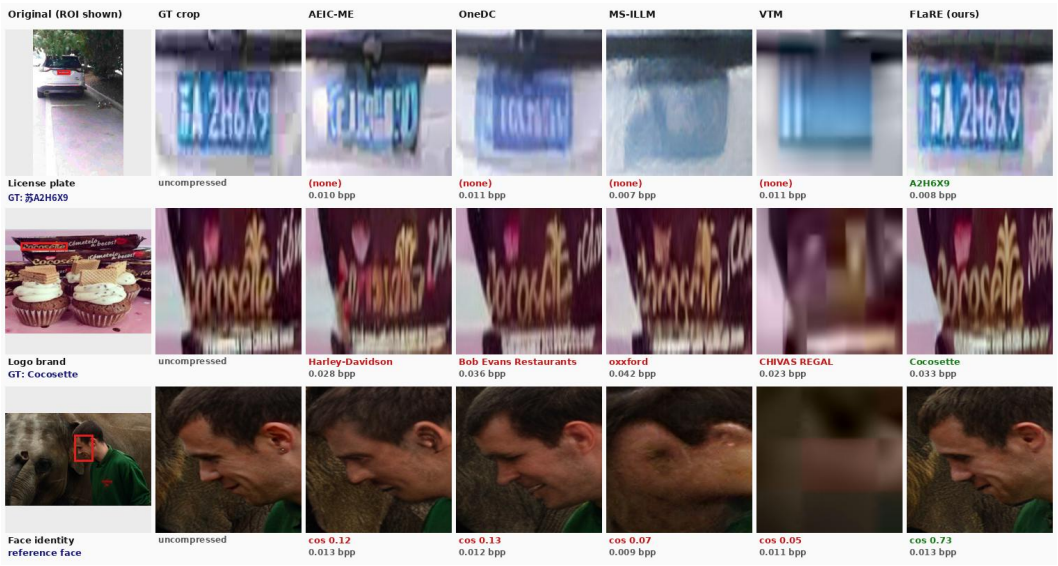
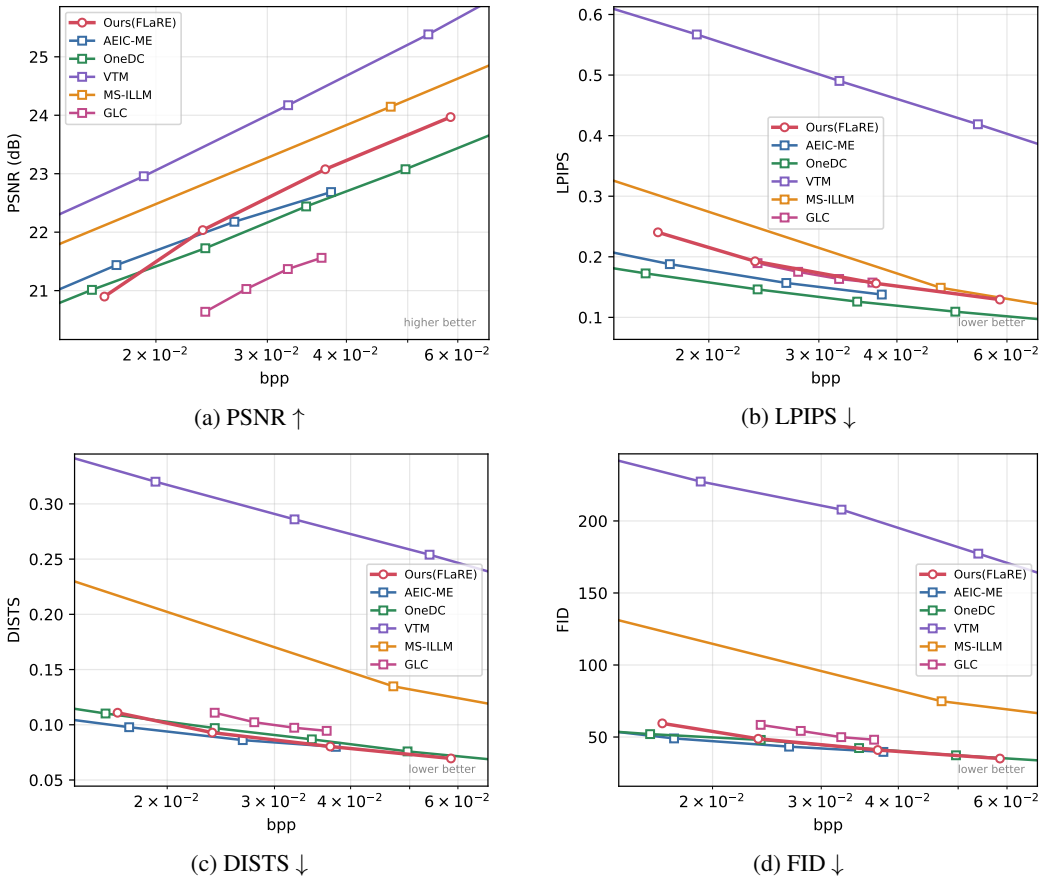


