CONTINUOUS DETAIL-EXPOSURE CONTROL VIA LATENT SKIPPING FOR VIEW-LEVEL PRIVACY IN LEARNED IMAGE COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned image compression typically optimizes rate–distortion performance, but offers limited control over how much visual detail is exposed at decoding time. We study controllable detail exposure: suppressing fine-grained texture while preserving a structurally interpretable view, thereby limiting what downstream tasks can extract from the decoded view, a guarantee we refer to as view-level privacy. We propose sCBH (s-Contract with Blockhead), a channel-level skipping framework trained without any downstream-task objective, in which a single scalar s (the detail-exposure strength) determines how many latent channels are retained in the bitstream. At $s = 1$, sCBH is bit-exact to the base codec, adding zero quality tax; decreasing s progressively removes texture channels from the bitstream and reduces bitrate. A lightweight structural decoder reconstructs a low-detail representation from the structural channels, preserving coarse scene structure under aggressive skipping. Skipping and stuffing strategies are adapted to the causal order of each entropy model. Experiments on ELIC, Hyperprior, CCM, and HPCM show continuous control over bitrate and detail; across four downstream tasks, performance degrades according to each task's reliance on texture versus structure. The source code is publicly available on our anonymous repository.

1 INTRODUCTION

Learned image compression has advanced rapidly, with methods such as ELIC (He et al., 2022) and HPCM (Li et al., 2025b) approaching or surpassing traditional codecs in rate-distortion (RD) performance. However, these frameworks are designed solely for compression efficiency and provide no explicit control over how much visual detail is exposed at decoding time. In applications such as image sharing, facial photo storage, and document scanning, anyone with access to the decoded view can exploit fine-grained details even when only the coarse structure is needed. As we show, the same frozen codec can reduce the face detection rate from 99.25% to 0% while remaining bit-exact at full fidelity, simply by controlling how much detail is transmitted.

Existing approaches mainly fall into three categories with fundamental limitations. Global methods uniformly degrade the image, destroying information related to the tasks while often remaining partially invertible and permanently forfeiting full quality. Regional methods depend on external detectors, introducing additional computation and failure modes. Learning-based methods embed downstream task objectives in the loss function and limit generality to unseen scenarios.

We propose sCBH (s-Contract with Blockhead), a channel-level selective skipping framework that provides continuous control over detail exposure. Our key insight is that a frozen pre-trained compressor's latent space can be split into structural and texture channels, and that texture channels can be progressively removed without affecting the structural representation. sCBH provides three contributions:

1. A single scalar $s \in [0, 1]$ continuously controls the detail-exposure strength, quality, and bitrate, achieving **continuous detail-exposure control with zero additional tax at full fidelity**. When $s = 1$, sCBH is bit-exact to the base compressor (e.g., ELIC). As s decreases, a hard 32/288 structural/texture split is applied to the latent representation. Masked channels are physically excluded from entropy coding, with the exact implementation respecting

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.

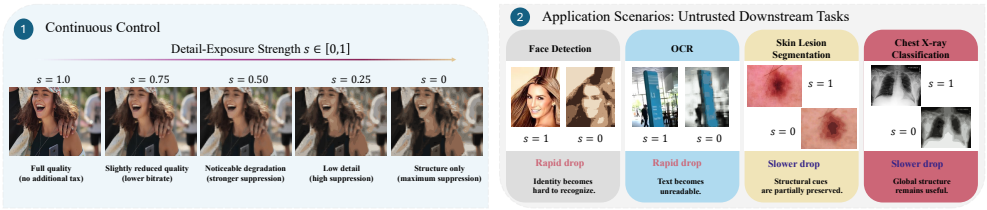


Figure 1: Continuous detail-exposure control via a single scalar s. Our method achieves bit-exact reconstruction at full quality and gradually suppresses fine-grained detail. High-frequency tasks suffer rapid performance drops, while structural medical tasks remain partially functional, demonstrating task-agnostic control of downstream task performance.

the causal order of the context model, achieving architecture-level channel splitting with variable-length skipping. Unlike pre-processing-based degradation, the removal happens inside the codec, costs exactly zero bits, and preserves the full-quality mode exactly.

2. The loss function and the Blockhead, a CNN net, are specially designed, achieving **task-agnostic generalization**: the training loss contains no downstream task terms, yet downstream performance degrades in a controlled way across face recognition, OCR, medical image segmentation, and chest X-ray classification. Unlike adversarial or task-guided methods that embed specific objectives, sCBH draws inspiration from the self-supervised paradigm.
3. A lightweight Blockhead decoder is attached to the frozen backbone without modifying its parameters. We validate sCBH on four architectures and five variants: ELIC (He et al., 2022) with/without AuxT (Li et al., 2025a) (SCCTX), Hyperprior (Ballé et al., 2018) (sigma-only), CCM (He et al., 2021) (checkerboard) and HPCM (Li et al., 2025b) (progressive context), confirming that **the approach generalizes across entropy model designs**.

Scope and definition. We define view-level privacy under the threat model in which downstream tasks access only the decoded view, not the latent representation, and we evaluate against standard, non-adaptive downstream models. sCBH targets scenarios where the concern is unauthorized exploitation of the decoded view rather than an untrusted decoder. Due to the frozen backbone design, the structural channels still contain identity information; this does not compromise our threat model, since downstream tasks cannot access the latent. We do not defend against adversaries that adapt to the mechanism or access the latent. We make no cryptographic or bitstream-level privacy claim.

2 Related Work

2.1 Learned Image Compression

Learned image compression has evolved from hyperprior-based models (Ballé et al., 2018; Minnen et al., 2018) to context-adaptive architectures. CCM (He et al., 2021) introduces mask convolution for spatial context modeling. ELIC (He et al., 2022) achieves effective compression with spatially adaptive context (SCCTX), with AuxT (Li et al., 2025a) introducing auxiliary transforms. HPCM (Li et al., 2025b) achieves state-of-the-art performance by employing multi-scale progressive context for hierarchical entropy coding. These methods optimize rate-distortion trade-offs but do not consider privacy. sCBH can be applied to any of the architectures mentioned above with the skipping strategy adapted to the causal order of each context model.

2.2 Privacy-Preserving Compression

Existing privacy-aware compression methods differ along three axes: the granularity at which detail exposure is controlled (binary vs. continuous), the external dependencies they introduce (detectors, task labels, discriminators, or keys), and the level at which privacy is defined (view-level vs. bitstream-level). Table 1 compares representative methods along these axes.

ROI-based anonymization detects and perturbs sensitive regions, e.g., by injecting a region-of-interest loss so that faces become unrecognizable while the surrounding scene remains in-

tact (Liebender et al., 2024). It requires an external detector, adding computation and failure modes, and provides only binary control. Task-guided information bottlenecks organize the latent space according to the mutual information of specific downstream tasks, trading reconstruction quality against task accuracy (Alvar & Bajić, 2021); privacy is thus defined by the tasks seen during training, and changing the task requires retraining.

Adversarial feature coding trains an autoencoder against a task network to prevent accurate reconstruction while preserving machine vision accuracy (Azizian & Bajić, 2022), but the privacy objective is limited to the tasks represented in the loss and training can be unstable. Conditional bitstream-level decoding embeds a trigger so that a single bitstream decodes either a protected or an authorized view (Shen et al., 2025); control is binary, and the design targets vision-language pretrained models rather than general visual data. Information-theoretic methods with keys combine the information bottleneck with shared-secret management (Razeghi et al., 2022), at the cost of key distribution infrastructure and variational inference assumptions.

In contrast, sCBH controls the detail-exposure strength with a single scalar s and introduces no external dependency: no detector, no task label, no discriminator, and no key. Its mechanism is architectural: a hard 32/288 structural/texture split of the latent, in which skipped texture channels are physically excluded from entropy coding following the causal order of the context model and thus incur exactly zero bitrate. At $s = 1$ the method is bit-exact to the base compressor, and privacy is defined at the view level, where unauthorized downstream tasks can only access the decoded view rather than the latent. We validate sCBH on ELIC, Hyperprior, CCM, and HPCM, where downstream performance degrades according to each task's reliance on texture versus structure.

Table 1: Comparison of Privacy-Preserving Compression Methods

Method	Mechanism	Control granularity	External dependency	Privacy level
sCBH (ours)	Channel skip	**Continuous**	**None**	View-level (variable bitstream)
ROI anon. (Liebender et al., 2024)	ROI loss	Binary	Detector	View-level
Scalable(Alvar & Bajić, 2021)	MI bottleneck	Task-level	Task labels	View-level
Adversarial (Azizian & Bajić, 2022)	Adv. training	Binary	Discriminator	View-level
PSIC (Shen et al., 2025)	Cond. decode	Binary	None (VLP)	Bitstream-level
IB+Key (Razeghi et al., 2022)	Theory+Key	Continuous†	Key mgmt	Bitstream-level

†Continuous in theory, but requires key infrastructure.

2.3 Channel Selection in Compression

Channel pruning has been studied for computational acceleration (He et al., 2017). Channel selection is inherently shaped by the causal order of the context model. SCCTX enforces a spatial-channel priority ordering, driving the model to concentrate information in the earliest channels, while channels of Checkerboard treat neighboring positions symmetrically, resulting in the uniform channel importance distribution. HPCM extends this by adopting multi-scale progressive contexts. Channels are organized in groups with varying quantization precisions across scales. sCBH preserves the channel priorities learned by the base models, applying skipping strategies adapted to each causal order. This ensures that skipping does not break the autoregressive conditioning, while simultaneously achieving exact zero bitrate for dropped channels.

