THINKING-THEN-COMPRESS WITH PERCEPTION-INFORMED SPATIAL BUDGETING FOR GENERATIVE IMAGE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative image compression introduces a generative model at the decoder to fill in details missing from the bitstream using image priors, and has become a prominent approach for perceptual compression at low bitrates. However, when a large generative decoder cannot be jointly optimized with the encoder, the encoder's compression decisions are no longer directly adapted to the decoder's reconstruction behavior. To address this issue, we propose **Thinking-then-Compress**, a closed-loop image compression paradigm instantiated as encoder-side search over transmitted discrete spatial schemes using reconstruction feedback from a deployed decoder. The encoder first analyzes perceptually important content, proposes a regional scheme, evaluates the resulting reconstruction, and revises the scheme within a fixed feedback budget. Within this paradigm, Perception-Informed Spatial Budgeting proposes regional tier updates from image importance, codec error, and observed generative recovery, then verifies joint proposals using reconstruction feedback and their rate cost. Reconstruction-Aware Generative Enhancement further fine-tunes the generation branch and generation-side adapters for compressed inputs and uses reconstruction feedback to compose codec and generative outputs spatially, preserving reliable codec details while introducing generation where it improves recovery. Under the reported evaluation protocol, our method achieves favorable performance against existing perceptual image compression methods. Our project will be publicly available after publication.

1 INTRODUCTION

Image compression aims to balance bitrate against reconstruction quality. With the progress of deep learning, end-to-end learned image compression has substantially improved this trade-off through joint optimization of the encoder and decoder (Ballé et al., 2018; He et al., 2022). At extremely low bitrates, however, the limited bit budget makes it increasingly difficult to preserve fine image details, and distortion-oriented codecs often produce overly smooth reconstructions with reduced perceptual quality (Mentzer et al., 2020).

Generative image compression emerged to address this limitation. It introduces a generative model at the decoder, which uses an image prior to synthesize plausible details absent from the bitstream (Mentzer et al., 2020; Kuang et al., 2024). By complementing the transmitted code with a learned image prior, reconstruction depends on both transmitted content and model-supplied details. In jointly trained generative codecs, the encoder learns what to preserve into a compact representation while the decoder learns to reconstruct from that representation (Agustsson et al., 2019; Mentzer et al., 2020).

Recent generative systems have become larger and structurally more complex, with some combining multimodal understanding and image generation within a single model (Xie et al., 2025; Deng et al., 2025; Zhou et al., 2025). Integrating such models into compression makes full joint optimization with the encoder increasingly costly. An alternative is to fine-tune a pretrained generative model to restore compressed images while keeping the codec fixed. However, this adapts the generative model to the codec's outputs without jointly adapting the encoder's compression decisions to the

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.

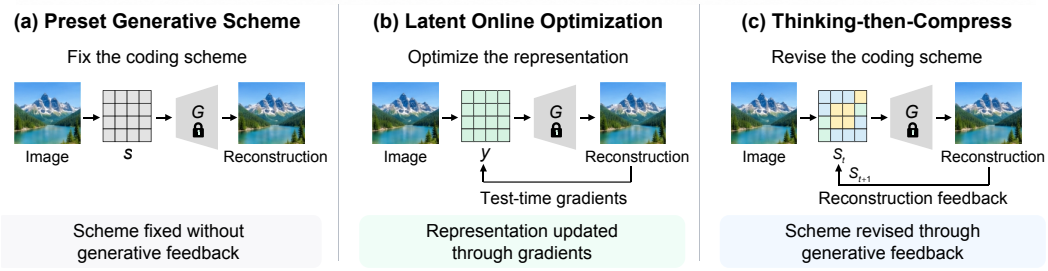

Figure 1: Comparison of inference-time adaptation strategies for compression with a fixed generative decoder. (a) A preset generative scheme is not revised after observing the generated reconstruction. (b)Latent online optimization updates a continuous representation through test-time gradients, while (c) Thinking-then-Compress revises the spatial coding scheme according to the observed recovery behavior of the generative decoder. The distinction lies in using generative reconstruction to adjust the division of information between the bitstream and the generative prior.

generative model's reconstruction capabilities. The core problem is therefore not merely how to compress an image, but how to determine which visual content to preserve in the bitstream and which to leave to the generative prior.

The decision about what to preserve in the bitstream depends on the decoder's reconstruction behavior. Yet this behavior is conditioned on the compressed bitstream produced by the same decision, making the compression scheme and reconstruction mutually dependent. A scheme selected without observing the resulting reconstruction may not make effective use of the decoder's capabilities, which motivates the encoder to revise the scheme through iterative reconstruction feedback.

Because the decoder's restoration capability varies across regions, additional transmitted information may provide limited benefit where recovery is already effective, but remain valuable where reconstruction is unreliable. A spatially uniform scheme may therefore spend bits where they provide little benefit while leaving other regions insufficiently represented, motivating spatially selective decisions guided by reconstruction feedback. Regional variation also means that generation may improve poorly preserved content while altering reliable codec details, so the codec and generative reconstructions should be combined according to their observed regional performance.

These considerations motivate **Thinking-then-Compress**, a closed-loop image compression paradigm that derives the compression scheme through feedback from the deployed decoder. The encoder performs a rate-aware search over discrete regional compression schemes whose side information is explicitly accounted for. In Thinking-then-Compress, image understanding identifies perceptually important content, and iterative reconstruction feedback determines how that content should be represented in the bitstream. Figure 1 contrasts this generative-feedback scheme revision with a preset generative scheme and latent online optimization.

Within this paradigm, **Perception-Informed Spatial Budgeting** iteratively refines regional compression decisions using image analysis and reconstruction feedback. **Reconstruction-Aware Generative Enhancement** further adapts the generative decoder to compressed inputs and combines codec and generative reconstructions according to their observed regional performance. Extensive experiments demonstrate the effectiveness of our method for perceptual image compression at extremely low bitrates. Our contributions are summarized as follows:

- We propose Thinking-then-Compress, a closed-loop image compression paradigm that searches over transmitted discrete spatial schemes during encoding using reconstruction feedback from a fixed deployed enhancer. The encoder uses image analysis to propose a scheme, observes the decoder's reconstruction, and revises the scheme until no candidate is accepted or the feedback budget is exhausted.
- Within this paradigm, we devise Perception-Informed Spatial Budgeting, which provides regional tier updates from importance, codec error, and generative recovery, then verifies their effectiveness through reconstruction feedback from the fixed decoder.
- To realize region-dependent reconstruction at the decoder, we develop Reconstruction-Aware Generative Enhancement, which adapts the generative model using paired codec reconstructions and original images. Using the codec and converged provisional reconstructions, the encoder selects and transmits quantized regional mixing weights, allowing

the decoder to preserve reliable codec details while exploiting full-step generation where it improves recovery.

2 Related Work

2.1 Learned Image Compression

Learned image compression replaces hand-crafted transforms and entropy models with neural networks optimized under a rate-distortion objective. Early studies established end-to-end nonlinear transform coding and introduced hierarchical priors to improve entropy estimation (Ballé et al., 2017; 2018). Subsequent methods further improved coding efficiency through autoregressive dependencies, expressive transforms, and spatial and channel context modeling (Minnen et al., 2018; Cheng et al., 2020; He et al., 2022). These methods achieve strong rate-distortion performance by jointly adapting the encoder and decoder to a shared training objective.