Figure 22: **Task-specific reconstructions at matched (or lower) rate.** For each scenario, we show the full image with the ROI, the uncompressed GT crop, the four rate-matched baselines, and FLARE. The competitors garble the task-critical content while FLARE, given a binary ROI field, preserves it: the plate string is read correctly (and at *lower* rate than every baseline), the logo brand is classified correctly, and face identity is retained (ArcFace cosine 0.73 vs. $\leq$0.13 for every baseline).

values and the rate savings derived from it carry substantial finite-sample noise; we report it for completeness. The RD panels are drawn on a log-rate axis restricted to a window that tightly brackets FLARE's operating range, so that the comparison is read at the rates where FLARE actually operates rather than being dominated by the far higher-rate regimes of VTM and MS-ILLM.

The radar (Fig. 24) and Table 13 summarize these curves as BD-rate against FLARE on the objective axes. The subjective study provides only two points per codec, so the standard four-point cubic BD-rate procedure does not apply; for the subjective axis we connect each codec's two points linearly in log-rate, integrate the horizontal rate difference over the MOS range common to both codecs, and report the average, which we call the *two-point rate saving*. The picture is deliberately honest: on full-frame objective metrics FLARE trades wins and losses with the realism codecs (e.g., FLARE needs less rate than OneDC at equal PSNR but AEIC-ME edges FLARE on LPIPS/DISTS/FID), and against the high-rate VTM and MS-ILLM anchors the objective comparison is partly out of overlap. The decisive axis is the subjective one: at matched rate, human observers prefer FLARE by a large margin, corresponding to a two-point rate saving of $+83.4\%$ over AEIC-ME and $+55.9\%$ over OneDC (overlap-only, no extrapolation). In other words, the rate that FLARE redirects toward



(a) PSNR ↑ (b) LPIPS ↓

(c) DISTS ↓ (d) FID ↓

Figure 23: Per-metric RD curves on the 40-image subjective set (bpp including required color statistics, log-rate axis bracketing FLARE's operating range). FLARE is the anchor (red). On whole-image objective metrics FLARE is competitive but not dominant: because its coded fidelity is concentrated in attention-salient regions rather than spread uniformly, gains that are visible to observers are partly averaged out by full-frame metrics. This is the expected signature of allocation driven by an attention-derived excess fidelity-need field and motivates the subjective evaluation.

Table 13: BD-rate (%) of each competitor relative to FLARE (anchor) on the 40-image subjective set; the last column is the two-point rate saving on MOS (defined in the text), not a BD-rate. Positive = competitor needs more rate for equal quality = FLARE better; negative = FLARE needs more rate. "∗": quality ranges do not overlap, sign by strict dominance (magnitude capped, integration not defined). "–": subjective point unavailable.