3 Method

3.1 Architecture Overview

Fig. 2 illustrates the sCBH framework. The encoder g_a produces y with $M = 320$ channels. 32 structural channels and 288 texture channels are split. Then only k_t texture channels and structural channels are concatenated and passed to the entropy model. Other texture channels are physically excluded from the bitstream. At the decoder, the entropy model reconstructs n_{keep} channels, and the channels skipped at the encoder are filled according to a convention determined by the causal order of the context model, on which both encoder and decoder agree. The Blockhead decoder, which is a lightweight module compared with the backbone decoder g_s, reconstructs the color-block image x_{rough} from structural channels and the K-means clustering conducted at the inference stage only improves the quality of vision, generating I_s. The Blockhead is trained on the training data pre-

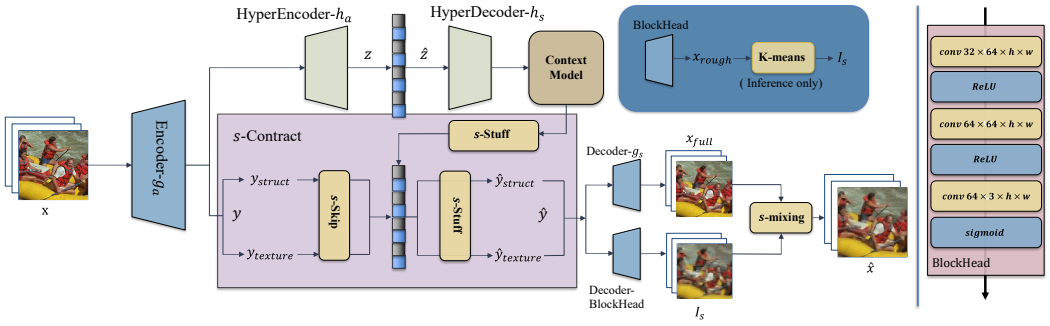


Figure 2: sCBH architecture overview. The input image x is encoded into the latent space and hard-split into structural and texture channels. $k_t = \text{round}(s \cdot 288)$ texture channels are retained and the rest are physically absent from the bitstream. The entropy model encodes only $n_{keep} = 32 + k_t$ channels. At the decoder side, the skipped channels are filled according to the value of s. The skipping and stuffing strategies are designed according to the causal order of the context model. The full 320-channel latent is passed to the base decoder to produce x_{full}. The structural channels are fed into the Blockhead decoder to produce a structural color-block image x_{rough}. Non-differentiable K-means is applied at inference only to generate I_s for better representation and effective detail suppression. The final reconstruction is obtained via s-mixing, ensuring bit-exact reconstruction at $s = 1$ and pure color-block image at $s = 0$. The exact number of blocks adapts according to the size of x and the value of s. The baseline backbone is frozen. The Blockhead decoder is trained on the training data to learn a feature representation that facilitates effective K-means clustering at inference.

processed by K-means with a dedicated loss function to learn a feature representation that facilitates effective K-means clustering.

3.2 Causal-Order-Aware Skipping and Stuffing

At $s < 1$, some texture channels are not needed at the decoder. Encoding them would waste bitrate, so the encoder skips them and the decoder stuffs the skipped positions. Because context models impose different causal structures on the channel dimension, the strategies must follow the causal order to keep autoregressive conditioning intact. We identify three cases, as illustrated in Fig. 3.

ELIC (SCCTX). SCCTX enforces a spatial-channel priority ordering that concentrates information in the earliest channels, so texture channels reside at the tail. Skipping is simple truncation: only the first $32 + k_t$ channels are encoded, and the decoder zero-fills the skipped positions.

Hyperprior & CCM (Checkerboard). Hyperprior has no context model, and the checkerboard context treats neighboring positions symmetrically; both exhibit a uniform channel importance distribution without an inherent ordering. Channels are sorted by their variance on the training data, and the encoder truncates after sorting while the decoder applies the inverse permutation before zero-filling.

HPCM. HPCM codes progressively across three resolution levels (2, 3, and 6 steps), quantizing all 320 channels at each step against context-predicted means and scales. We therefore skip at the step level: for a given s, the first $2 + n_{\text{tex}}$ steps (with $n_{\text{tex}} = \text{round}(s \cdot 9)$) encode all 320 channels, while later steps encode only the 32 structural channels, with texture positions filled by the context model's predicted means—a better choice than zero-filling given the cross-attention and residual prediction in the context model. The two Scale-1 steps always code fully, and truncated steps are placed last in coding order, which guarantees rate monotonicity.

3.3 The s-Contract

We consider a learned image compressor with encoder g_a, decoder g_s, a hyperprior bottleneck, entropy model with quantizer, and context model. Given an input image $x \in \mathbb{R}^{3\times H\times W}$, the encoder

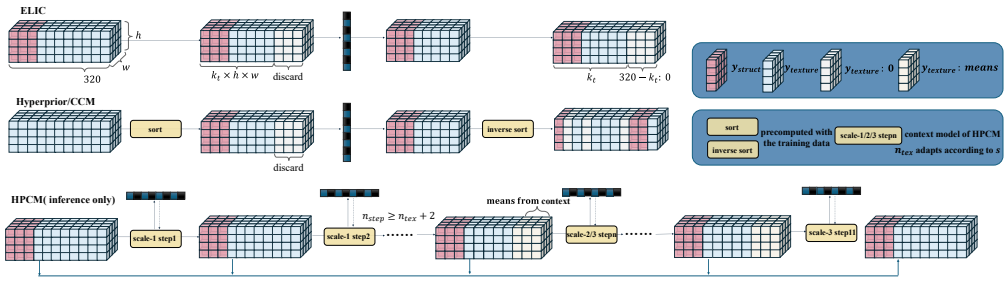


Figure 3: Skipping and stuffing follow the causal order of the context model: truncation for SCCTX, variance-based sorting with an inverse permutation for Hyperprior and CCM, and step-level skipping with predicted-mean stuffing for HPCM. During training we follow the ELIC-style truncation strategy to keep the optimization identical.

produces a latent tensor $y = g_a(x) \in \mathbb{R}^{M \times h \times w}$, where M is the number of latent channels and h, w are the downsampled spatial dimensions from H, W. We define a detail-exposure strength scalar $s \in [0, 1]$, establishing the following contract between the encoder and decoder:

$$k_t = round(288 \cdot s), \quad n_{keep} = 32 + k_t \tag{1}$$

where k_t is the number of retained texture channels and n_{keep} is the total number of channels passed to the entropy model. Ignored texture channels are filled by the decoder according to the value of s. The contract specifies that at $s = 1$, all 320 channels are encoded, ensuring bit-exact reconstruction to the base model. At $s = 0$, only 32 structural channels are encoded, and 288 texture channels are physically excluded from the bitstream.

3.4 Blockhead and s-Mixing

The Blockhead decoder takes the 32 structural channels as input and reconstructs a structural color-block image x_{rough}. Compared with the backbone decoder g_s, the Blockhead is significantly smaller in parameter count, enabling efficient inference. The Blockhead is trained on data pre-processed by K-means with a dedicated loss function, learning a feature representation that facilitates effective K-means. At inference, K-means is applied on the Blockhead output to further improve the quality of x_{rough}, generating I_s. This design leverages the differentiability of the Blockhead for end-to-end training while benefiting from the non-differentiable K-means during inference.

The number of K-means clusters $K(s)$ adapts to the detail-exposure strength s and the spatial resolution of x. When s is large, more texture channels are retained, and the color-block image should preserve finer details; thus $K(s)$ is large. When s is small, coarse color blocks suffice, and $K(s)$ is small. Specifically, we define

$$\begin{aligned} K_0 &= K_{\text{base}} \cdot h \cdot w/256, \\ K(s) &= \max\left\{\text{round}(K_0 \cdot s^2),\, K_{\min},\, 1\right\}. \end{aligned} \tag{2}$$

where K_0 is proportional to the spatial dimensions of x, and $K_{\min} = K_0/16$ ensures a minimum number of blocks. The final reconstruction blends the color-block image I_s with the full reconstruction x_{full}, produced by the base decoder from the latent in which all 320 channel positions are filled (the retained n_{keep} channels plus the stuffed skipped channels):

$$\hat{x} = \sqrt{1-s} \cdot I_s + \left(1 - \sqrt{1-s}\right) \cdot x_{full} \tag{3}$$

At $s = 1$, the weighting coefficient for I_s is zero, yielding $\hat{x} = x_{full}$ and ensuring bit-exact reconstruction to the base compressor. At $s = 0$, the coefficient for x_{full} is zero, yielding $\hat{x} = I_s$, a pure color-block image with no texture detail. We choose the square-root function which provides a smoother perceptual transition and prevents abrupt texture cut-off in the mid-range, while ensuring that I_s is fully faded out as s approaches 1 and still maintaining a visible color-block effect at low s values.

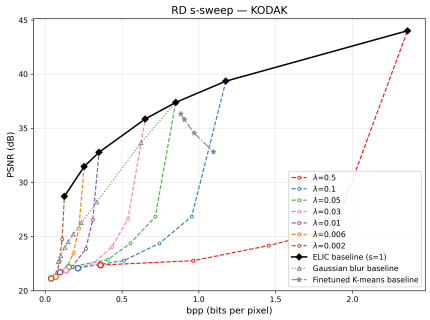


Figure 4: Rate-distortion performance on Kodak. Results from 7 λ checkpoints of sCBH are shown at $s = 0, 0.25, 0.5, 0.75, 1$. PSNR and bpp increase as s increases. The shape of the curve is jointly governed by multiple factors, including the s-mixing mechanism and the adaptive $K(s)$ schedule. The Gaussian blur and fine-tuned K-means baselines are discussed in Sec. 4.5.

