

MixCompress: Mixture of Experts for Variable Rate Learned Image Compression

Calvin-Khang Ta* , Praneet Singh* , Tong Shao , and Peng Yin 

Dolby Laboratories, USA
{czta, pusing, tshao, pyin}@dolby.com

Abstract. Learned image compression (LIC) is bottlenecked by the need to store independent models for each rate-distortion operating point. Existing variable bit-rate (VBR) methods aim to reduce this overhead via dense parameter modulation, but forcing a shared backbone to approximate divergent mappings causes severe feature entanglement. Specifically, low-rate smoothing gradients inherently conflict with the preservation of high-frequency textural details, leading to sub-optimal performance. To resolve this, we propose MixCompress, a unified VBR framework based on sparse structural specialization. While sparsely gated Mixture-of-Experts (MoE) routing successfully mitigates gradient conflict, it operates on a fixed computational budget. To address the increased representational demands of higher bit-rates we introduce a Mixture-of-Depths (MoD) extension to dynamically scale model capacity. Combined with Conditional Auxiliary Transforms (CAT) for dynamic sub-band energy modulation, our hierarchical framework effectively dynamically scales capacity. Extensive evaluations demonstrate that MixCompress not only matches individually optimized single-rate baselines but can even surpass them, establishing a new Pareto frontier for computationally efficient image coding.

Keywords: Representation Learning · Image Compression · Mixture of Experts

1 Introduction

Learned image compression (LIC) has emerged as a powerful alternative to conventional codecs by jointly optimizing nonlinear transforms and entropy models under a rate-distortion objective [5]. Hierarchical hyper-priors [6], joint autoregressive modeling [25], and refined context modeling [7, 15, 16] progressively closed the gap to traditional codecs. More recent Transformer and linear-attention architectures [13, 22, 24] further expanded representational power while maintaining practical decoding complexity, establishing LIC as a high-performance learned alternative for image coding. Despite these advances, state-of-the-art LIC models require separate models for each operating point, leading to increased training costs, storage overhead, and deployment complexity.

* Equal Contribution

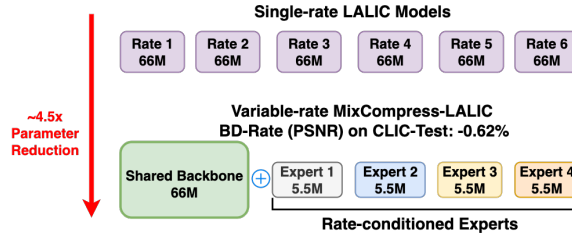


Fig. 1: Parameter Comparison. Traditional approaches require separate models for each rate point (top), while MixCompress shares 75% of parameters across all rates with only rate-specific expert modules (bottom). The anchor is the single rate LALIC.

Variable-rate LICs aim to address this limitation by supporting multiple operating points within a single unified model. Existing approaches rely on conditioning a shared dense backbone [9, 30] via latent gain scaling [10] or global affine transformations [29]. While effective for unified deployment, these methods underperform their single rate optimized counterparts. This performance gap can be attributed to a fundamental optimization flaw: dense parameter modulation forces the single shared backbone to approximate divergent rate-distortion mappings. At the extremes of the rate-distortion trade-off, dense parameter sharing induces severe feature entanglement. Specifically, the parameters responsible for preserving high-frequency textural details at peak bit-rates are subjected to destructive smoothing gradients from low bit-rate optimization. This results in gradient conflict and creates a strict capacity bottleneck.

To overcome this limitation, we introduce MixCompress, a unified conditional framework that combines structural specialization with rate-aware representation learning. At the core of our approach is the use of sparsely gated Mixture-of-Experts (MoE) or Mixture-of-Depths (MoD) modules within key bottlenecks of the analysis transform, synthesis transform, and hyper-prior networks. Rather than forcing a shared backbone to compromise across the rate-distortion curve, the routing mechanism provides gradient conflict mitigation. By isolating the gradients from low bit-rate samples we shield the high bit-rate experts from learning suboptimal features. This allows for the model to more effectively model the entire rate-distortion operating curve.

We complement this structural routing with Conditional Auxiliary Transforms (CAT), a rate-aware wavelet shortcut that modulates sub-band energy as an explicit function of the Lagrange multiplier λ . While MoE and MoD specialize structure, CAT enables feature energy compaction—introducing rate-dependent behavior without altering the entropy model or arithmetic coding pipeline.