Competitor	PSNR	LPIPS	DISTS	FID	Subjective MOS (two-point)
AEIC-ME	−2.5	−31.6	−16.6	−25.4	+83.4
OneDC	+20.2	−47.0	+11.9	−11.1	+55.9
VTM	−48.3	+1025	+150*	+150*	–
MS-ILLM	−29.2	+6.7	+371.1	+376.8	–
GLC	+67.8	−2.5	+49.8	+44.6	–

attention-salient fidelity yields perceptual quality that whole-image metrics understate but observers clearly register.

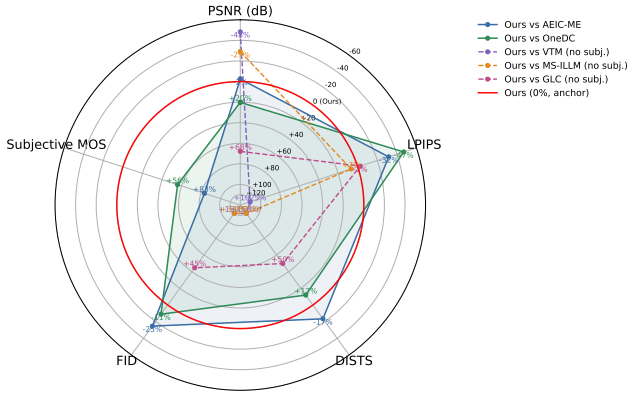


Figure 24: BD-rate radar with FLARE as the anchor. Each objective axis is the BD-rate of a competitor relative to FLARE, and the subjective axis is the two-point rate saving on MOS defined in the text (positive = the competitor needs *more* rate for equal quality = FLARE better), so FLARE-advantage is compressed toward the center and the red ring marks FLARE's 0% parity. Concentric circles are equally spaced in BD-rate. The subjective (MOS) axis is computed on overlapping rates only, without extrapolation, and is available only for AEIC-ME and OneDC (solid polygons); competitors without a subjective point are drawn as dashed open paths. "$*$" marks axes where the objective quality ranges do not overlap, for which the sign is set by strict dominance and the magnitude capped. Vertex labels give each BD-rate; magnitudes above 110% are clipped to the outer ring and labeled with the true value.

H CHANNEL SPLIT IMPROVES OPTIMIZATION UNDER A MATHEMATICALLY IDENTICAL OBJECTIVE

This appendix isolates the effect of the base/message *channel split* (Sec. 4.1) from any difference in the training objective, and shows that the split alone yields a better optimum.

Setup. We compare two architectures finetuned from their respective 370k-step pretrains: FLARE (the 48/272 split of Eq. 7) and a *no-split* ablation in which the entire 320-channel latent is one message group—the scale network acts on all channels, there is no gate on the coded path, and only the training-only auxiliary-decoder input is gated by the excess fidelity-need field. Crucially, we make the two *training objectives mathematically identical*. The deployed loss (adversarial term aside) is $L = R + \lambda_{\ell_2}\ell_2 + \lambda_{\mathrm{D}}\mathrm{DISTS} + \lambda_{\mathrm{C}}\mathrm{CLIP} + A\,\mathrm{MSE}_{\mathrm{aux}}$. We set the rate weight to unity, so FLARE's groupwise rate $R(\mathbf{y}^{\mathrm{base}}) + R(\mathbf{y}^{\mathrm{msg}})$ equals the no-split total rate R; and we set a common excess-fidelity weight $A = \lambda_{\mathrm{fid}} \cdot 255^2$ shared by both models. All other terms, hyperparameters, data, and schedule are identical; only the architecture differs. We evaluate on a held-out set the *adv-excluded* objective (the adversarial term uses a per-model discriminator and is not comparable across runs), driving each model with excess fidelity-need fields drawn from the training distribution.