4 Experiments

Datasets We use Kodak (24 images) as the test dataset and flicker2W (20769) as the training dataset. CLIC (102, both mobile and professional) and Tecnick (120) are used for cross-dataset validation. **Baselines** We compare sCBH against: (1) the base ELIC compressor; (2) K-means, where the original image is clustered before ELIC encoding; and (3) Gaussian blur, where the original image is blurred before ELIC encoding. For downstream probe comparisons, all baselines are used with their backbones frozen, and both the standard and fine-tuned K-means variants are evaluated. For rate-distortion comparisons, only the fine-tuned variant is used, as it achieves better performance. **Metrics** Rate-Distortion: PSNR (dB) and bpp (bits per pixel) calculated using bitstreams from arithmetic coding, which better reflects the true coding cost than likelihood-based estimation. Privacy probes: face detection rate, OCR character retention, skin lesion segmentation, and chest X-ray classification accuracy. The details are provided in Appendix A.2. We choose the detection rate instead of the verification accuracy because detection rate directly measures the detector's failure, thus serving as a more faithful privacy metric.

4.1 Continuous control via s-Sweep

We first examine the rate-distortion performance of sCBH across different compression levels. Fig. 4 presents the RD curves on Kodak across 7 λ checkpoints, each evaluated at $s = 0, 0.25, 0.5, 0.75, 1$. Both PSNR and bpp are monotonic over s, with the exact curve shape jointly governed by multiple factors including the s-mixing mechanism and the adaptive $K(s)$ schedule. Visual examples are provided in Appendix A.7.

To further verify the fine-grained controllability, Fig. 5 shows a denser sweep of s at $\lambda = 0.05$, with step size 0.05 from 0 to 1. Both metrics remain monotonic over the entire range, confirming that the continuous control is not limited to coarse discrete steps. The sweep is best read as a family of operating points: as $s \to 1$ the reconstruction approaches the base codec, while at low s suppression dominates; the intermediate range trades reconstruction quality for detail suppression, and the appropriate point is task-dependent. Beyond PSNR and bitrate, we also examine perceptual metrics on the same dense sweep. SSIM increases and LPIPS decreases monotonically with s for all five variants, tracing smooth trajectories in both s and bitrate (Appendix A.6); at $s = 1$ the metrics recover each base codec, while at $s = 0$ SSIM drops to 0.48–0.60 and LPIPS rises to 0.45–0.67, quantifying the amount of fine-grained detail that is deliberately removed.

4.2 Four Tasks Privacy Probes

We evaluate sCBH on four downstream tasks that cover both high-frequency-dependent and low-frequency-reliant scenarios. These tasks are selected to reflect the structural and texture channel split. Face recognition and OCR rely heavily on high-frequency details including facial landmarks, skin texture, and character strokes, while medical imaging tasks primarily depend on low-frequency structural information including organ shapes and intensity distributions.

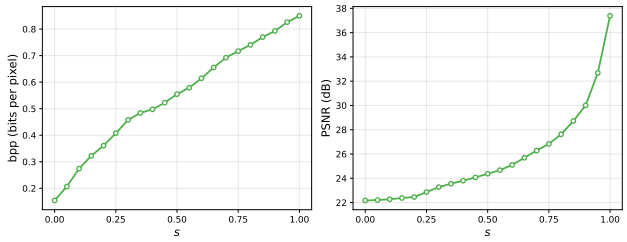


Figure 5: Continuous control via s-sweep on Kodak ($\lambda = 0.05$). PSNR and bpp are plotted against s with step size 0.05. Both metrics vary monotonically over the entire range, demonstrating fine-grained controllability of the single scalar s.

Fig. 6 reports the performance of all four tasks as s varies from 0 to 1. Face detection and OCR drop rapidly as s decreases, reaching near-zero performance at $s = 0$, while skin lesion segmentation and chest X-ray classification degrade more gradually. This behavior is consistent with the frequency interpretation of the structural and texture split: removing texture channels eliminates high-frequency information critical for face and text recognition, whereas the 32 structural channels preserve sufficient low-frequency information for coarse medical reasoning. For skin lesion segmentation (ISIC DSC), we observe an increase from $s = 1$ to $s = 0.75$ for most λ checkpoints, followed by a decline at lower s; the initial rise may be attributed to the removal of high-frequency texture noise, which simplifies the segmentation boundary, and as s approaches 0 the loss of structural detail dominates and DSC drops (the coarse s-sampling step of 0.25 does not resolve the exact peak, but the trend is consistent across checkpoints). We stress that we do not claim low-s reconstructions to be perceptually natural; the claim is that they preserve coarse structural information while suppressing fine-grained texture, consistent with the asymmetry observed in Fig. 6.

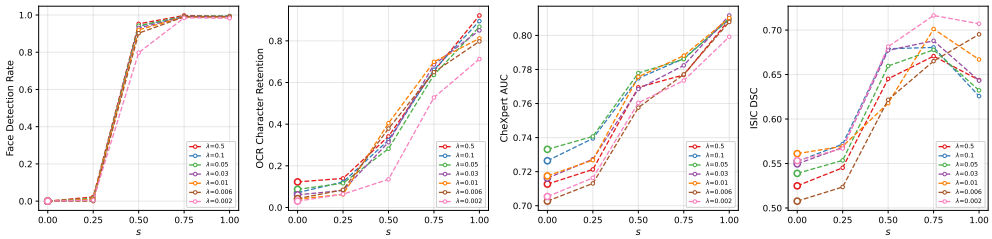


Figure 6: Privacy probes as a function of s. Four downstream tasks are evaluated. High-frequency tasks (face detection, OCR) exhibit rapid degradation, while low-frequency tasks (ISIC, CheXpert) show more gradual declines. The rise in ISIC DSC from $s = 0.25$ to $s = 0.75$ suggests that removing distracting texture details can improve segmentation performance, but the performance drops once the structural representation becomes degraded at lower s.

4.3 Zero Additional Quality Tax

Table 2: Comparison between sCBH s=1 and ELIC Baseline ($\lambda = 0.05$)

Dataset	Variant	bpp	PSNR (dB)	Δbpp / ΔPSNR
Kodak	ELIC baseline	0.8499	37.3849	—
	sCBH s=1	0.8499	37.3849	0.00 / 0.0000
CLIC	ELIC baseline	0.5825	38.1959	—
	sCBH s=1	0.5825	38.1959	0.00 / 0.0000
Tecnick	ELIC baseline	0.3620	40.6291	—
	sCBH s=1	0.3620	40.6291	0.00 / 0.0000

Table 2 confirms that sCBH at s=1 is bit-exact to ELIC. ΔPSNR and Δbpp are zero. Across all 28 downstream probe points, 26 are exactly identical and the remaining 2 differ within 0.0005 due to the floating-point precision. At $s = 1$, the s-mixing coefficient I_s is zero and $n_{keep} = 320$ encodes all channels.

4.4 CROSS-ARCHITECTURE GENERALITY

Table 3: Cross-Architecture s=1 Consistency (Kodak)

Metric	ELIC (SCCTX)	ELIC with AuxT	Hyperprior	CCM (checkerboard)	HPCM (progressive context)
ΔPSNR	0.00†	0.00	0.00	0.00	0.00
Δbpp	0.000	0.000	0.000	0.000	0.000

† All the average differences within 0.0001 are reported as zero.

We validate the generality of the sCBH channel-skip mechanism by applying it to four distinct architectures with different entropy models, all with M=320 latent channels. Table 3 confirms that s=1 is consistent across all architectures: the channel masking has no effect when all channels are retained. The key finding is that the entropy models all handle the partial-zero input correctly under frozen backbones, because the masked channels are bypassed during arithmetic entropy coding. Appendix A.5 shows the s-sweep across architectures; all exhibit PSNR degradation from s=1 to s=0, confirming that the channel-skip mechanism generalizes across different entropy model designs.

4.5 COMPARISON WITH BASELINES

Table 4: Comparison with Baselines ($\lambda = 0.05$).

Metric	sCBH s=0	s=0.5	Gaussian σ=2.5	K-means(Std) K=4	K-means(Std) K=8	K-means(FT) K=4	K-means(FT) K=8	ELIC
PSNR	22.1733	24.3765	24.5227	31.3077	33.5342	29.9578	32.8152	37.3849
bpp	0.1541	0.5547	0.1503	1.0216	0.9424	1.1613	1.0960	0.8499
CelebA det	0.0000	0.9425	0.9850	0.4825	0.9525	0.3125	0.9175	0.9950
OCR ret	0.0870	0.2820	0.3395	0.8194	0.8417	0.7804	0.8385	0.8677

Table 4 shows the comparison between sCBH and baselines mentioned in Sec. 4 at different levels of detail suppression. Gaussian blur preserves nearly all faces (det = 0.985) despite similar bpp (0.150) and PSNR (24.52 dB) to sCBH $s = 0$ (0.154 bpp, 22.17 dB), and it is invertible in principle via Wiener deconvolution, making it a weak protection. Lowering the rate alone is equally ineffective: the base ELIC model at 0.126 bpp still detects 98.3% of faces. At matched suppression, sCBH is also the most efficient: fine-tuned K-means requires 1.16 bpp to reduce face detection to 0.3125 (31.25%). Fine-tuned with the same data and iterations as sCBH, ELIC on blockified data does not improve the trade-off: K-means remains above 31 dB PSNR, and the fine-tuned codec loses compression efficiency, confirming that the ELIC architecture is ill-suited for color-block representations. Because these baselines adopt fundamentally different design objectives, we compare system-level overhead and suppression effectiveness rather than numerical alignment (Appendix A.3).

4.6 ABLATION STUDIES

We conduct ablation studies to dissect two critical design choices: (i) freezing the pre-trained backbone versus end-to-end (E2E) FineTune, and (ii) applying K-means blockification solely at inference.

Table 5: Ablation Studies ($\lambda = 0.05$, Kodak)

Metric	sCBH s=1	End-to-end training	w/o channel skip
PSNR	37.3849	33.5407	—
bpp	0.8499	1.5417	Skipping reduces bpp

Freezing the backbone is mandatory. As reported in Table 5, end-to-end training with the same number of iterations causes a 3.5 dB PSNR drop at s=1 and an irreversible bitrate inflation: the shared backbone distorts its pre-trained representation to accommodate low-s masked inputs, breaking bit-exact equivalence. Freezing eliminates this drift, guaranteeing zero additional quality tax; the variable-length skipping removes tail texture channels from the bitstream, saving bits effectively.