At low bitrates, generative compression introduces an image prior to synthesize details that cannot be fully preserved in the bitstream. Adversarial methods improve perceptual quality by matching the distribution of reconstructed and natural images (Agustsson et al., 2019; Mentzer et al., 2020; Muckley et al., 2023; Agustsson et al., 2023). Diffusion-based methods provide a stronger generative prior for perceptual reconstruction (Careil et al., 2024; Lei et al., 2023; Kuang et al., 2024; 2026), while GLC and DLF perform compression with structured generative latent representations (Jia et al., 2024; Xue et al., 2025). Despite their architectural differences, these methods require the encoder to determine what information should be transmitted and what can be recovered from the generative prior. This decision is generally learned through joint training or produced in a single forward pass before the reconstruction of the deployed decoder is observed.

Content-adaptive compression further adjusts the representation to individual images or spatial regions. Spatially adaptive methods predict regional coding decisions from the input image or encoder-side features (Xu et al., 2025; Sheng et al., 2026). Instance-adaptive methods optimize latent representations, model parameters, compensation information, or diffusion decoding parameters for each image (Campos et al., 2019; Jiang et al., 2022; Kuang et al., 2025; Brenig & Timofte, 2025). These approaches adapt a representation or decoder configuration, whereas our setting asks which content should be transmitted after observing the fixed enhancer's reconstruction. Our method instead performs instance-specific, rate-aware optimization over a transmitted discrete spatial compression scheme while keeping the generative enhancer fixed. Candidate schemes are evaluated using reconstruction feedback from the fixed generative enhancer.

2.2 Large Generative Models

Large generative models have substantially advanced image synthesis through expressive visual priors and scalable architectures. Diffusion models establish high-quality generation through iterative denoising (Ho et al., 2020; Dhariwal & Nichol, 2021), while latent diffusion performs this process in a compact latent space to reduce computational cost (Rombach et al., 2022). Large-scale text-conditioned models further extend these priors to diverse image synthesis and restoration tasks (Ramesh et al., 2022; Saharia et al., 2022). More recent unified models integrate multimodal understanding and image generation within a shared architecture (Xie et al., 2025; Wu et al., 2025; Zhou et al., 2025; Deng et al., 2025). These developments make pretrained generative models increasingly attractive as reconstruction priors for perceptual image compression (Lei et al., 2023; Careil et al., 2024; Kuang et al., 2024).

Existing generative codecs commonly train their generative decoders jointly with the encoder (Agustsson et al., 2019; Mentzer et al., 2020; Muckley et al., 2023), while methods based on pretrained generators typically incorporate them under a predefined coding formulation (Lei et al., 2023; Careil et al., 2024). Pretrained generative models can also be adapted to downstream inputs through task-specific fine-tuning or additional conditioning modules (Zhang et al., 2023; Mou et al., 2024). For image compression, such adaptation can teach a generator to restore codec outputs, but it does not jointly adapt the encoder's compression decisions to the resulting reconstruction behavior. Fully optimizing an increasingly large generator together with the encoder also incurs substantial training cost (Rombach et al., 2022; Deng et al., 2025). An alternative is to adapt the encoder-side compression decisions while keeping the generative model fixed. Our method follows this setting

and uses reconstructions from the adapted model to search over deployable spatial compression schemes.

3 THINKING-THEN-COMPRESS PARADIGM

3.1 PRELIMINARIES AND MOTIVATIONS

Classical rate-distortion theory characterizes the minimum coding rate required to satisfy a prescribed expected distortion (Shannon, 1959).

In a learned codec, the encoder maps the source to a quantized representation $Y = Q(E_\phi(X))$, and the decoder reconstructs the image as $\hat{X} = D_\theta(Y)$. Joint optimization under a rate-distortion objective allows the encoder to learn what information to preserve and the decoder to learn how to reconstruct it, thereby adapting the transmitted representation to the decoder's reconstruction capabilities.

Generative compression reconstructs the image as $\hat{X} = G_\theta(Y, \xi)$, where $Y = Q(E_\phi(X))$ is the quantized representation, G_θ is the generative decoder, and ξ denotes sampling randomness. The resulting reconstruction depends on both the information retained in Y and the learned image prior encoded by G_θ.

When the encoder and generative decoder are jointly trainable, they can be optimized through

$$\min_{\phi,\theta} \mathbb{E}_{X,\xi} \left[R_\phi(X) + \lambda_{\mathrm{RD}} \, \ell \left(X, G_\theta \left(Q(E_\phi(X)), \xi \right) \right) \right], \tag{1}$$

where $R_\phi(X)$ denotes the estimated coding rate, ℓ denotes the reconstruction error, and $\lambda_{\mathrm{RD}} > 0$ controls their relative weight. Under this objective, the value of retaining information in the bitstream depends on how the generative decoder uses that information and recovers the missing content, so the encoder's compression decisions are adapted to the decoder's reconstruction capabilities.

In practice, full joint optimization in Eq. 1 may be unavailable when a large pretrained generative decoder is introduced into the compression system. A practical alternative is to adapt the generative decoder to compressed inputs while keeping the codec fixed. This improves the decoder's ability to reconstruct codec outputs, but does not adapt the encoder's compression decisions to the decoder's reconstruction capabilities, so the mutual adaptation provided by joint optimization is not established.

To approximate decoder-aware adaptation without updating the trained decoder or reproducing joint training, we adjust the transmitted scheme for each input according to its observed reconstruction behavior. Let $s \in \mathcal{S}$ denote an admissible compression scheme. With the decoder parameters fixed at $\bar{\theta}$, the reconstruction under scheme s is expressed as

$$y_s = E_s(x), \qquad c_s = D_{\mathrm{c}}(y_s; s), \qquad \hat{x}_s = \mathcal{R}_{\bar{\theta}} \left(c_s, s, \bar{\xi} \right), \tag{2}$$

where c_s is the codec reconstruction, $\mathcal{R}_{\bar{\theta}}$ denotes the fixed reconstruction process, and $\bar{\xi}$ denotes fixed sampling randomness used for consistent scheme evaluation.

For a scheme s and regional mixing weights $\boldsymbol{\alpha}$, we formalize the final reconstruction as

$$z_s = G_{\bar{\theta}}^{(T_{\mathrm{f}})}(c_s; \bar{\xi}), \qquad a_{\boldsymbol{\alpha}} = \mathcal{F}(\mathcal{Q}(\boldsymbol{\alpha})), \qquad \hat{x}_{s,\boldsymbol{\alpha}} = (1 - a_{\boldsymbol{\alpha}}) \odot c_s + a_{\boldsymbol{\alpha}} \odot z_s. \tag{3}$$

The total rate used for optimization and evaluation is

$$R_{\mathrm{total}}(s, \boldsymbol{\alpha}; x) = R_y(s; x) + R_s(s) + R_a(\boldsymbol{\alpha}), \tag{4}$$

where R_y, R_s, and R_a denote the accounted lengths of the image representation, spatial scheme, and quantized regional weights, respectively. The corresponding ideal constrained target is

$$(s^\star, \boldsymbol{\alpha}^\star) \in \underset{s \in \mathcal{S}, \boldsymbol{\alpha} \in [0,1]^N}{\arg\min} \ell(x, \hat{x}_{s,\boldsymbol{\alpha}}) \quad \text{s.t.} \quad R_{\mathrm{total}}(s, \boldsymbol{\alpha}; x) \le B. \tag{5}$$

Here, B is the ideal total bit budget and b_α is the fixed number of bits used for each regional composition weight. Because the fixed partition contains N regions, $R_a(\boldsymbol{\alpha}) = N b_\alpha$ regardless of the weight values, leaving $B_s = B - N b_\alpha$ as the ideal budget for the coded image and spatial