Phenomenon. Under this identical objective, the split model attains a lower loss. Fig. 26 (held-out, term-decomposed) gives $L_{\text{split}} < L_{\text{no-split}}$ (0.498 vs. 0.548 at 20k steps), and Fig. 28 shows the split loss below the no-split loss throughout training—both with and without the adversarial term—with the gap *widening* as training proceeds ($0.044 \to 0.050$ from 10k to 20k). Spatially (Fig. 27), the split concentrates coded rate on the regions selected by the excess fidelity-need field (target-region bit share $\approx$29–34$\%$ of a $\approx$3–9$\%$ area), whereas the no-split model spreads rate into the background and codes more bits in total (e.g., 70.5k vs. 56.1k on kodim05).

Cause. Decomposing the loss gap term-by-term (Fig. 26), the advantage is dominated by the auxiliary reconstruction distortion $A\,\mathrm{MSE}_{\mathrm{aux}}$: the split's coded latent conveys source fidelity in the requested region far more efficiently (lower aux MSE) at *less* rate, conceding only a small amount on the global perception terms (ℓ_2/DISTS/CLIP). Mechanistically, $\mathbf{y}^{\mathrm{msg}}$ specializes in source-specific fidelity through scaling and gating by the excess fidelity-need field, while $\mathbf{y}^{\mathrm{base}}$ remains always coded and is not directly scaled or gated by that field. Both groups still contribute to the delivered reconstruction and receive floor supervision; the distinction is architectural, not a mapping of the objective-level floor and excess to separate latent groups. The no-split model instead uses one latent