K-means blockification and channel skipping together form the core suppression mechanism, yet the effect is task-dependent. Since it is applied only at inference, it does not alter the bitstream (Table 6). Removing it exposes a dichotomy: face detection explodes at s=0 and s=0.25,

Table 6: Ablation study of K-means ($\lambda = 0.05$). Columns alternate between sCBH and the same setting without inference-time K-means (w/o); s increases from 0 to 0.75 from left to right.

Metric	sCBH 0	w/o 0	sCBH 0.25	w/o 0.25	sCBH 0.5	w/o 0.5	sCBH 0.75	w/o 0.75
PSNR	22.1733	22.6249	22.8665	23.2617	24.3765	24.4735	26.8332	26.8764
bpp	0.1541	0.1541	0.4076	0.4076	0.5547	0.5547	0.7168	0.7168
CelebA det	0.0000	0.8675	0.0150	0.9525	0.9425	0.9875	0.9925	0.9950
OCR ret	0.0870	0.0893	0.1163	0.1141	0.2820	0.2958	0.6367	0.6285

confirming that the hard color-block boundaries are essential to disrupt the facial layout (e.g., facial contours) on which detectors rely, whereas OCR is nearly unaffected because text strokes are already smoothed by the raw Blockhead output. K-means improves PSNR only marginally; without it, the raw Blockhead output produces hazy gradients with blocking artifacts, while with it the reconstruction is crisp and visually hides the low-bitrate degradation (Appendix A.7). The structural-channel ablation is discussed in Appendix A.4.

5 Conclusion

We presented sCBH (s-Contract with Blockhead), a channel-level selective skipping framework for continuous detail-exposure control in learned image compression. sCBH achieves zero quality tax at full quality. The method is task-agnostic. The training loss contains no downstream task terms, yet generalizes to unseen tasks with varying degrees of protection determined by their reliance on high-frequency texture information. Our results demonstrate that view-level structural/texture separation, combined with physical channel skipping, offers a more effective detail-quality-bitrate trade-off than conventional baselines such as Gaussian blur or K-means preprocessing. sCBH provides a feasible framework for compression with controllable detail suppression, where the degree of suppression is continuously adjustable via a single parameter without retraining or external detectors.

Several limitations remain. High fidelity and strong suppression cannot be achieved at the same operating point: sCBH guarantees the zero-cost full-quality mode at $s = 1$ and efficient suppression at low s, with a continuous transition in between. Protection on medical tasks is moderate, reflecting that diagnostic features largely reside in structural channels. The s parameter is global across the image; region-selective control would require additional localization. Large-image inference requires K-means clustering at high spatial resolution, which can be slow on CPU; GPU acceleration or hierarchical clustering is required. Future work includes extending sCBH to bitstream-level privacy through adversarial identity removal or invertible separation, region-selective privacy using semantic localization, and acceleration methods for large-image deployment. Delicate designs are required for scenarios with an untrusted decoder.

AI use statement

The authors are non-native English speakers. During the preparation, we used a large language model solely for language polishing and repetitive tasks. The LLM was employed to improve the clarity and accuracy of our written expression; it was not involved in any scientific aspects of this work, including research design, data processing, experimental execution, result interpretation, or conclusion formation. All authors have reviewed the final content.

Reproducibility statement

We provide all necessary details to reproduce our experiments. The training configuration, including architecture specifications, optimization hyperparameters, loss functions, and data preprocessing, is fully described in Appendix A.1. Our implementation is built upon (Bégaint et al., 2020). Evaluation protocols for all four downstream tasks, including dataset splits and metric definitions, are given in Sec. 4 and Appendix A.2. The source code and pre-trained models are available in an anonymized repository (`https://anonymous.4open.science/r/sCBH`) and will be open-sourced upon publication.

REFERENCES

Saeed Ranjbar Alvar and Ivan V Bajić. Scalable privacy in multi-task image compression. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5. IEEE, 2021.

Bardia Azizian and Ivan V Bajić. Privacy-preserving feature coding for machines. In *2022 Picture Coding Symposium (PCS)*, pp. 205–209. IEEE, 2022.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja. Compressai: a pytorch library and evaluation platform for end-to-end compression research. *arXiv preprint arXiv:2011.03029*, 2020.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5202–5211, 2020. doi: 10.1109/CVPR42600.2020.00525.

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

Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5708–5717. IEEE, 2022.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1398–1406. IEEE, 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pp. 1156–1160. IEEE, 2015.

Han Li, Shaohui Li, Wenrui Dai, Maida Cao, Nuowen Kan, Chenglin Li, Junni Zou, and Hongkai Xiong. On disentangled training for nonlinear transform in learned image compression. In *International Conference on Learning Representations*, 2025a.

Yuqi Li, Haotian Zhang, Li Li, and Dong Liu. Learned image compression with hierarchical progressive context modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18834–18843. IEEE, 2025b.

Christoph Liebender, Ranulfo Bezerra, Kazunori Ohno, and Satoshi Tadokoro. Region of interest loss for anonymizing learned image compression. In *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pp. 3569–3576. IEEE, 2024.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018.

Behrooz Razeghi, Shideh Rezaeifar, Sohrab Ferdowsi, Taras Holotyak, and Slava Voloshynovskiy. Compressed data sharing based on information bottleneck model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3009–3013. IEEE, 2022.

Xuelin Shen, Jiayin Xu, Kangsheng Yin, and Wenhan Yang. Privacy-shielded image compression: Defending against exploitation from vision-language pretrained models. *arXiv preprint arXiv:2506.15201*, 2025.

A Appendix

A.1 Training Details

Blockhead architecture. The Blockhead is a lightweight 3-layer CNN: it first applies bilinear $16\times$ upsampling to restore spatial resolution, then three 3×3 convolutions ($3\to64\to64\to3$ channels) with ReLU activations and a final Sigmoid, producing a smooth RGB output $x_{\text{rough}} \in [0,1]^{3\times H\times W}$. The total parameter count is $\sim$57K, compared to $\sim$12M for the ELIC backbone decoder g_s.

Table 7: Training Configuration

Item	Value
Dataset	flicker2W (20769 images), RandomCrop 256
Epoch / Batch	30 / 8
Optimizer	Adam, lr= 10^{-4} (aux 10^{-3})
Initialization	Transfer from ELIC re3 (λ=5×10^{-2})
Frozen modules	g_a, h_a, h_s, SCCTX, g_s, EB
Trained modules	Blockhead
s sampling	25% s=1 + 25% s=0 + 50% Beta(0.5, 0.5)

The training configuration is summarized in Table 7. The key design choice is *freezing the backbone*: end-to-end training (even after carefully tuning the rate and loss coefficients) causes s=1 to drift by -3.5 dB with higher bitrate, because the shared backbone distorts the pre-trained representation. Freezing guarantees that s=1 is bit-exact to ELIC.

The sCBH's training is continued from the base model. The training loss of sCBH combines quality, structure, smoothness, channel penalty, and rate terms:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{struct}} \cdot 255^2 \cdot L_1(x_{\text{rough}}, \text{teacher}) \cdot (1-s) \\ & + \lambda_{\text{tv}} \cdot \text{TV}(x_{\text{rough}}) \\ & + \text{bpp}_y + \text{bpp}_z \text{ (End-to-End Only)} \\ & + \lambda_{\text{chan}} \sum_{c=k_t}^{287} \frac{c}{288} \cdot \|z_{t0}[:,c]\|^2 \text{ (End-to-End Only)} \\ & + \lambda \cdot 255^2 \cdot \text{MSE}(x_{\text{full}}, x) \cdot \alpha(s) \text{ (End-to-End Only)}\end{aligned} \tag{4}$$

where the structural teacher is the original image processed by K-means at $K(s)$ clusters, weighted by $(1-s)$ to vanish at s=1; $\text{TV}(x) = \frac{1}{HW}\sum_{i,j}(|x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}|)$ encourages spatial smoothness in the Blockhead output for cleaner K-means boundaries at inference. The last three terms are active only during end-to-end training for fair ablation study in Sec. 4.6. $\alpha(s) = s^2$ suppresses the quality path at low s to prevent bitrate inflation. The channel penalty applies position-weighted magnitude regularization to skipped channels, pushing the encoder to concentrate information in the early structural channels.

Hyperparameters: λ=0.05, λ_{struct}=0.03, λ_{tv}=0.01, λ_{chan}=0.0005, K_{base}=64. At 256×256 resolution, K_0=64 and $K_{\min}=K_0/16=4$.

The specific choice of 32 structural channels is architecturally motivated by the grouped-channel design of SCCTX, which partitions the 320-channel latent into five groups. The first two groups, totaling 32 channels, are designated to encode the most foundational spatial priors. Empirically, this threshold is validated by the reconstruction performance of the Blockhead decoder: utilizing only these 32 channels yields a structural reconstruction I_s with approximately 24 dB PSNR, which reliably preserves coarse shapes and global intensity distributions while effectively suppressing high-frequency textures. This 24 dB baseline serves as a practical sweet spot, ensuring sufficient structural fidelity for low-frequency tasks (e.g., medical imaging) without compromising the disruption effect on texture-dependent tasks.

The training device is NVIDIA RTX 3090 24G. The device for evaluation and downstream tasks is NVIDIA RTX 2080Ti 11G. The CPU is Intel Xeon E5-2650 v4 @ 2.20GHz.

A.2 Experimental Metrics

Rate-Distortion: PSNR (dB) and bpp (bits per pixel) calculated using bitstreams from arithmetic coding, which better reflects the true coding cost than likelihood-based estimation.

Privacy probes: face detection rate (det_rate, CelebA(Liu et al., 2015), 400 pairs chosen), OCR character retention (Karatzas et al., 2015) (100 images chosen), skin lesion segmentation (DSC on ISIC (Codella et al., 2019), 200 images chosen), and chest X-ray classification accuracy (AUC on CheXpert (Irvin et al., 2019), 200 images chosen).