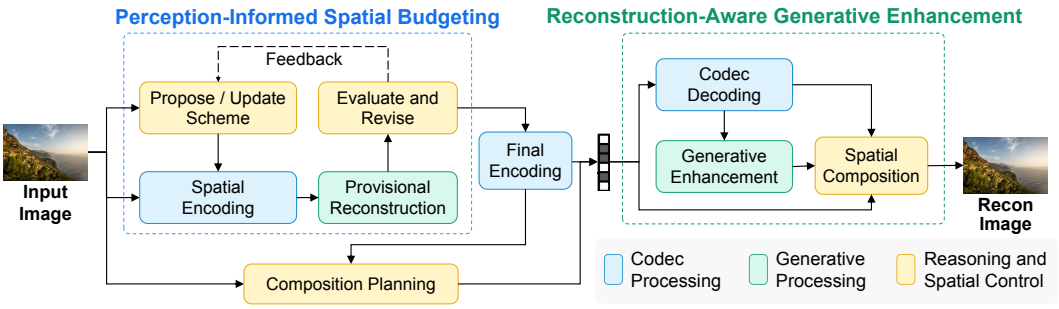

Figure 2: Overview of the proposed Thinking-then-Compress paradigm. Perception-Informed Spatial Budgeting iteratively refines the compression scheme using reconstruction feedback. After the call-budget-limited, rate-aware search, the selected regional weights guide the spatial composition of the codec and generative reconstructions. Blue, green, and yellow denote codec processing, generative processing, and scheme and composition control, respectively.

scheme. Equation 5 is the ideal constrained target, while the reported implementation uses a two-stage Lagrangian approximation and reports the achieved rate rather than guaranteeing an exact prescribed bitrate. The first stage searches for s with provisional generation and fixed tier anchors, and the second selects $\boldsymbol{\alpha}$ from the converged provisional reconstruction after fixing s. The region partition and fixed sampling configuration are shared protocol elements and require no transmitted description. Unlike Eq. 1, the decoder parameters remain fixed and the optimization selects the compression scheme according to the reconstruction produced by the deployed decoder.

For clarity, let $\boldsymbol{\alpha}^\star(s)$ denote the composition weights selected for a fixed scheme s, and define $\hat{x}_s\hat{x}_{s,\boldsymbol{\alpha}^\star(s)}$ and $L(s)\ell(x,\hat{x}_s)$. For a candidate update from s to s', the resulting reconstruction improvement is

$$\Gamma(s \to s') = L(s) - L(s') = \ell(x,\hat{x}_s) - \ell(x,\hat{x}_{s'}). \tag{6}$$

The update relation in Eq. 6 suggests two practical requirements for scheme selection.

- **Candidates must be evaluated by their resulting reconstructions.** Changes in codec error need not predict changes in the output of the generative decoder, so candidate assessment must account for the generated reconstruction rather than codec quality alone.
- **The benefit of an update is state-dependent.** Because the generative reconstruction is conditioned on the current codec output, the benefit of a regional update depends on the current scheme. Candidate updates must therefore be evaluated using their resulting reconstructions, with feedback refreshed after the scheme changes.

Thinking-then-Compress therefore evaluates candidate schemes through reconstructions produced by the fixed deployed decoder, as instantiated by the spatial budgeting procedure below.

3.2 Overview

Fig. 2 shows the framework built on a base image codec and an adapted generative model. Perception-Informed Spatial Budgeting determines the compression scheme, while Reconstruction-Aware Generative Enhancement produces and composes the codec and generative reconstructions.

Given an image x and a nominal rate operating point λ, Perception-Informed Spatial Budgeting returns a region-dependent scheme using importance, perceptual error, and reconstruction feedback. This process produces a selected scheme s, its coded representation y, and the corresponding codec reconstruction c:

$$s = \mathcal{P}(x,\lambda;G_{\bar{\theta}}), \qquad y = E_s(x), \qquad c = D_c(y;s), \tag{7}$$

where $\mathcal{P}$ denotes the complete spatial budgeting procedure, detailed in Section 3.3.

Reconstruction-Aware Generative Enhancement adapts the generative model using paired codec reconstructions and original images, then keeps its parameters fixed during encoding and decoding.

Once the coding scheme is determined, reconstruction feedback is used at the encoder to select regional mixing weights. The decoder recovers the codec image c and the weight field a from the transmitted representation and side information, and reconstructs the image as

$$z = G_{\bar{\theta}}(c), \qquad \hat{x} = (1 - a) \odot c + a \odot z, \tag{8}$$

where $\odot$ denotes element-wise multiplication. The model adaptation and spatial composition mechanisms are described in Section 3.4. Scheme search and weight selection take place at the encoder. The decoder requires neither the original image nor repetition of these procedures.

3.3 Perception-Informed Spatial Budgeting

Perception-Informed Spatial Budgeting instantiates $\mathcal{P}$ in Eq. 7 by iteratively refining a region-dependent compression scheme for a nominal rate operating point. Regional updates are proposed from perceptual and reconstruction cues and screened by their contextual reconstruction outcomes and rate costs.

Spatial budgeting formulation. We divide the image into N nonoverlapping regions $\{\Omega_i\}_{i=1}^{N}$ and represent a scheme as $s = \{t_i\}_{i=1}^{N}$, where t_i is the compression tier assigned to Ω_i. Starting from a dominant-tier reconstruction, selected regions are encoded at higher tiers and smoothly composed to obtain

$$y_s = E_s(x), \qquad c_s = D_{\mathrm{c}}(y_s; s), \qquad R_y(s; x) + R_s(s) \le B_s, \tag{9}$$

the dominant tier is a uniform whole-image tier selected for the target rate, while regional encodings at higher tiers replace its corresponding regions. Here, $B_s = B - Nb_\alpha$ defines the corresponding ideal scheme budget after reserving the fixed composition-weight cost in Eq. 4. The corresponding syntax and rate calculation are detailed in Appendix A.1. For the reported operating points, we use the Lagrangian counterpart of Eq. 9, retaining rate-increasing updates when their verified gain per bit exceeds an operating-point parameter λ and reporting the resulting achieved rate.

We evaluate each scheme using a provisional reconstruction from the fixed generative decoder

$$\begin{aligned} \tilde{z}_s &= G_{\bar{\theta}}^{(T_{\mathrm{p}})}(c_s), \\ \tilde{x}_s &= A(s) \odot \tilde{z}_s + (1 - A(s)) \odot c_s, \\ J(s) &= \mathcal{D}(x, \tilde{x}_s). \end{aligned} \tag{10}$$

Here, T_{p} is a reduced sampling budget, $A(s) \in [0, 1]^{H \times W}$ is a fixed tier-conditioned composition field obtained by expanding the anchor associated with each regional tier, and $\mathcal{D}$ is a perceptual reconstruction criterion. The same fixed sampling randomness is used for all candidate evaluations so that changes in $J(s)$ reflect changes in the compression scheme. Final generation uses the full sampling budget, and the weights selected from the provisional reconstruction are applied to the final output. This reduced-step proxy lowers encoder computation but relies on regional recovery behavior remaining stable across provisional and final sampling budgets. The final composition weights are selected in Section 3.4.

Perception-informed proposal. Before scheme search, the understanding branch of BAGEL (Deng et al., 2025) is queried once per input, and its cached result provides a regional importance score I_i for each Ω_i. The scores remain fixed throughout the search, so the feedback loop makes no further calls to the understanding model. The codec error and observed generative recovery are defined as

$$d_i^{\mathrm{c}} = \mathcal{D}_i(x, c_s), \qquad g_i = d_i^{\mathrm{c}} - \mathcal{D}_i(x, \tilde{z}_s), \tag{11}$$

where $\mathcal{D}_i$ denotes the reconstruction error over region Ω_i and g_i measures the improvement provided by generation.