By aligning conditional capacity allocation directly with the rate-distortion objective, MixCompress effectively decouples the optimization trajectories of competing rate points. Extensive experiments on the Kodak, CLIC, and Tecnick datasets demonstrate our architecture not only closes the variable-rate performance gap, but significantly outperforms state-of-the-art single-rate baselines, including LALIC [13], LIC-TCM [24], and FAT [22]. Importantly, our variable-

rate models are trained for the same duration as one single-rate model, yet replaces multiple independently trained single-rate checkpoints, substantially reducing both checkpoint storage and total parameter footprint. MixCompress thus establishes a new Pareto frontier for practical neural image compression, achieving superior representational density through sparse conditional computation without a proportional increase in inference cost (Fig. 1).

2 Related Works

2.1 Learned Image Compression

Modern learned image compression (LIC) is based on end-to-end optimized non-linear transform coding frameworks [5, 6], commonly enhanced with hyper-prior and auto-regressive context models to estimate latent distributions [6, 25]. Recent work has improved representation capacity through Transformers [22, 24], linear attention mechanisms [13], and auxiliary wavelet shortcuts (AuxT) that accelerate training via lightweight parallel de-correlation branches [21]. Despite these advances, most LIC methods remain optimized for a single rate point, requiring separate model parameters for different operating bitrates.

2.2 Variable-Rate Compression

Variable-rate LIC enables multiple rate-distortion (R-D) operating points within a single model. Conditional autoencoders [9] and modulated autoencoders [30] condition the network on the Lagrange multiplier to enable coarse rate control. Gain-based approaches such as the asymmetric gained VAE [10] allow continuous bit-rate adaptation via latent scaling, while quantization-aware frameworks [29] and scalable decoding methods [32] further enhance flexibility.

While these approaches simplify deployment, they largely rely on conditioning or latent modulation within a shared dense backbone. As a result, unified models often underperform individually optimized single-rate counterparts, revealing a gap between parameter sharing and per-rate specialization. Importantly, most methods adapt representations but not model structure, limiting conditional capacity allocation across bit-rate regimes.

2.3 Mixture-of-Experts and Conditional Computation

Mixture-of-experts (MoE) architectures enable conditional computation by routing inputs to specialized expert subnetworks [17]. Sparsely gated MoE layers activate only a subset of experts per sample, dramatically increasing effective capacity without proportional computational cost [28]. Large-scale systems such as GShard and Switch Transformer [12, 20] depict the scalability of conditional routing, and vision MoE models [26] extend these benefits to image tasks. In image compression, mixture-based representations have appeared in alternative coding paradigms [14], but sparsely gated MoE has not been integrated into hierarchical transform-based LIC, particularly for variable-rate modeling.

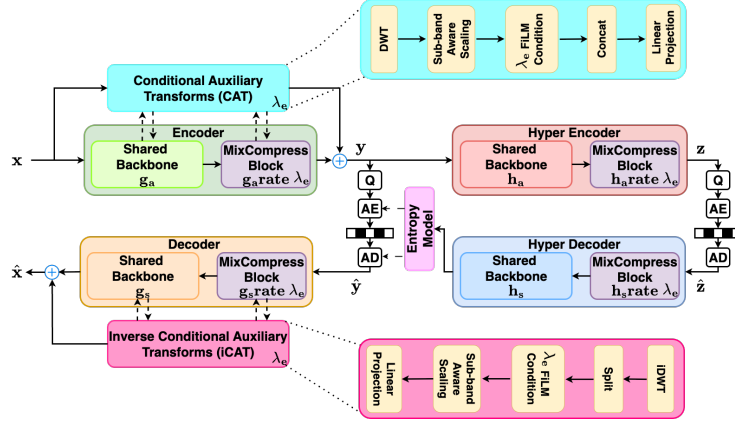


Fig. 2: Overview of the MixCompress architecture. We adapt a LIC codec and replace dense bottleneck blocks in the analysis, synthesis, and hyper-prior transforms with rate-conditioned (λ_e) MixCompress blocks. Sub-band energy compaction is further modulated through Conditional Auxiliary Transforms (CAT/iCAT), enabling dynamic adaptation across rate-distortion operating points.