We evaluate face privacy using the InsightFace toolkit, which combines RetinaFace (Deng et al., 2020) for face detection with ArcFace (Deng et al., 2019) for identity embedding. The det_rate measures the fraction of image pairs where both faces are successfully detected by the detector. This metric directly captures the dominant privacy mechanism of sCBH at low s: the color-block reconstruction prevents the face detector from locating faces, making subsequent identity verification impossible. We do not report verification accuracy because, at low s, most faces are undetectable and accuracy would be computed on an unreliable survivor subset; reporting accuracy on such a biased survivor subset would misrepresent the method's actual privacy protection in real-world scenarios.

A.3 Discussion on Baseline Comparability and System Complexity

A perfectly aligned comparison is challenging because existing approaches adopt different design objectives and assumptions, as prior works (Sec. 2) adopt fundamentally different design philosophies. Specifically, ROI-based methods rely on external object detectors (e.g., YOLO or MTCNN) that introduce heavy computational pre-processing; task-guided bottlenecks and adversarial frameworks incorporate task-specific labels or discriminators, which require non-trivial re-training and destabilize the original compression objective.

In contrast, sCBH deliberately refrains from incorporating any such external modules. Our design introduces only a lightweight Blockhead decoder with $\sim$57K additional parameters (compared to $\sim$12M for the backbone decoder and $\gg$10M for typical detectors), and our training loss contains no downstream task terms or adversarial signals. The optimization signals of our method and these baselines are therefore in entirely different orders of magnitude and complexity regimes.

Consequently, we argue that forcing a direct one-to-one numerical comparison (e.g., PSNR at the exact same bpp) against these heavily augmented baselines would be misleading rather than informative. The key takeaway is that sCBH achieves competitive or superior downstream task degradation while introduces substantially less architectural and training overhead, making it genuinely deployable in resource-constrained scenarios where external detectors are unavailable or too expensive.

A.4 Ablation study for different structural channels

Structural channel count ablation. We evaluate four structural channel configurations $c_s \in \{16, 32, 48, 64\}$ (with $c_t = 320 - c_s$ texture channels) at λ=0.05. All configurations achieve identical reconstruction quality at s=1, confirming that the split ratio does not affect full-quality decoding.

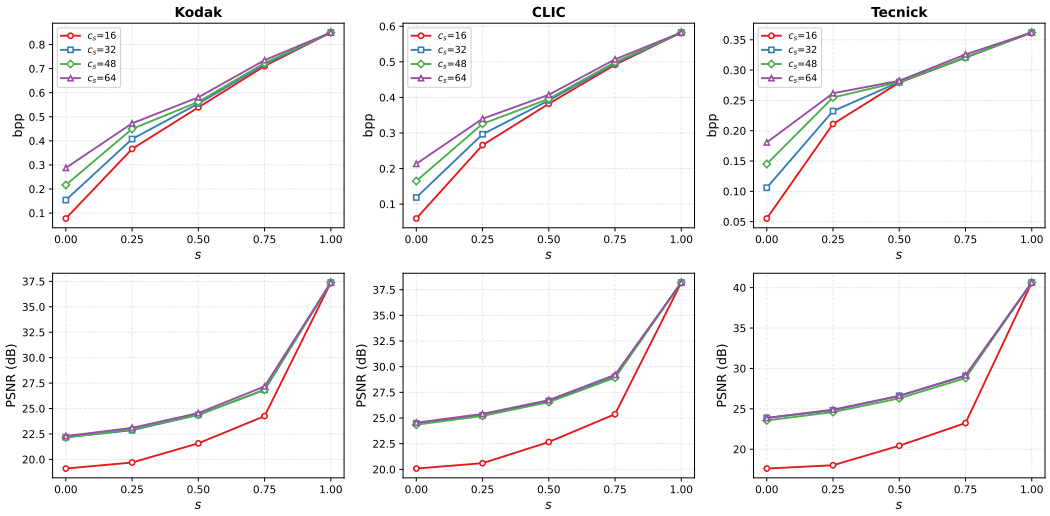


Figure 7: Effect of structural channel c_s count on rate-distortion performance at λ=0.05. Up: bpp vs. s; Down: PSNR vs. s. All configurations converge to identical ELIC-level quality at s=1. At s=0, c_s=16 suffers a 3 dB structural quality deficit, while c_s=48/64 incur higher mandatory bitrate than c_s=32 without quality improvement.

At s=0, the c_s=16 configuration yields only 19.10 dB, 3.1 dB below c_s=32, indicating insufficient structural information for faithful layout reconstruction. Increasing c_s beyond 32 (to 48 or 64) provides negligible PSNR gains at s=0 (22.16–22.29 dB vs. 22.17 dB for c_s=32) while incurring proportionally higher mandatory bitrate: 0.217 and 0.288 bpp respectively, compared to 0.154 bpp for c_s=32. This confirms that 32 structural channels capture the full structural information of the latent, and additional channels encode only texture redundancy that increases the control-irrelevant bitrate floor.

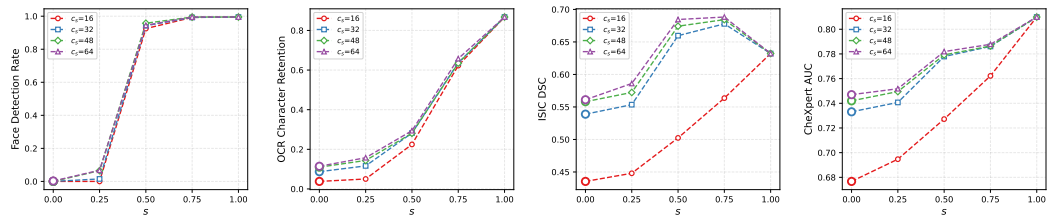


Figure 8: Privacy probe metrics across c_s configurations at λ=0.05. Each subplot shows one downstream task's performance-s.

Privacy probe analysis across c_s configurations. We measure four downstream task performances as a function of s for each c_s value. At s=0 (maximum suppression), c_s=16 and c_s=32 both achieve complete face identity protection (face detection rate = 0), while c_s=64 exhibits marginal leakage (det_rate = 0.0025). The OCR retention rate at s=0 increases monotonically with c_s, confirming that more structural channels inevitably leak more textual content. For medical probes (ISIC DSC and CheXpert AUC), the s=0 values also rise with c_s, reflecting the same trend: structural channels carry task-relevant information that cannot be selectively removed. Importantly, the transition from s=0 to s=0.5 is steepest for c_s=32 on CelebA (Δdet_rate = 0.943), indicating finer-grained control over the suppression–utility tradeoff. We select c_s=32 as it achieves: (1) zero face detection rate at s=0, (2) the lowest OCR retention among configurations with sufficient structural quality, (3) the sharpest transition, and (4) minimal mandatory bitrate.

A.5 Additional Rate-Distortion and Privacy Curves

The absolute values differ across architectures, reflecting their different base RD performance. But the relative behavior is consistent. Since ELIC with AuxT incorporates the wavelet-based linear shortcut (WLS) module compressing the energy of high frequencies, the R-D curve has slight fluctuations. However, this does not affect our analysis or conclusion.

A.5.1 Rate-Distortion Curves

Fig. 9 shows the Rate-Distortion curves of all the five variants.

A.5.2 Privacy Probes

Fig. 10 shows the four downstream tasks' performance of the four additional variants (the ELIC variant is shown in Fig. 6).

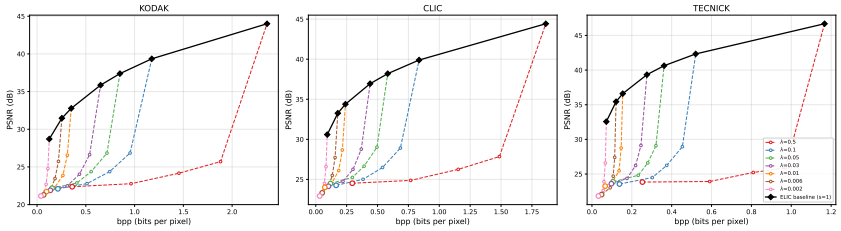
A.6 Perceptual Metrics

Fig. 11 shows SSIM and LPIPS of the decoded view as functions of the detail-exposure strength s (left column) and of the measured bitrate (right column), for all five variants on Kodak ($\lambda = 0.05$, s sampled with step 0.05). Across all variants, SSIM increases and LPIPS decreases monotonically over the entire range of s, and both metrics trace smooth trajectories in both coordinates, confirming that the single scalar s provides continuous control over the perceptual detail of the decoded view rather than a binary switch. At $s = 1$ the metrics match each variant's base codec, consistent with bit-exact reconstruction. At $s = 0$, SSIM falls to 0.48–0.60 and LPIPS rises to 0.45–0.67, quantifying the deliberate removal of fine-grained detail. We stress that these metrics measure the departure of the decoded view from the original image, i.e., how much detail is no longer exposed; they are not presented as evidence of perceptual quality, since low-s reconstructions are intentionally low-detail. Structural interpretability at low s is supported by the downstream task analysis (Fig. 6) and the visual examples in Appendix A.7. Absolute values differ across variants, reflecting their different base RD performance, while the relative behavior is consistent.

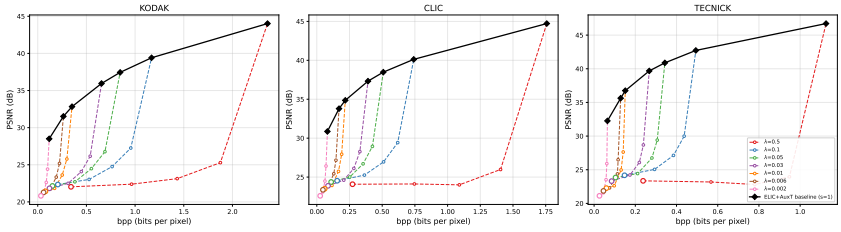
A.7 Visual s-Sweep

Fig. 12, 13, 14, and 15 show the visual s-sweep on four Kodak images. Each 6×5 grid shows I_s (row 1), x_{full} (row 2), $\hat{x}$ (row 3) at full resolution, and their crops (rows 4–6), across $s \in \{0, 0.25, 0.5, 0.75, 1\}$. At s=0, only K=4 color blocks are visible; at s=1, the output is bit-exact to ELIC. Best viewed at full resolution.