We combine these observations into a proposal score

$$q_i = \eta_{\mathrm{c}} \operatorname{Norm}(d_i^{\mathrm{c}}) - \eta_{\mathrm{g}} \operatorname{Norm}(g_i) + \eta_{\mathrm{I}} \operatorname{Norm}(I_i), \tag{12}$$

where Norm denotes min–max normalization across regions. A high q_i favors regions that are important, poorly preserved by the codec, or insufficiently recovered by generation. The score ranks regional upgrades, downgrades, and paired exchanges for feedback verification rather than directly determining the next scheme. The resulting regional proposals are subsequently evaluated through reconstruction feedback.

Feedback-verified refinement. For each candidate s', low-cost screening recomputes its codec reconstruction while reusing the current provisional generative reconstruction, yielding

$$\widehat{\Delta J}(s' \mid s) = J(s) - \mathcal{D}\left(x, A(s') \odot \tilde{z}_s + (1 - A(s')) \odot c_{s'}\right). \tag{13}$$

Infeasible or clearly degrading candidates are removed, and the remaining candidates are ranked using their screened gains and rate changes. Compatible high-ranking updates are merged into a joint scheme s_{J} and verified with one provisional generative reconstruction. For each affected region Ω_i, the implementation measures the contextual gain

$$\gamma_i = \mathcal{D}_i(x, \tilde{x}_s) - \mathcal{D}_i(x, \tilde{x}_{s_{\mathrm{J}}}). \tag{14}$$

For the screened candidate rate change ΔR_i, a rate-increasing update is retained when $\gamma_i > \tau$ and $\gamma_i / \Delta R_i > \lambda$, while a non-increasing update requires only $\gamma_i > \tau$. Here, ΔR_i is the rate change of the screened candidate that introduces the update, τ is the minimum accepted regional gain, and λ is the operating-point-specific gain-per-bit threshold. Because the updates are evaluated jointly, γ_i is a contextual verification signal rather than an isolated causal contribution, and may include cross-region interactions from the generator. The retained updates define the next scheme. When only a subset of a joint proposal is retained, we recompute its provisional reconstruction before it is used for subsequent refinement or composition planning, ensuring that the feedback matches the retained scheme. Complete search pseudocode and generator-call accounting are provided in Appendix A.2.

3.4 Reconstruction-Aware Generative Enhancement

We adapt BAGEL (Deng et al., 2025) to restore details from codec reconstructions using paired compressed and original images. Since the benefit of generation varies spatially, reconstruction feedback further guides regional composition of the codec and generative outputs.

Generative enhancement model. Following BAGEL (Deng et al., 2025), the codec reconstruction c is encoded into latent and semantic conditions

$$h_{\mathrm{c}} = E_{\mathrm{VAE}}(c), \qquad f_{\mathrm{c}} = E_{\mathrm{ViT}}(c). \tag{15}$$

Together with a fixed restoration instruction, these conditions guide the generative branch to produce a latent that is decoded into the enhanced image. We use θ to denote the trainable enhancement parameters and $\bar{\theta}$ to denote their frozen values after fine-tuning. We denote the complete training-time enhancement process by $G_\theta(c)$ and omit the fixed instruction and sampling randomness for brevity.

Reconstruction-oriented training. We construct pairs (c, x) from original images and their codec reconstructions under different spatial compression settings, using c as the condition and x as the target. Following the flow matching formulation of BAGEL (Deng et al., 2025), we encode the target as $h_0 = E_{\mathrm{VAE}}(x)$ and define

$$h_t = (1 - t)h_0 + t\epsilon, \qquad v^\star = \epsilon - h_0, \tag{16}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$ is sampled during training. Conditioned on the compressed image, the generative branch predicts

$$\hat{v}_\theta = v_\theta(h_t, t, h_{\mathrm{c}}, f_{\mathrm{c}}), \qquad \hat{h}_0 = h_t - t\hat{v}_\theta. \tag{17}$$

Because $\hat{h}_0 - h_0 = -t(\hat{v}_\theta - v^\star)$, the velocity and clean-latent terms can be written compactly as

$$\mathcal{L}_{\mathrm{train}} = \mathbb{E}_{(c,x),t,\epsilon}\left[w(t)\|\hat{v}_\theta - v^\star\|_2^2\right], \qquad w(t) = 1 + \lambda_{\mathrm{rec}} t^2. \tag{18}$$

This reconstruction-oriented weighting emphasizes clean-latent recovery from compression-conditioned inputs. We optimize the generative branch and its adapters while keeping the VAE, vision encoder, and codec fixed, and freeze the complete enhancement model after training.

Reconstruction-aware spatial composition. Since generation does not improve every region equally, its regional contribution is determined from observed reconstruction behavior. After spatial budgeting selects the scheme, we use its converged provisional reconstruction $\tilde{z} = G_{\bar{\theta}}^{(T_{\mathrm{P}})}(c)$ to optimize continuous surrogate composition weights.

We assign each region Ω_i a mixing weight $\alpha_i \in [0, 1]$ and select the weights at the encoder through

$$\boldsymbol{\alpha}^\star \in \arg\min_{\boldsymbol{\alpha}\in[0,1]^N} \frac{1}{N} \sum_{i=1}^{N} \ell\left(x_i,\, (1-\alpha_i)c_i + \alpha_i \tilde{z}_i\right), \tag{19}$$

where x_i, c_i, and $\tilde{z}_i$ denote the corresponding regional crops and ℓ is the reconstruction criterion. The codec and generative reconstructions remain fixed during selection, requiring no additional generative sampling. The selected weights are quantized and spatially smoothed before transmission.

At the decoder, the quantized regional weights are expanded into a spatial field with smooth boundary transitions.

$$a = \mathcal{F}(\mathcal{Q}(\boldsymbol{\alpha}^\star)), \qquad \hat{x} = (1-a) \odot c + a \odot z. \tag{20}$$

Here, $\mathcal{Q}$ denotes weight quantization and $\mathcal{F}$ denotes boundary-aware feathering. The decoder generates the full-step reconstruction $z = G_{\bar{\theta}}^{(T_{\mathrm{f}})}(c)$ and applies the transmitted weights to it as in Eq. 20. The composition retains reliable codec content while introducing generated details according to regional reconstruction feedback. Weight selection is performed at the encoder, while the decoder only applies the transmitted weights without accessing the original image.

4 Experiments

4.1 Implementation

We use the pretrained DLF (Xue et al., 2025) as the base codec and keep its parameters fixed. The generative enhancer is initialized from BAGEL (Deng et al., 2025) and fine-tuned on paired original and DLF-reconstructed images from a separate training corpus that is image-disjoint from both evaluation sets. During fine-tuning, the VAE and understanding branch remain frozen, while the generation branch and generation-side latent adapters are optimized. During inference, spatial budgeting searches over fixed codec tiers using reduced-step provisional reconstructions, followed by full-step generation and spatial composition. The reported bitrate adds the image representation and fixed-length protocol costs for scheme and composition side information. We evaluate our method on 41 images from CLIC 2020 (Toderici et al., 2020) and the 100-image DIV2K validation set (Agustsson & Timofte, 2017). Training, search, sampling, rate accounting, and computational details are provided in Appendix A.1.

4.2 Quantitative Comparison

We compare against traditional codecs BPG (Bellard, 2014) and VTM (Bross et al., 2021); learned codecs ELIC (He et al., 2022), HiFiC (Mentzer et al., 2020), MS-ILLM (Muckley et al., 2023), MRIC (Agustsson et al., 2023), and CRDR (Iwai et al., 2024); and generative methods PerCo (Careil et al., 2024), PICS (Lei et al., 2023), CGDM (Kuang et al., 2024), CGDM+ (Kuang et al., 2026), GLC-Image (Qi et al., 2025), and DLF (Xue et al., 2025), using public pretrained models for all learned baselines.