Compared to prior variable-rate methods based solely on conditioning or gain modulation, MoE introduces structural specialization through conditional parameter activation. When applied to the analysis transform, synthesis transform, and hyperprior networks, MoE enables bit-rate-dependent capacity allocation throughout the compression pipeline. Extending this idea with depth-wise adaptivity (MoD) further enhances conditional computation. Together with recent advances in structured energy compaction such as AuxT, these developments motivate a unified variable-rate image compression framework.

3 Methodology

3.1 Preliminaries: Variable Rate Learned Image Compression

Variable Learned Image Compression Formulation: Modern learned image compression relies on end-to-end optimized non-linear transform coding framework [5, 6]. An input image x is mapped to a latent representation y via an analysis transform g_a , and is reconstructed as $\hat{x}$ via a synthesis transform g_s . The latent y is further mapped to a hyper-latent z through the hyper-encoder h_a . Both the latent and hyper-latent are quantized and denoted as $\hat{y}$ and $\hat{z}$ respectively. In order to entropy code $\hat{z}$ we use the standard factorized prior [6]. The hyper-decoder (h_s) then estimates the entropy parameters μ, σ from $\hat{z}$ which is refined with the space-channel context entropy model [15]. A standard single rate model optimizes a parameter set θ for the rate-distortion objective:

$$\min_{\theta} \mathcal{L}_{\lambda}(\theta) = \mathbb{E}_{x \sim p_{data}} [\lambda D(x, \hat{x}; \theta) + R(\hat{y}; \theta) + R(\hat{z}; \theta)] \quad (1)$$

where D is our distortion metric and λ is a fixed Lagrange multiplier. In order to support variable-rate control without having a set of parameters for each λ , unified frameworks sample $\lambda \sim p(\lambda)$ during training, effectively optimizing the expectation over a distribution of tasks:

$$\min_{\theta} \mathcal{L}_{VBR}(\theta) = \mathbb{E}_{\lambda \sim p(\lambda), x \sim p_{data}} [\lambda D(x, \hat{x}; \theta) + R(\hat{y}; \theta) + R(\hat{z}; \theta)] \quad (2)$$

Limitations of Dense Parameter Modulation: To adapt a shared backbone to varying λ , existing methods rely on dense parameter modulation, such as conditional convolutions [9] or latent scaling [29]. Given an embedded rate parameter λ_e , a shared feature map f is modulated via a learned affine transformation $f' = \sigma(MLP_s(\lambda_e)) * \text{Conv2D}(f) + MLP_b(\lambda_e)$ where σ is an activation function (e.g., SoftPlus).

While dense conditioning enables variable rate functionality, it consistently under-performs models optimized for single rate points. We hypothesize that this performance gap is due to a multi-objective optimization failure: Gradient Interference. Let $g_{low} = \nabla_{\theta} \mathcal{L}_{low}$ and $g_{high} = \nabla_{\theta} \mathcal{L}_{high}$ represent the gradients for updating a shared set of parameters for both low and high bit-rate targets. High bit-rate optimization requires the model preserve high-frequency structural details, whereas low bit-rate optimization strictly penalizes detail in favor of blurrier reconstructions thus lowering the bit-rate. This results in a scenario where the cosine similarity between task gradients is potentially negative [31]. When cosine similarity is negative the gradients for either task will degrade the model performance for the opposite task. Applying affine transformations in the form of conditional convolutions or latent scaling is not sufficient for resolving this conflict as this only scales the feature magnitude and will force the model to learn a suboptimal compromise between the two rate points.

3.2 Conditional Auxiliary Transforms (CAT)

Improving feature energy compaction has been shown to accelerate training and enhance performance in LIC. Auxiliary linear shortcuts, such as AuxT [21], leverage wavelet-based decompositions to offload coarse feature de-correlation and sub-band energy modulation from the main nonlinear transform, stabilizing optimization and improving convergence. However, existing shortcuts use static sub-band scaling, limiting flexibility in variable-rate LIC where the optimal energy distribution depends on the target rate: low-rate regimes benefit from stronger compaction, while high-rate regimes require finer detail preservation.