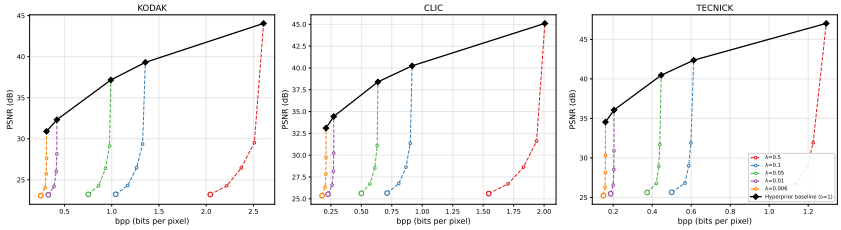
Fig. 16, 17, 18, and 19 show the comparison with and without K-means at inference. Top row: sCBH with K-means, producing sharp color boundaries. Bottom row: sCBH without K-means, the raw Blockhead output retains smoother gradients, reducing the suppression effect. The image is cropped into 256×256.



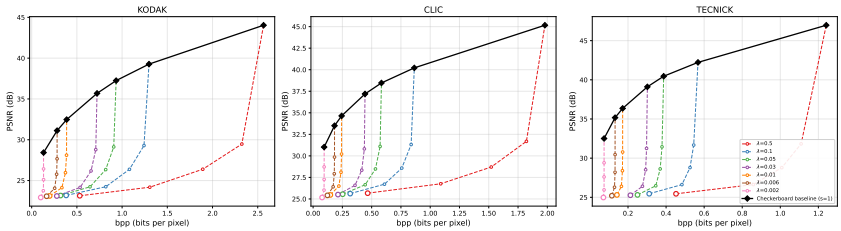
(a) ELIC



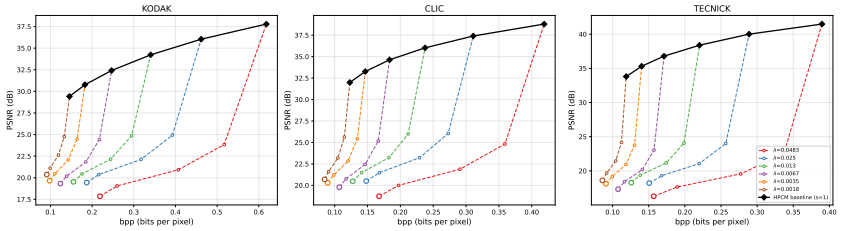
(b) ELIC with AuxT



(c) Hyperprior

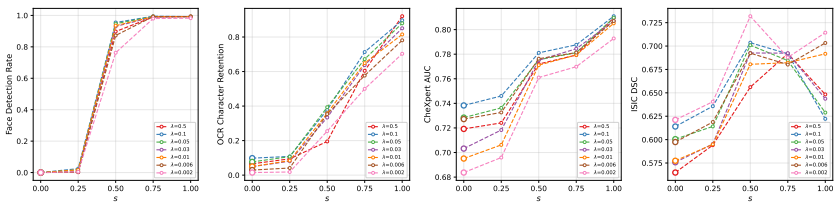


(d) CCM

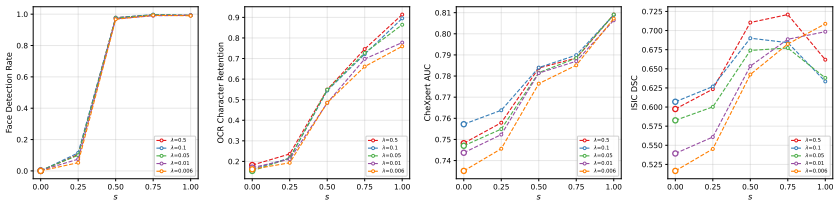


(e) HPCM

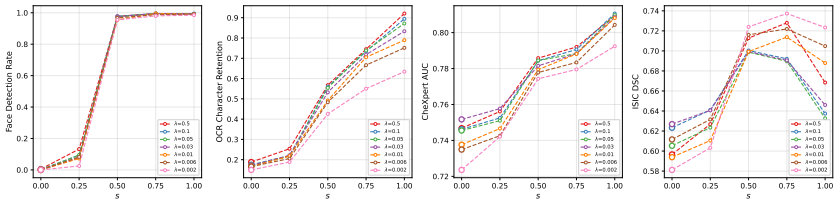
Figure 9: Rate-distortion curves for all five architectures on Kodak, CLIC, and Tecnick. Each plot shows the sCBH s-sweep against the base compressor's RD curve, confirming that $s = 1$ lies on the base RD curve and s provides continuous control along it.



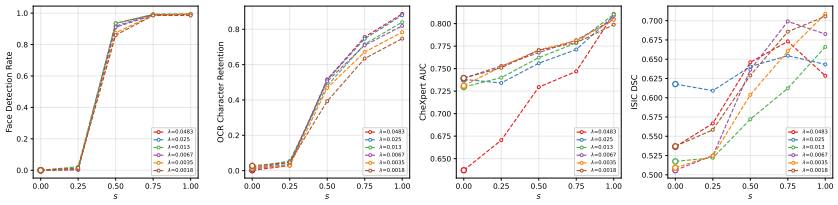
(a) ELIC with AuxT



(b) Hyperprior

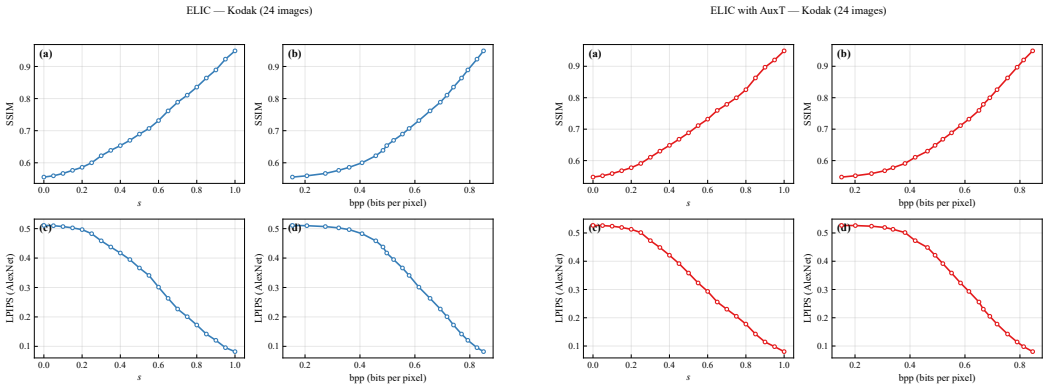


(c) CCM



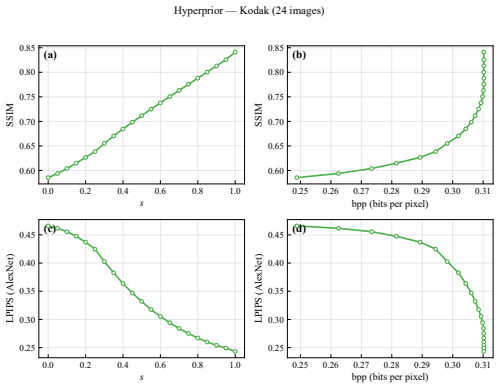
(d) HPCM

Figure 10: Four-task privacy probes across architectures. Each subplot shows CelebA detection rate, OCR retention, ISIC DSC, and CheXpert AUC as functions of s. All architectures exhibit downstream task degradation from s=1 to s=0, confirming cross-architecture generality.

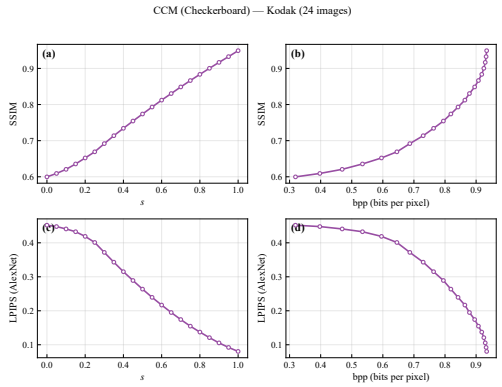


(a) ELIC

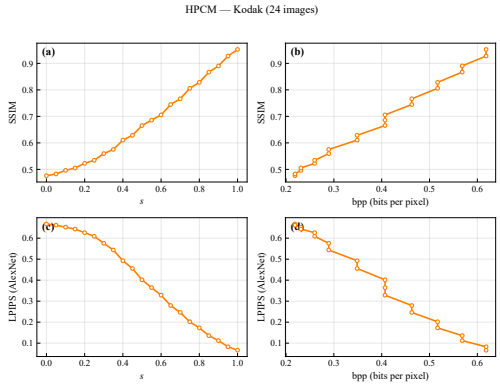
(b) ELIC with AuxT



(c) Hyperprior



(d) CCM



(e) HPCM

Figure 11: Perceptual metrics for all five variants on Kodak ($\lambda = 0.05$). Each panel shows SSIM (top row) and LPIPS (bottom row) as functions of s (left column) and bitrate (right column). All variants exhibit monotonic SSIM increase and LPIPS decrease over the full range of s.