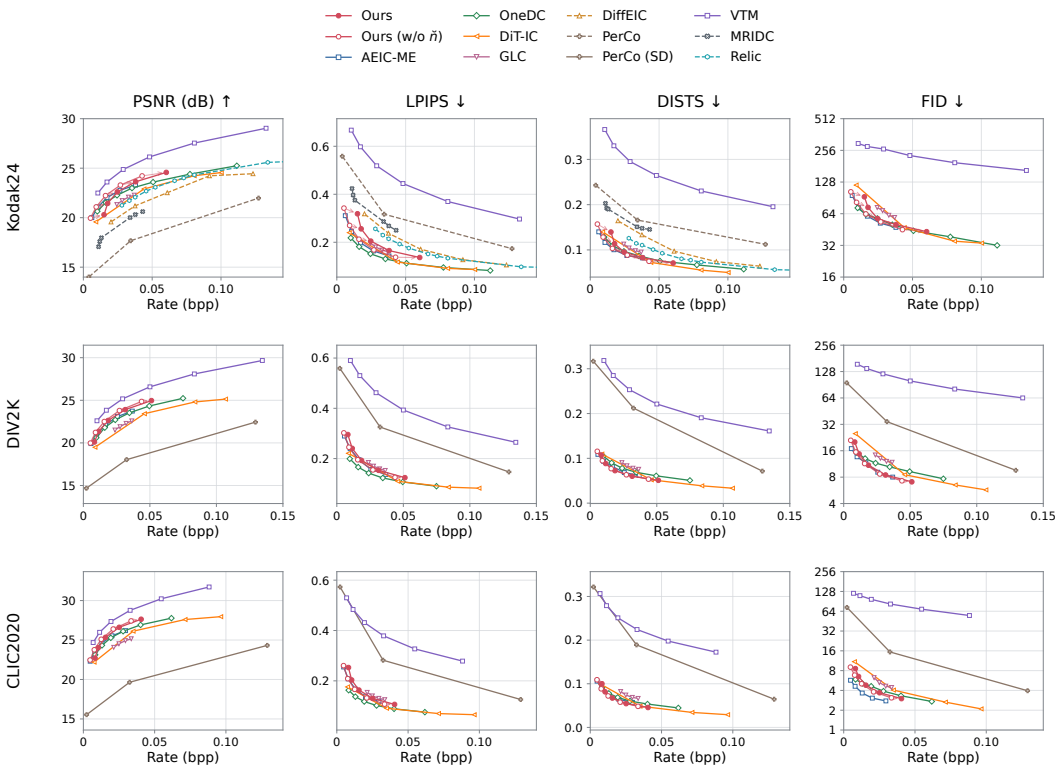


Figure 25: **Rate–quality comparisons on Kodak24, DIV2K, and CLIC2020.** Rows show Kodak (24 images), DIV2K validation (100), and CLIC2020 test (428), at native resolution for local evaluations; columns show PSNR, LPIPS, DISTS, and patch-FID. FLARE uses the subjective-study field construction at all five checkpoints ($\lambda \in \{32, 16, 8, 4, 2\}$). Filled/open red markers denote the study field/zero excess; translucent arrows connect zero excess to the study field at identical weights. Rate axes are linear; FID axes are logarithmic (base 2). Local FID uses 256×256 patches from regular and half-patch-shifted grids. Rates are mean per-image bpp. DIV2K/CLIC rates include container and required side-information bytes; the earlier Kodak baseline results retain their original accounting, including the entropy-estimated GLC rate, while both FLARE curves on Kodak use real payload bytes. GLC on DIV2K/CLIC uses a local rANS implementation with the released weights. PerCo (SD) is a separate implementation from the published PerCo curve. On DIV2K/CLIC, FLARE and AEIC-ME use the same tiled UNet/VAE decoding. DiT-IC (Shi et al., 2026) is run locally from its four released checkpoints with real entropy coding, full-frame at native resolution, rates as bitstream bytes over the original pixels. Dashed Kodak curves are published or digitized results for DiffEIC (Li et al., 2024), PerCo (Careil et al., 2024), MRIDC (Xu et al., 2025), and Relic (Relic et al., 2026) (uniform timestep map); their protocols are not fully harmonized with the local evaluations. Unavailable metrics or dataset results are omitted.

for both the global reconstruction/perception objective and the requested excess fidelity. Because its coded path has no gate, it cannot switch background rate off and spends 0.093 vs. 0.069 bpp even when the excess fidelity-need field is zero, yielding worse fidelity per bit. The comparison therefore contrasts the split together with the coded-path gate it enables against a single ungated group; the gate presupposes the split, since only a separate message group can be switched off without removing the base code, so the two are not separable ablation factors here. Hence, although the objectives are mathematically identical, the split provides an inductive/optimization structure that reaches a better optimum—justifying the channel split of Sec. 4.1 on optimization, not merely interface, grounds.

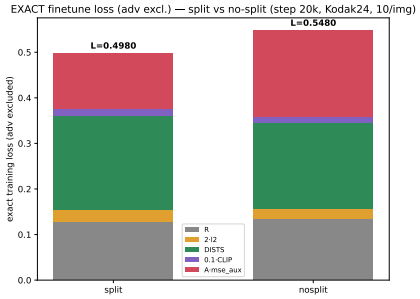


Figure 26: **Identical-objective loss (adv excluded), held-out, term-decomposed.** The split reaches a lower total loss; the gap is dominated by the auxiliary reconstruction distortion $A \cdot \mathrm{MSE}_{\mathrm{aux}}$ while the no-split model spends slightly less on the global perception terms at the cost of more rate.