To address this, we introduce *Conditional Auxiliary Transforms (CAT)*, a rate-aware extension that conditions sub-band energy modulation on the Lagrangian multiplier λ . CAT retains the structural benefits of wavelet shortcuts while enabling adaptive energy reweighting across operating points.

Wavelet Decomposition: Given an intermediate feature map $\mathbf{P} \in \mathbb{R}^{B \times C \times H \times W}$, CAT applies a 2D Haar discrete wavelet transform (DWT) as follows:

$$\mathbf{Z} = \text{DWT}(\mathbf{P}) \in \mathbb{R}^{B \times 4C \times \frac{H}{2} \times \frac{W}{2}}. \quad (3)$$

The transform decomposes $\mathbf{P}$ into four frequency subbands

$$\{\mathbf{P}_{LL}, \mathbf{P}_{LH}, \mathbf{P}_{HL}, \mathbf{P}_{HH}\}$$

, each in $\mathbb{R}^{B \times C \times \frac{H}{2} \times \frac{W}{2}}$, corresponding to low-frequency (LL) and high-frequency (LH, HL, HH) components [21]. This is then tokenized into $\mathbf{U} = \text{reshape}(\mathbf{Z}) \in \mathbb{R}^{B \times \frac{HW}{4} \times 4C}$.

Rate-Conditioned Subband Aware Scaling: Each CAT block maintains learnable scaling vectors for each sub-band $\mathbf{s}_{\text{base}} \in \mathbb{R}^{4C}$. Let $\lambda_e \in \mathbb{R}^{d_\lambda}$ be the learned embedding of the selected rate point. A FiLM head produces rate-conditioned scaling parameters:

$$[\gamma_\lambda, \beta_\lambda] = f_{\text{film}}(\lambda_e), \quad \gamma_\lambda, \beta_\lambda \in \mathbb{R}^{4C}. \quad (4)$$

These rate-conditioned scalars are then used to update the static subband aware scaling parameters of AuxT [21] as follows: $\mathbf{s}_\lambda^{\text{enc}} = \gamma_\lambda \odot \mathbf{s}_{\text{base}} + \beta_\lambda$. Encoder-side modulation is applied token-wise: $\hat{\mathbf{U}} = \mathbf{U} \odot \exp(\mathbf{s}_\lambda^{\text{enc}})$, where $\odot$ denotes channel-wise multiplication. A learned linear projection then maps the rate-conditioned wavelet tokens to the shortcut output channels: $\mathbf{V} = \hat{\mathbf{U}} \mathbf{W}_{\text{enc}}$, $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{4C \times C'}$. Reshaping yields $\mathbf{V} \in \mathbb{R}^{B \times C' \times \frac{H}{2} \times \frac{W}{2}}$, which is injected residually at the various locations of the encoder as done in AuxT.

Decoder Symmetry: iCAT mirrors the encoder operations on the decoder side. Given decoder features $\mathbf{Y} \in \mathbb{R}^{B \times C' \times H \times W}$, we compute

$$\mathbf{U}' = \text{reshape}(\mathbf{Y} \mathbf{W}_{\text{dec}}), \quad \mathbf{W}_{\text{dec}} \in \mathbb{R}^{C' \times 4C}. \quad (5)$$

A decoder FiLM head produces $\mathbf{s}_\lambda^{\text{dec}} \in \mathbb{R}^{4C}$, and reciprocal scaling is applied: $\tilde{\mathbf{U}}' = \mathbf{U}' \odot \exp(-\mathbf{s}_\lambda^{\text{dec}})$. After reshaping to $\mathbb{R}^{B \times 4C \times \frac{H}{2} \times \frac{W}{2}}$, inverse DWT reconstructs the spatial shortcut: $\hat{\mathbf{P}} = \text{IDWT}(\tilde{\mathbf{U}}')$. CAT/iCAT preserves the plug-and-play nature of AuxT, leaving the main transform and entropy model unchanged, while elevating auxiliary branches be fully rate-aware. By conditioning subband energy modulation on λ_e , our models can dynamically adjust energy compaction across all various rate points.

3.3 Mixture of Experts

To resolve the gradient interference inherent to dense parameter modulation, we introduce MixCompress. By replacing the last layer at key bottlenecks with sparsely gated Mixture-of-Experts (MoE) modules, we decouple the optimization trajectories of competing rate points.