Figure 12: kodim01 (architecture)

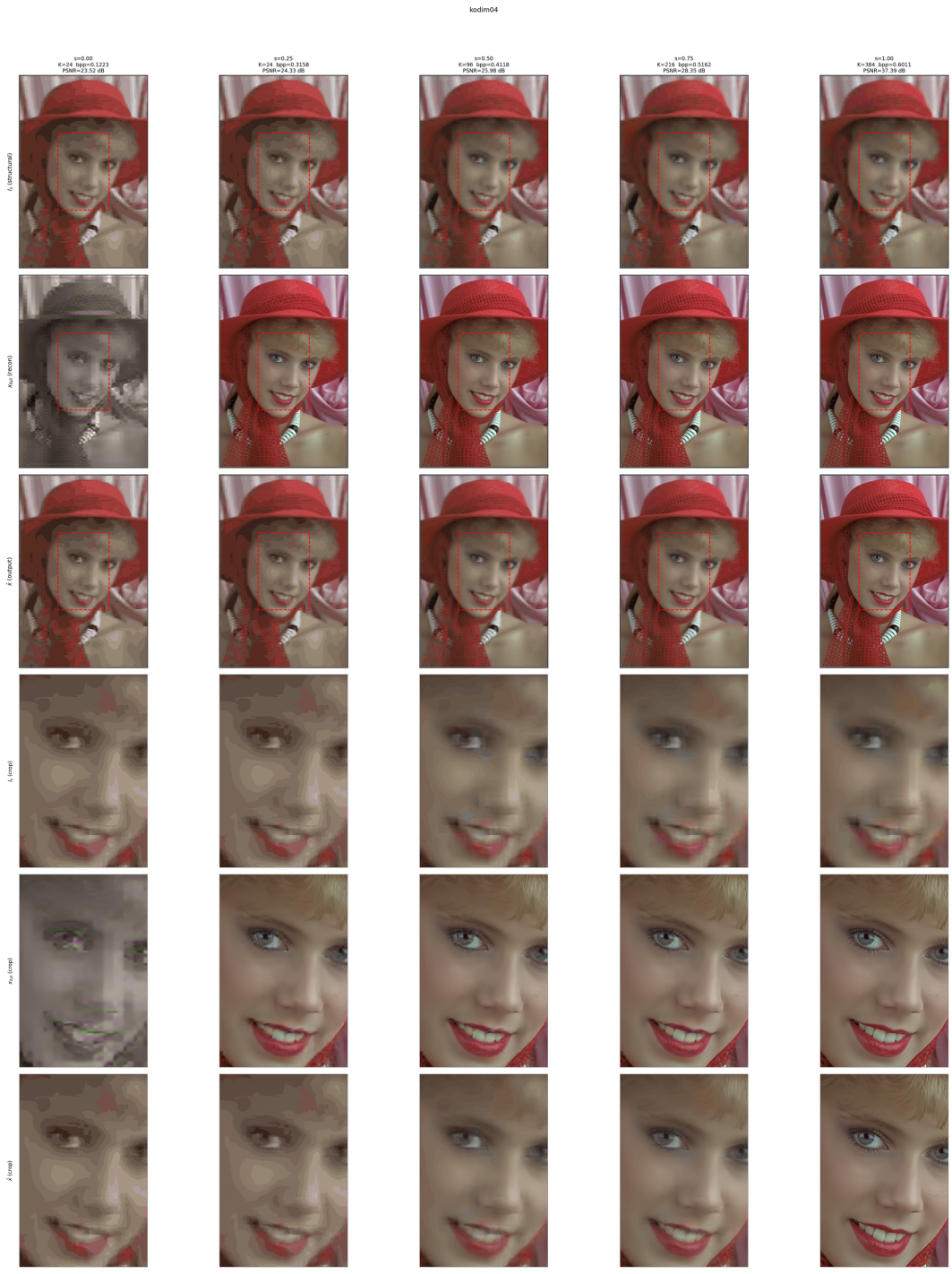


Figure 13: kodim04 (close-up face)

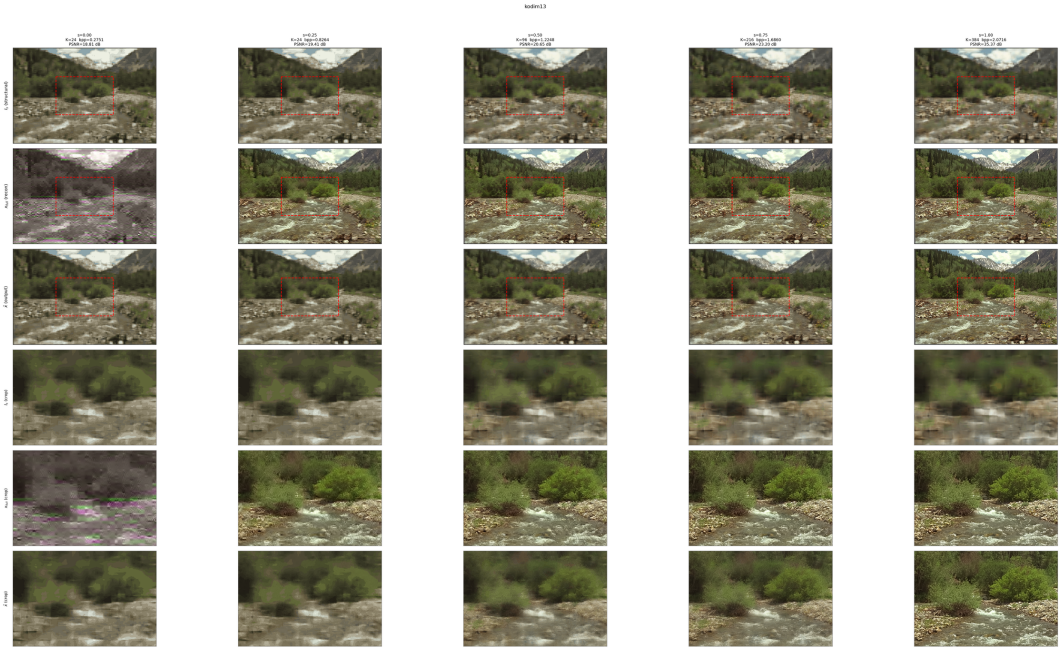


Figure 14: kodim13 (landscape)



Figure 15: kodim14 (group portrait)

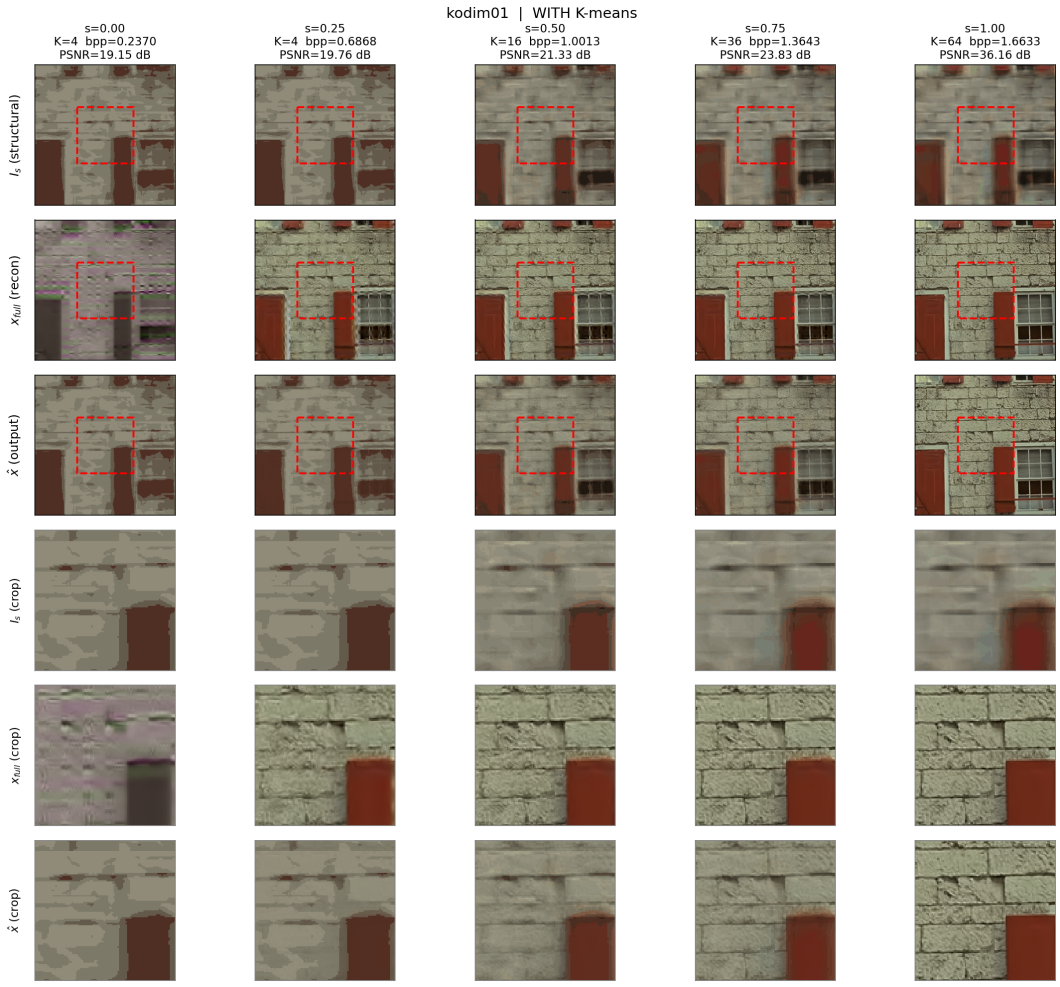


Figure 16: with K-means (kodim01)