We evaluate fidelity with PSNR and MS-SSIM (Wang et al., 2003), perceptual similarity with DISTS (Ding et al., 2022), TOPIQ-FR (Chen et al., 2024), and MS-SWD (He et al., 2024), and distributional realism with FID (Heusel et al., 2017). Figure 3 presents rate–quality results on both datasets, with enlarged low-bitrate regions; additional metrics are reported in Appendix A.3.

4.3 Qualitative Comparison

Figure 4 compares representative reconstructions from CLIC 2020. Although VTM, ELIC, and PerCo use substantially higher bitrates, their reconstructions blur or alter the facial and fur textures highlighted in red. DLF provides a stronger low-rate reconstruction, while our method further preserves local contours and fine fur structure at a slightly lower bitrate, yielding the lowest DISTS among the compared methods in this example.

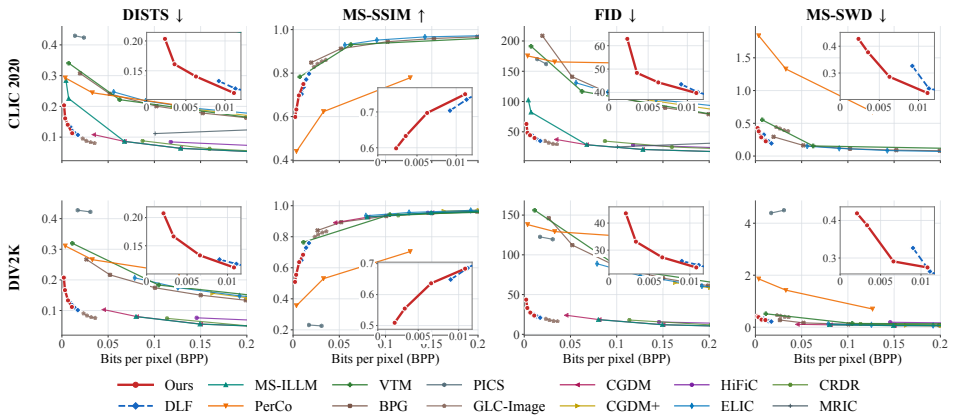

Figure 3: Rate-quality comparisons on CLIC 2020 and DIV2K. Upward and downward arrows indicate whether higher or lower values are preferred.

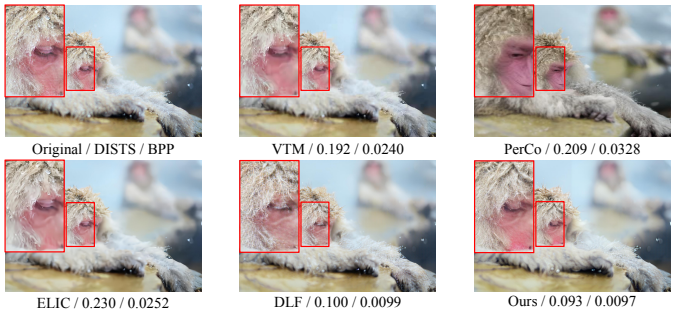

Figure 4: Qualitative comparison on CLIC 2020. Red boxes highlight local texture differences.

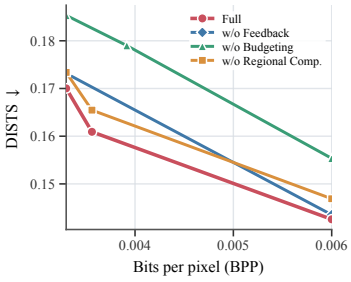

Figure 5: Ablation results on CLIC 2020.

4.4 Ablation Studies

We evaluate the three principal components of our method using DISTS over their common low-rate interval. The complete rate-quality curves are provided in the supplementary material.

Effectiveness of Feedback-Guided Scheme Revision. The w/o Feedback variant determines the spatial scheme from perceptual analysis but omits iterative revision using reconstruction feedback. As shown in Fig. 5, the full model achieves lower DISTS throughout the displayed range, showing that observed decoder behavior provides useful information beyond the initial scheme.

Effectiveness of Spatial Budgeting. The w/o Spatial Budgeting variant applies a uniform compression tier to the entire image instead of adapting the scheme across regions. Its consistently higher DISTS indicates that regional budgeting better matches transmitted information to spatially varying restoration behavior under the reported settings.

Effectiveness of Regional Composition. The w/o Regional Composition variant replaces regional mixing weights with a single image-level weight. Regional composition produces lower DISTS by preserving reliable codec content while using generation where it improves reconstruction.

5 Conclusion

We present Thinking-then-Compress, a closed-loop image compression paradigm that adapts compression decisions to the observed reconstruction behavior of a fixed generative decoder. Perception-Informed Spatial Budgeting combines perceptual importance with reconstruction feedback to iteratively determine a regional compression scheme, while Reconstruction-Aware Generative Enhancement adapts generation to compressed inputs and composes codec and generative outputs according to regional reconstruction quality. The resulting scheme search remains at the encoder and does not require joint optimization with the generative decoder or additional optimization at the decoder. Experiments demonstrate the superity of our proposed methods.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, provide feedback on research methodology and experimental design, and support the interpretation of experimental results. We also used these tools for language polishing, organizing literature and reproducibility-related details, preparing preliminary figure layouts, and generating or editing parts of the experimental code. The AI-assisted code was reviewed and tested by the authors, and all reported experiments were executed and verified by the authors. All other AI-assisted outputs were also reviewed and verified by the authors. We take responsibility for the final content of this work, including its text, claims, code, results, and artifacts.

REPRODUCIBILITY STATEMENT

The method formulation and optimization objectives are described in Section 3, and the datasets, comparison methods, evaluation metrics, and principal experimental settings are provided in Section 4. Appendix A.1 reports the training, spatial scheme, search, sampling, composition, and rate-accounting configurations. Appendix A.2 provides the complete spatial budgeting procedure and its computational accounting, Appendix A.3 contains additional quantitative results, and Appendix A.4 reports supplementary ablations. Our project will be released upon publication.

REFERENCES

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 Workshops*, pp. 126–135, 2017.

Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 221–231, 2019.

Eirikur Agustsson, David Minnen, George Toderici, and Fabian Mentzer. Multi-realism image compression with a conditional generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22324–22333, 2023.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

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

Fabrice Bellard. BPG image format. `https://bellard.org/bpg/`, 2014.

Jonas Brenig and Radu Timofte. Diffusion-based compression quality tradeoffs without retraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 5620–5629, 2025.

Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. doi: 10.1109/TCSVT.2021.3101953.

Joaquim Campos, Simon Meierhans, Abdelaziz Djelouah, and Christopher Schroers. Content adaptive optimization for neural image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Marlène Careil, Matthew J. Muckley, Jakob Verbeek, and Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=ktdETU9JBg`.

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

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

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

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

Jiaqi He, Zhihua Wang, Leon Wang, Tsein-I Liu, Yuming Fang, Qilin Sun, and Kede Ma. Multiscale sliced Wasserstein distances as perceptual color difference measures. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024. URL https://arxiv.org/abs/2407.10181.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Shoma Iwai, Tomo Miyazaki, and Shinichiro Omachi. Controlling rate, distortion, and realism: Towards a single comprehensive neural image compression model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2888–2897, 2024. doi: 10.1109/WACV57701.2024.00288.

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*, pp. 26088–26098, 2024.

Wei Jiang, Wei Wang, Songnan Li, and Shan Liu. Online meta adaptation for variable-rate learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 498–506, 2022.