Sparse Structural Routing: Given an input feature f and a target rate embedding λ_e , a routing network conditioned on λ_e generates a probability distribution over N parallel expert networks $\{E_1, E_2, \dots, E_N\}$. To ensure that universally useful features get through for all rate-points we maintain a separate shared expert E_{global} . The output of the MixCompress block is computed as the sum of the shared expert and the top- K selected experts: $f' =$

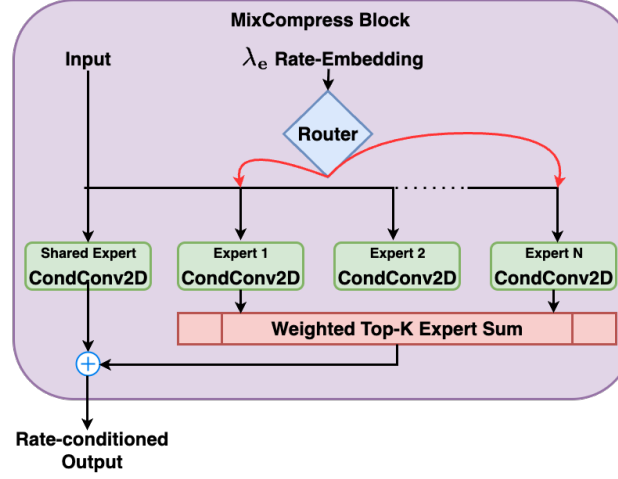


Fig. 3: The MixCompress block. The input feature is always processed by a shared expert to ensure stable information flow, and is summed with the outputs of the sparsely routed top-K experts, thus mitigating conflicting gradient updates.

$E_{global}(f) + \sum_{i \in \mathcal{T}_k} G(\lambda_e) E_i(f)$ where $\mathcal{T}_k$ is the set of top-K expert indices selected by the softmax gating mechanism $G(\lambda_e)$. In our implementation we set $K = 2$. Since our router is only conditioned on the λ_e we face the high likelihood of expert collapse [8, 26]. This results in a small subset of experts getting updated as other experts are completely ignored. To encourage balanced expert utilization we employ noise annealing and inject additive Gaussian noise ϵ into the router logits during training: $G(\lambda_e) = \text{Softmax}(\text{TopK}(\text{MLP}(\lambda_e) + \epsilon))$. This encourages the model to explore all experts and is gradually reduced to zero over time (schedule details in Sec. 4).

Task Conditioned Experts: A key architectural constraint in our framework is that both the gating networks $G(\lambda_e)$ and the CAT modules depend strictly on the rate embedding λ_e which is entirely independent from the input image x . Because these modulations are independent of the input data, they do not actually need to be computed at inference. For practical deployment, the exact expert assignment and modulation parameters can be precomputed and cached in a lookup table, introducing zero routing overhead at inference.

Gradient Conflict Mitigation: The primary advantage of this sparse structural routing is the capacity to mitigate gradient interference, a well documented optimization failure in multi-task and multi-objective learning [27, 31]. Let the parameters of a MixCompress block be partitioned into the shared expert θ_{shared} and the routed experts $\theta_{routed} = [\theta_1, \dots, \theta_N]$. For a given λ , the gradient of the loss $\mathcal{L}_\lambda$ decomposes into updates for these respective subspaces. The inner product between the gradients for two widely separated operating points λ_{low} and λ_{high} is the sum of their subspace inner products:

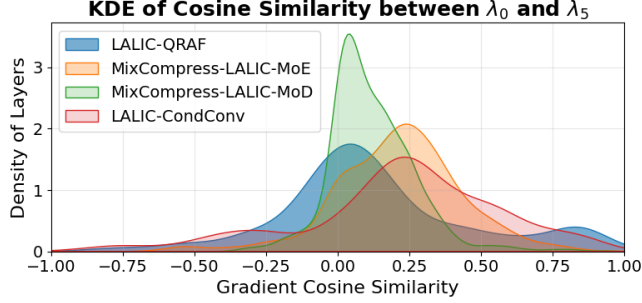


Fig. 4: Kernel Density Estimation of layer-wise gradient cosine similarity between our lowest and highest operating points λ_0 and λ_5 on 100 training images. The dense approaches (QRAF [29], CondConv [9]) exhibit long negative tails, indicating severe gradient conflict. MixCompress is able to mitigate the interference shifting the optimization dynamics towards orthogonality.