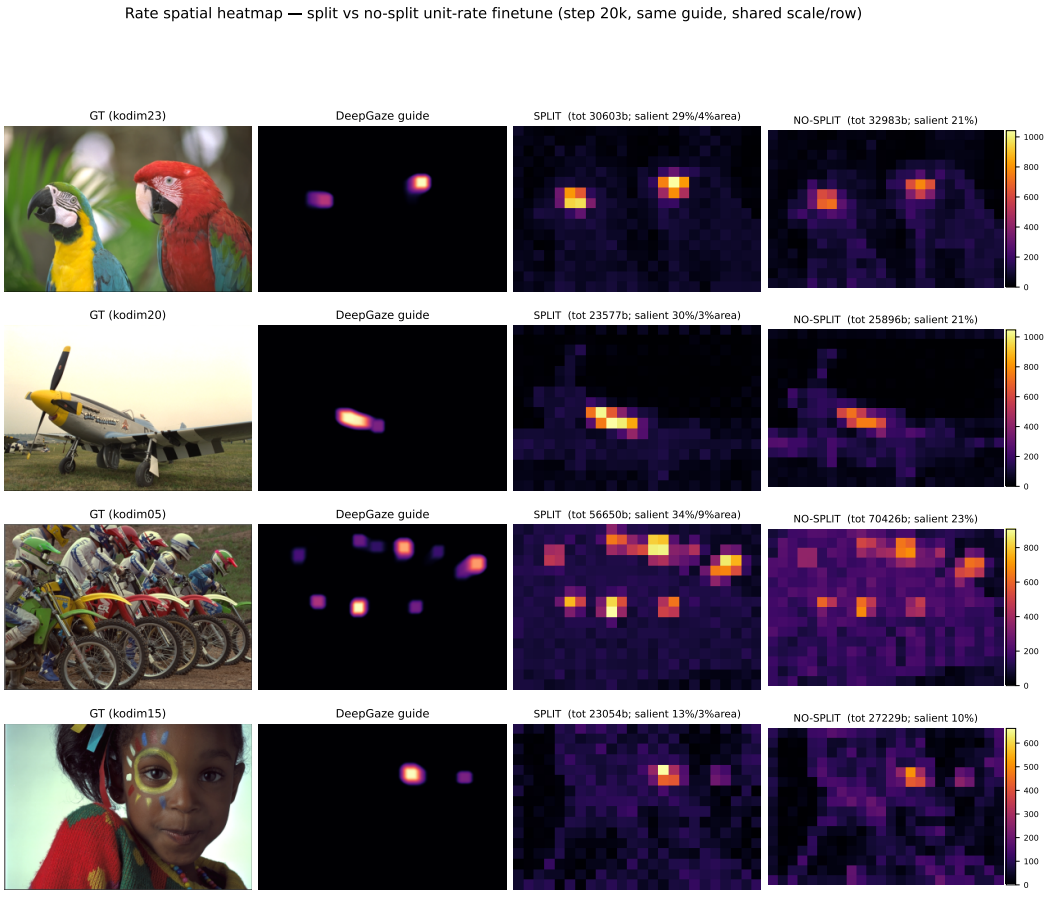


Figure 27: **Spatial rate allocation (same excess fidelity-need field, shared scale per row).** The split concentrates coded bits on the requested regions; the no-split model spreads rate into the background and codes more bits in total (per-panel: total bits and target-region bit share).

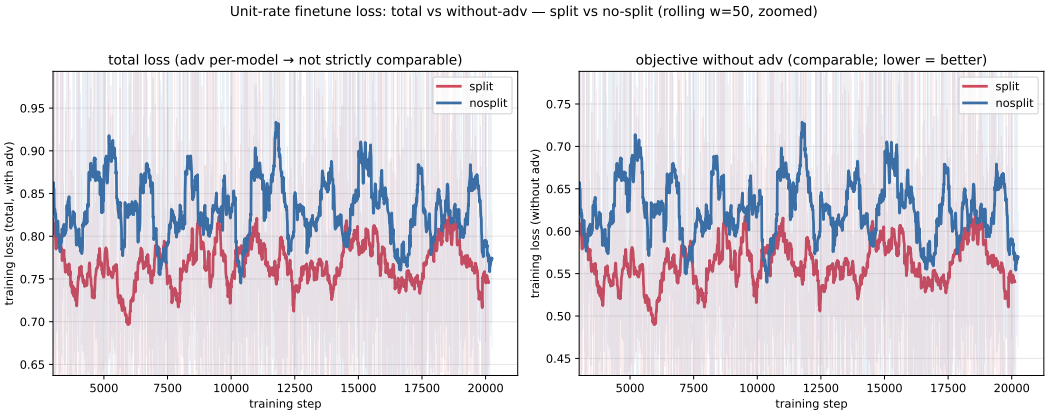


Figure 28: **Training-loss trajectory** (left: total; right: adversarial term excluded; rolling mean). The split loss stays below the no-split loss throughout, and the gap widens with training.