Haowei Kuang, Yiyang Ma, Wenhan Yang, Zongming Guo, and Jiaying Liu. Consistency guided diffusion model with neural syntax for perceptual image compression. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1622–1631, 2024. doi: 10.1145/3664647.3681336.

Haowei Kuang, Wenhan Yang, Zongming Guo, and Jiaying Liu. Cross-granularity online optimization with masked compensated information for learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16514–16523, 2025.

Haowei Kuang, Wenhan Yang, Zongming Guo, and Jiaying Liu. Syntax-driven multi-realism image compression with consistency guided diffusion model. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(5):6211–6224, 2026. doi: 10.1109/TCSVT.2026.3651547.

Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti. Text + sketch: Image compression at ultra low rates. In *ICML Workshop on Neural Compression*, 2023. URL https://arxiv.org/abs/2307.01944.

Fabian Mentzer, George D. Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11913–11924, 2020.

David Minnen, Johannes Ballé, and George D. Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10794–10803, 2018.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4296–4304, 2024.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25426–25443. PMLR, 2023.

Linfeng Qi, Zhaoyang Jia, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Generative latent coding for ultra-low bitrate image and video compression. *arXiv preprint arXiv:2505.16177*, 2025.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

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*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494, 2022.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 4:142–163, 1959.

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*, pp. 32936–32946, 2026.

George Toderici, Wenzhe Shi, Radu Timofte, Lucas Theis, Johannes Ballé, Eirikur Agustsson, Nick Johnston, and Fabian Mentzer. Workshop and challenge on learned image compression. CVPR Workshop, 2020. URL http://www.compression.cc.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12977, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

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*, pp. 18051–18061, 2025.

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

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, 2025.

A Appendix

A.1 Implementation Details

Enhancement training. We train the BAGEL enhancement model on paired original and DLF-reconstructed images for 2,000 steps using AdamW with a learning rate of 2×10^{-4}. We use the EMA checkpoint at step 2,000, obtained from training on six A800 GPUs. The velocity and clean-latent losses have unit weights, while the fixed instruction contributes no language-token loss. We optimize the MoT generation branch and the VAE-to-language, language-to-VAE, latent-position, and time adapters, while keeping the VAE, ViT, understanding branch, and remaining parameters frozen. The compressed condition is injected through both the VAE and ViT pathways, and the original image provides the VAE target.

Spatial scheme and perceptual analysis. Spatial budgeting uses 512×512 non-overlapping regions and 13 compression tiers derived from DLF without retraining the codec. Writing each tier as (q, K_{cb}, d) for the DLF quality index, effective codebook size, and spatial downsampling factor, respectively, the tier set is L0p $= (3, 4096, 4)$, L1b $= (0, 2048, 2)$, L1–L4 $= (0$–$3, 4096, 2)$, L5–L8 $= (0, 512$–$4096, 1)$, and L9–L11 $= (1$–$3, 4096, 1)$. The BAGEL understanding branch is queried once per input at the encoder before scheme search using a fixed prompt, and the resulting semantic boxes b_m with normalized relevance scores r_m are cached for that input. Regional importance is computed as

$$I_i = \max\left(\{r_m \mid \mathrm{center}(b_m) \in \Omega_i\} \cup \{0\}\right). \tag{21}$$

For a region assigned tier k, the provisional composition field uses an anchor a_k calibrated once on a held-out set and fixed for all test images. The anchors are expanded using the same boundary smoothing rule as final composition. Images are padded to a multiple of 512 pixels for regional coding and cropped back to their original dimensions after reconstruction. The four nominal operating points initialize the uniform dominant tier with L0p, L1, L4, and L8, respectively, before regional refinement over the 13 fixed DLF tiers.

Understanding prompts. The understanding branch uses a three-stage prompting procedure that first summarizes the scene, then identifies compression-sensitive regions, and finally localizes them with normalized bounding boxes. We use the following system prompt.

> You should first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process is enclosed within thinking response tags.

The three stage-specific prompts are as follows.

> **Scene understanding.** Look at this image carefully. Reason step by step about the whole scene, including the main subjects, the objects, people, animals, or text present, and the overall spatial layout. Then finish with a short scene summary of at most two sentences, prefixed with exactly `SCENE:` on its own line. Do not output any coordinates yet.

> **Region identification.** Based on the scene summary, identify the 1–4 most important local regions that would suffer most from heavy compression and must be preserved, including faces, text, fine textures or fabrics, sharp edges, and salient objects. Output exactly one line per region in the format `Region k: <short semantic label>`. Do not output coordinates at this stage.

> **Region localization.** For each region listed above, provide its bounding box within the full image. Coordinates are (x_1, y_1, x_2, y_2) normalized to $[0, 1]$. Output exactly one line per region in the format `Region k: bbox=[x1,y1,x2,y2]`. Each box must tightly enclose the corresponding image region and have positive width and height, with each dimension at least 0.03. Different regions must not share identical boxes. If a region is not spatially localizable, output a reasonable box. Output only the region bounding-box lines without additional reasoning or description.

Understanding decoding. Scene and region responses are sampled for at most three attempts with temperatures 0.8, 1.1, and 0.6, while bounding boxes are decoded greedily. Parsing retains at most four valid boxes, clamps coordinates to $[0, 1]$, and discards boxes with normalized area below 0.01.

Search configuration. We set $\eta_{\mathrm{c}} = 0.8$, $\eta_{\mathrm{g}} = 0.2$, and $\eta_{\mathrm{I}} = 1.0$ in the proposal score, with $\tau = 0$ and $\tau_{\mathrm{screen}} = 0.001$. Each round proposes at most three upgrades, two downgrades, two paired exchanges, and one exploration candidate, giving at most eight candidate schemes containing at most ten atomic tier changes. In the reported Lagrangian runs, an upgrade or downgrade moves a region by up to two tiers, a paired exchange couples one upgrade with one downgrade, and exploration tests an unvisited important region. The default configuration permits three provisional generative calls in total, including the initial reconstruction of the starting scheme, and terminates when this limit is reached, no candidate survives, or two consecutive proposals are rejected after exploration. Experiments use Lagrangian rate control with $\lambda \in \{2 \times 10^{-5}, 5 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-7}\}$ for the four nominal operating points, respectively, and report the achieved rate. The implementation also applies a loose safety cap of four times the nominal rate budget.

Sampling and composition. Provisional and final reconstructions use 8 and 80 sampling steps, respectively. All reported results use this single deterministic sampling configuration. Generation uses 16-pixel tile overlap and a fixed restoration instruction for all images. The same instruction is used during enhancement training, provisional reconstruction, and final decoding.

> Restore the original image from this heavily compressed reconstruction. Keep the content and layout unchanged and recover faithful colors, textures, edges and details without adding artifacts.

Composition weights are optimized against the converged 8-step reconstruction and then applied to the 80-step reconstruction produced for final decoding. This optimization is a continuous surrogate evaluated before 6-bit quantization and boundary expansion, so its solution need not minimize the final transmitted-weight objective exactly. Both $\mathcal{D}$ in spatial budgeting and ℓ in composition planning are implemented with LPIPS (Zhang et al., 2018). The regional weights are optimized from the provisional reconstruction using Adam for 200 steps with a learning rate of 0.05, quantized to 6 bits, and expanded with 16-pixel linear feathering. The regional LPIPS objective excludes a 16-pixel boundary margin before feathering is applied.

Rate accounting. For an $n_h \times n_w$ grid with $N = n_h n_w$ regions and M rectangles, the reported side rate contains a tier map, rectangle descriptors, a 512-bit header, and $6N$ composition-weight bits. The tier-map cost is the smaller of a fixed 4-bit-per-region representation and run-length coding, while the rectangle cost is $M\,(4 + 4\lceil \log_2 \max(n_h, n_w) \rceil)$ bits.