$$\begin{aligned} \langle g(\lambda_{low}), g(\lambda_{high}) \rangle &= \langle g_{shared}(\lambda_{low}), g_{shared}(\lambda_{high}) \rangle \\ &\quad + \langle g_{routed}(\lambda_{low}), g_{routed}(\lambda_{high}) \rangle \end{aligned} \quad (6)$$

The shared expert processes gradients from all operating points and is still subject to gradient conflicts. As shown in prior multi-objective optimization literature [23, 27], forcing conflicting updates through shared parameters limits the model’s ability to reach the true Pareto frontier. However, the inclusion of a routing mechanism isolates gradient conflicts within conditional capacity based on expert overlap. Let $\mathcal{T}_{low}$ and $\mathcal{T}_{high}$ be the selected experts for their respective rate point. The inner product of the routed subspace is strictly bounded by the intersection of these active experts:

$$\langle g_{routed}(\lambda_{low}), g_{routed}(\lambda_{high}) \rangle = \sum_{i \in \mathcal{T}_{low} \cap \mathcal{T}_{high}} \langle \nabla_{\theta} \mathcal{L}_{low}, \nabla_{\theta} \mathcal{L}_{high} \rangle \quad (7)$$

The shared expert θ_{shared} is still subject to potential gradient conflict but by reducing the volume of parameters subject to conflicting updates, we expect the overall interference to diminish. We empirically validate this in Fig. 4, where MixCompress shifts the gradient cosine similarity distribution towards zero, confirming that sparse structural routing mitigates gradient conflicts. Decomposing the conflict by network layers reveals that the encoder is disproportionately affected under dense modulation: 78% and 79% of the encoder layers in QRAF and CondConv respectively exhibit negative gradient cosine similarity. MixCompress-LALIC-MoE and MixCompress-LALIC-MoD reduces this to 31% and 22% respectively. This is consistent with visual quality improvements observed in high-bit rate reconstructions (Additional per layer statistics in Supplemental Material).

Mixture of Depths(MoD) Extension: While standard MoE frameworks allocate distinct parameters, maximizing fidelity at peak bit-rates often requires

additional representational capacity. We extend the proposed framework by progressively increasing expert complexity. In the MoD variant, experts will remain compatible by outputting the same tensor shapes but contain varying depths. We formulate the i -th expert as a composite function of i nested transformation blocks (in our case Bi-RWKV blocks) $E_i(f) = (h_i \circ h_{i-1} \dots h_1)(f)$. By routing high-rate samples to higher complexity experts the model can dynamically expand its representational capacity when required.

4 Experiments

4.1 Experimental Setup

We follow the training protocol of LALIC [13] and train all models on the first 400K images of OpenImages [19]. OpenImages provides diverse high-resolution natural content, making it well suited for learned image compression. During training, images are randomly cropped to 256×256 patches. All models are optimized using Adam with a batch size of 8. Rate-distortion optimization is performed under the MSE objective using the following multipliers: $\lambda \in \{0.0018, 0.0035, 0.0067, 0.0130, 0.0250, 0.0483\}$.

In contrast to single-rate baselines where each λ requires an independently trained model, in our approach we jointly learn all operating points within a single unified model by uniformly sampling one λ per mini-batch during training. This strategy ensures balanced supervision across rates while eliminating the need for multiple checkpoints. As a result, our approach substantially reduces training cost, storage overhead, and deployment complexity.

Importantly, although our model supports multiple rate points, we follow the same training schedule as LALIC and train for the same total duration as one single-rate model. Training is conducted in three stages. We first train for 40 epochs with a learning rate of 1×10^{-4} , followed by 4 epochs at 1×10^{-5} . Finally, we fine-tune the model for 4 additional epochs using 512×512 random crops to improve high-resolution reconstruction quality. All experiments are performed on a single NVIDIA H100 GPU.

4.2 Model Configuration

We integrate the proposed framework into two strong single-rate backbones, LALIC [13] and LIC-TCM [24]. To ensure a strictly controlled comparison, we preserve the original architecture of each backbone and introduce our variable-rate modules only at the end of the analysis transform, synthesis transform, and hyperprior network, as illustrated in Fig. 2. All remaining components are kept identical to their original implementations.