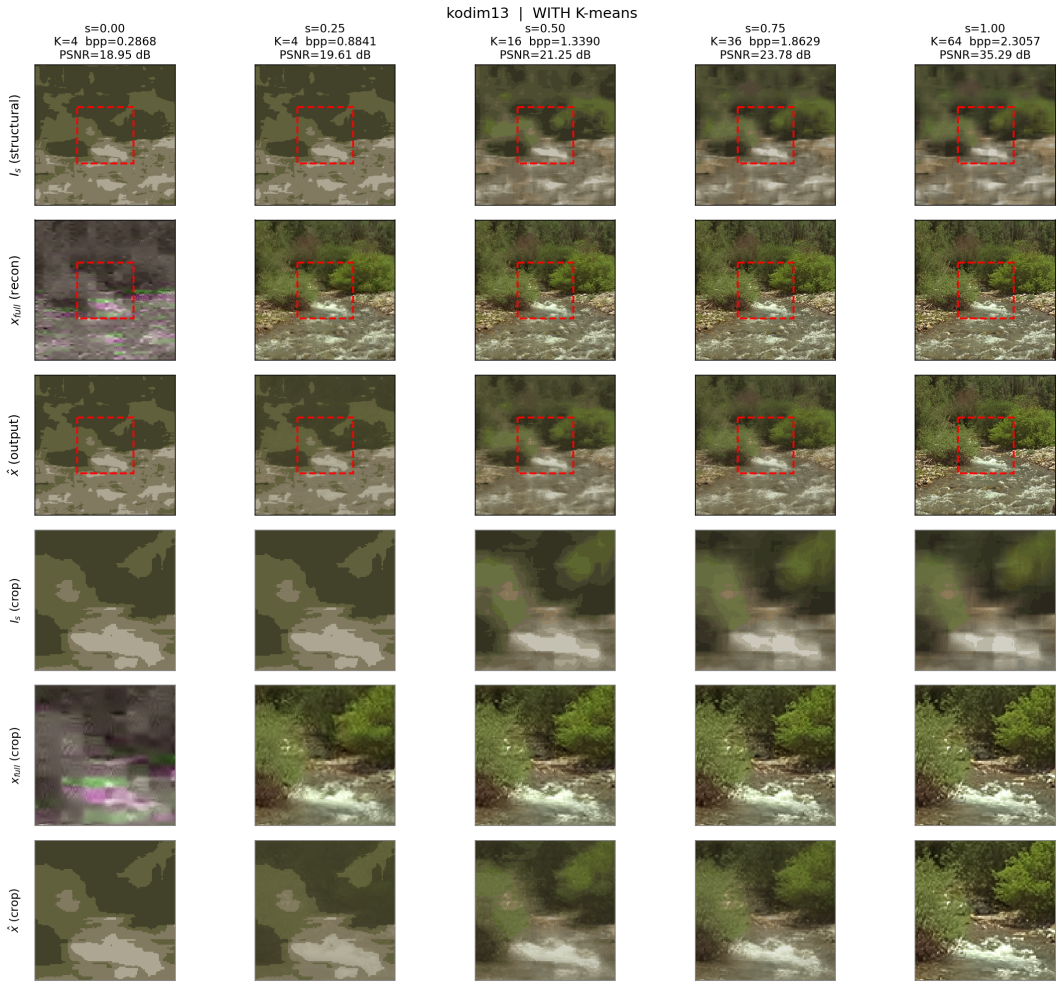


Figure 17: with K-means (kodim13)

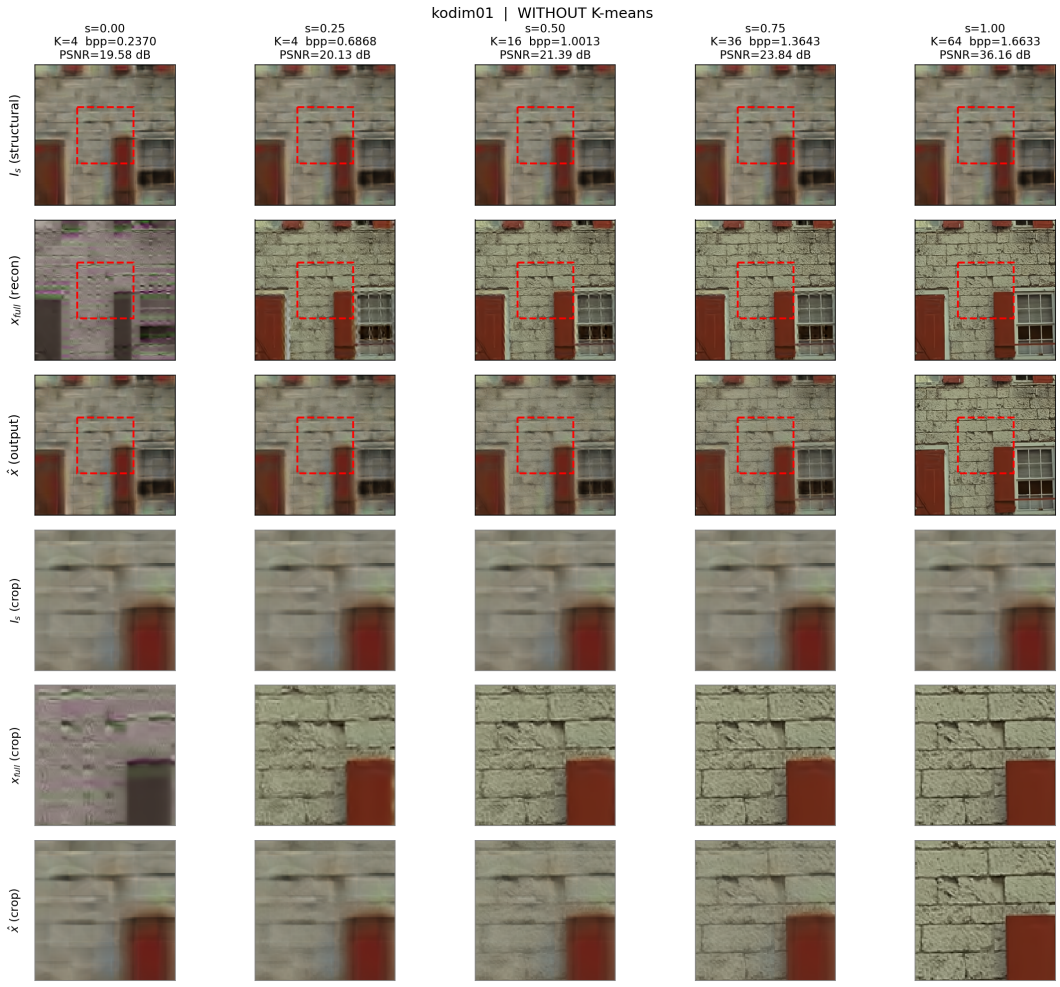


Figure 18: without K-means (kodim01)

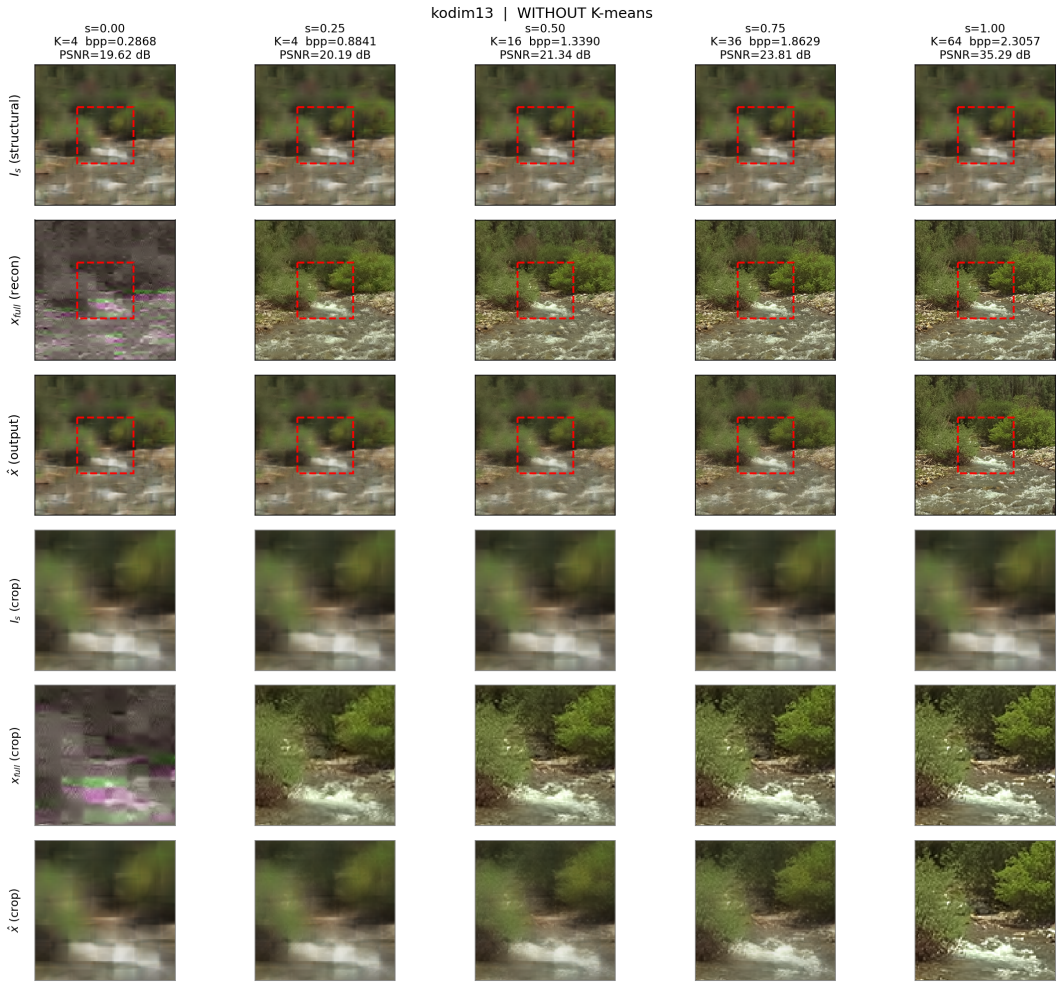


Figure 19: without K-means (kodim13)