Computational accounting. Encoding additionally invokes the BAGEL understanding branch once per input before search. With the provisional-call budget $K = 3$, scheme search uses $KT_{\mathrm{p}} = 24$ generative sampling steps, together with a conditional T_{p}-step refresh when joint verification retains only a subset of its proposed updates. The decoder uses one $T_{\mathrm{f}} = 80$ step call and spatial composition. Unlike latent online optimization, the scheme search requires no backward pass through the generative model, but it does require repeated forward sampling. Composition planning differentiates only with respect to the regional weights and does not backpropagate through the generative model.

Supplementary implementation. The anonymous code package contains the encoder-side search, understanding-region extraction, composition planning, fixed tier and rate configurations, paired-data interface, and an executable command template. BAGEL and DLF source code, pretrained weights, datasets, and generated outputs are not redistributed and must be obtained under their original licenses.

A.2 Spatial Budgeting Procedure

Given an image x, a nominal rate point and its λ, precomputed regional importance scores, and the fixed generative decoder, spatial budgeting proceeds as follows.

For a candidate s', define its rate change as

$$\Delta R(s' \mid s) = R_y(s'; x) + R_s(s') - R_y(s; x) - R_s(s). \tag{22}$$

The screening efficiency used to rank surviving candidates is

$$e_{\mathrm{screen}}(s' \mid s) = \begin{cases} +\infty, & \Delta R(s' \mid s) \le 0 \text{ and } \widehat{\Delta J}(s' \mid s) > 0, \\ \widehat{\Delta J}(s' \mid s)/\Delta R(s' \mid s), & \Delta R(s' \mid s) > 0 \text{ and } \widehat{\Delta J}(s' \mid s) > 0, \\ -\infty, & \text{otherwise.} \end{cases} \tag{23}$$

Candidates are ranked by higher efficiency, then smaller rate increase, and finally larger screened gain.

1. Initialize all regions with the dominant tier associated with the target rate, obtain its codec reconstruction, and evaluate its provisional reconstruction with fixed sampling randomness.
2. At each round, compute the regional proposal scores in Eq. 12, construct upgrade, downgrade, paired exchange, and exploration candidates, and remove duplicate or previously visited schemes.
3. Apply codec-only screening with Eq. 13, reject infeasible and strongly negative candidates, and rank the remaining candidates using Eq. 23.
4. Merge compatible changes from the top-ranked candidates, retain only the highest-ranked change when candidates conflict, remove the least efficient rate-increasing changes until the safety cap is met, and evaluate the joint scheme with one provisional generative call.
5. Measure the contextual regional gains in Eq. 14, retain rate-increasing changes whose gain per bit exceeds λ, and retain non-increasing changes with positive regional gain.
6. If only a subset of the joint proposal is retained, recompute the provisional reconstruction for the retained scheme before further refinement or composition planning.
7. Stop when the provisional-call budget is exhausted, no candidate remains, or two consecutive proposals are rejected after exploration.

The implementation first computes one provisional reconstruction of the initial scheme and counts it toward K. It then performs at most $K - 1$ verified revision calls, while codec-only screening requires no generative call. After search, composition planning reuses the converged provisional reconstruction, and final decoding performs one T_{f}-step reconstruction. When joint verification retains only a subset of the proposed regional changes, one additional provisional call refreshes the reconstruction for the retained scheme before it is used by subsequent search or composition planning.

A.3 Additional Quantitative Results

Figure 6 reports complementary comparisons using PSNR, TOPIQ-FR, LPIPS, and NIQE on CLIC 2020 and DIV2K. Full-reference metric points are averaged over all images in each dataset, while FID is computed at the dataset level. Because FID is estimated from 41 CLIC and 100 DIV2K images, we treat it as a descriptive dataset-level indicator rather than statistically conclusive evidence.

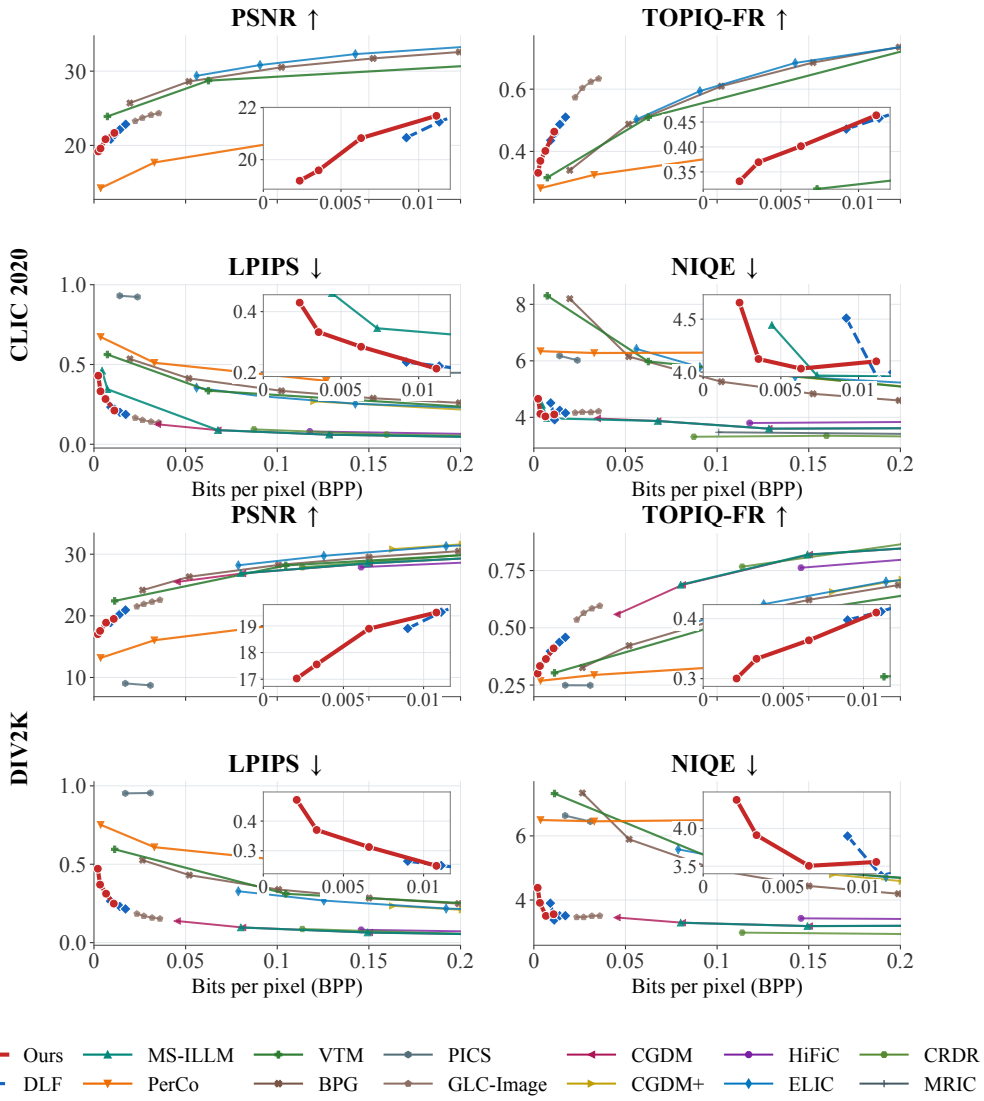

Figure 6: Additional rate-quality comparisons on CLIC 2020 and DIV2K. Each inset magnifies the low-rate region containing the complete curve of our method. Upward and downward arrows indicate whether higher or lower values are preferred.

A.4 Additional Ablation Studies

We further refine the first two main ablations on the 41 CLIC 2020 test images at four operating points. All supplementary curves use DISTS against the actual achieved bitrate, avoiding direct comparison between unmatched operating points.

Search budget and verification. Figure 7(a) and (b) vary the number of provisional generative calls from zero to the default budget of three. Zero and one call produce similar curves, while two calls recover most of the improvement obtained with the full budget, indicating that the main benefit

appears after reconstruction feedback has been refreshed more than once. For joint verification, we denote the number of retained regional changes by K_J to distinguish it from the provisional-call budget K. Figure 7(b) shows that reducing K_J from nine to one weakens perceptual quality at comparable rates.

Dominant-layer construction within spatial budgeting. The second main ablation removes spatially selective budgeting altogether. We further examine the construction used by the full spatial-budgeting scheme. In the default construction, the dominant layer is encoded for the image before higher-tier regional updates are applied. The real-dominant-drop variant re-encodes the dominant layer after removing regions represented by higher-tier rectangles. The alternative re-encodes only the remaining regions after removing areas represented by higher-tier rectangles. Figure 8 shows a moderate low-rate degradation under this stricter construction, with a smaller difference at the remaining operating points.

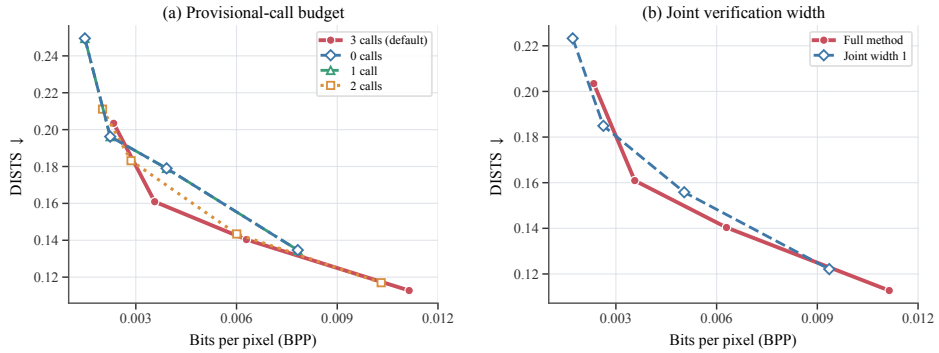

Figure 7: Supplementary feedback-search ablations on CLIC 2020 using 41 images and four operating points. The curves report DISTS against the actual achieved bitrate. Panel (a) varies the provisional-call budget, while panel (b) varies the joint verification width.

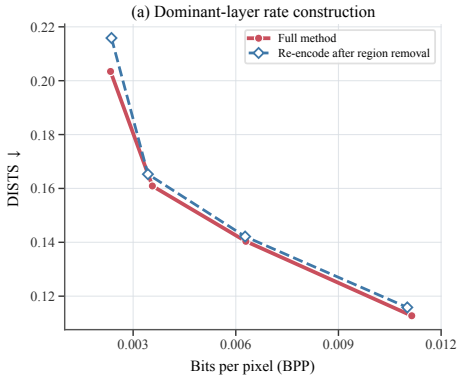

Figure 8: Supplementary refinement of the spatial-budgeting ablation on CLIC 2020 using 41 images and four operating points. The default method encodes the dominant layer before applying higher-tier regional updates, whereas the alternative re-encodes only the remaining regions after higher-tier regions are removed. The curves report DISTS against the actual achieved bitrate.

A.5 Additional Qualitative Results

Figure 9 provides additional low-bitrate comparisons covering facial details, human figures, and natural textures. The bitrate used by each reconstruction is reported below the corresponding image.

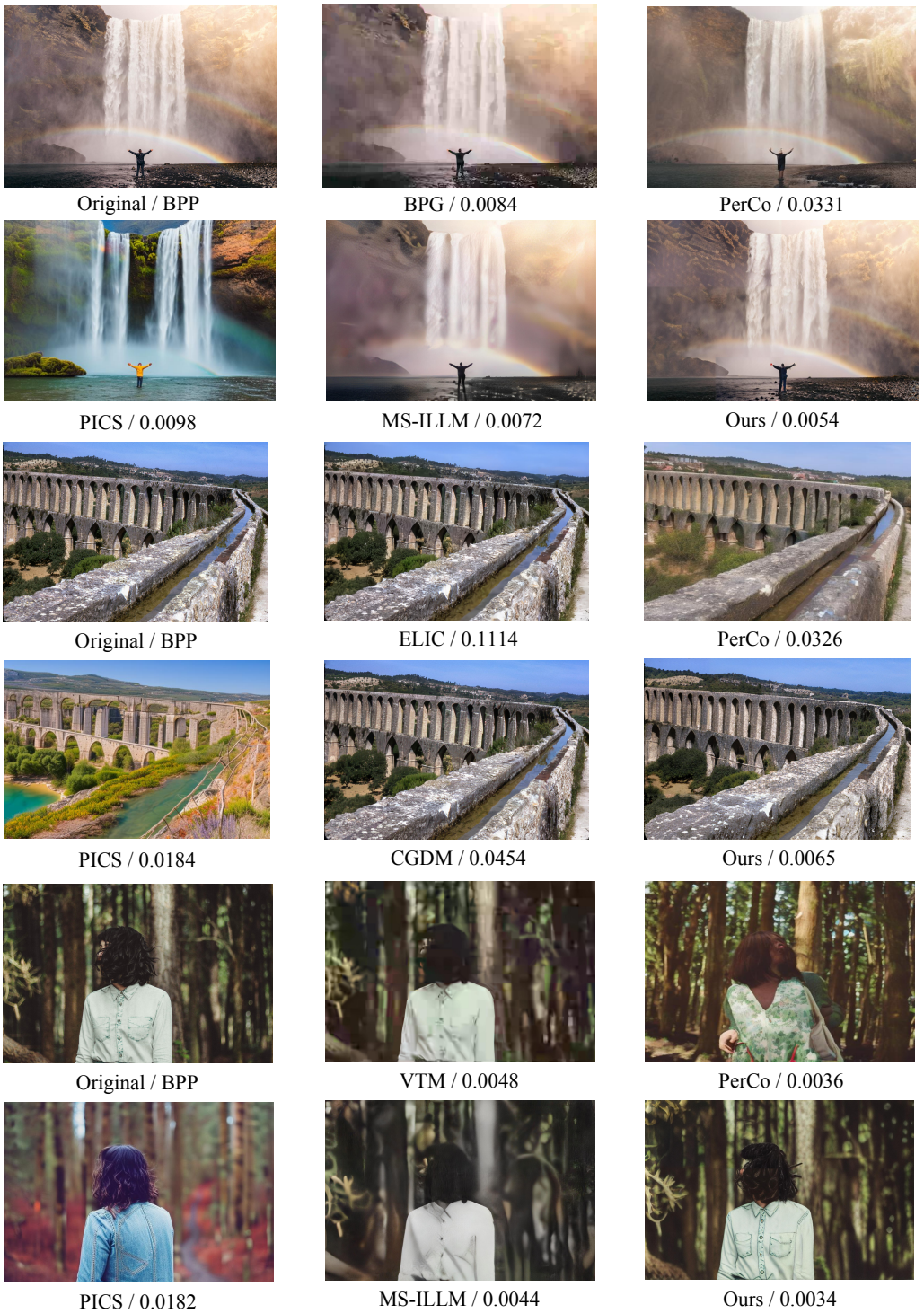

Figure 9: Additional qualitative comparisons. Zoom in for best view.