Variable-rate adaptation is enabled through a 32-dimensional rate embedding λ_e , which conditions routing decisions within each MixCompress block. Each block contains N routed experts with top- K sparse selection, together with one additional always-active shared expert that ensures stable information

flow. For the main experiments, we use $N = 4$ and $K = 2$ for all MixCompress variants unless specified otherwise. All experts share identical input-output tensor interfaces, enabling seamless integration without modifying the surrounding architecture. The effect of different expert counts N and routing sparsity levels K is analyzed in the supplementary material.

MixCompress Variants: We evaluate two expert configurations here:

- **MoE (Homogeneous Experts).** In both MixCompress-LALIC-MoE and MixCompress-TCM-MoE, all routed experts share the same internal architecture and differ only in parameters. This setting isolates the effect of parameter specialization under a fixed computational graph.
- **MoD (Progressive-Depth Experts).** In MixCompress-LALIC-MoD, routed experts progressively stack additional $2 \times$ Bi-RWKV [13] blocks, forming a shallow-to-deep hierarchy. In MixCompress-TCM-MoD, depth is increased more conservatively using $1 \times$ convolutional [24] blocks to control overall model size.

CAT Shortcuts: We enable CAT and iCAT in the encoder and decoder, respectively. Four conditional wavelet shortcuts are inserted symmetrically in both branches of the MixCompress architecture, providing rate-adaptive auxiliary pathways without altering the primary codec structure.

Routing Stabilization: To encourage balanced expert utilization in our rate-conditioned routers during training, we inject additive Gaussian noise into the router logits using a linearly annealed schedule, $\sigma(e) = \sigma_0 \cdot \max\left(1 - \frac{e}{E_{\text{anneal}}}, 0\right)$, where e denotes the current epoch. We use $\sigma_0 = 1.0$ and $E_{\text{anneal}} = 20$ by default. The noise is therefore large during the early epochs to promote exploration, gradually decays over training, and becomes zero after epoch 20, allowing stable expert specialization. To validate the effectiveness of this stabilization strategy, we provide comprehensive expert utilization and routing statistics in the Supplemental Material, confirming that noise annealing mitigates catastrophic routing collapse and maintains diverse expert utilization across most of the network.

4.3 Evaluation Datasets and Metrics

We evaluate on four widely used image compression benchmarks: Kodak [11] (24 images at 768×512 resolution), Tecnick [3] (100 images at 1200×1200 resolution), and the CLIC Professional validation and test sets [1], containing 41 and 60 high-resolution images (up to 2K resolution), respectively. Rate-distortion performance is measured using bits per pixel (BPP) and peak signal-to-noise ratio (PSNR). We report BD-rate savings computed using PSNR as the distortion metric and VTM-23.1 [18] as the anchor. All R-D curves are generated using the same set of λ values across all methods to ensure strict fairness.

4.4 Rate-Distortion Performance

Table 1 reports BD-rate results relative to VTM-23.1 [18]. We evaluate MixCompress against two categories of methods: (i) single-rate baselines, where each

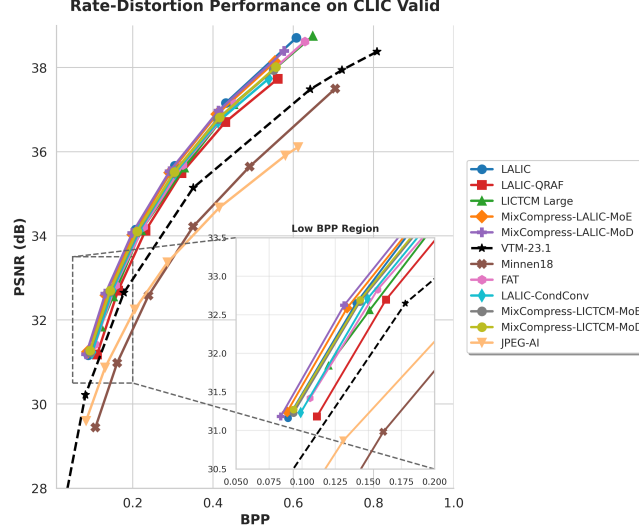


Fig. 5: RD curves demonstrating the improvement achieved by MixCompress on CLIC Validation. JPEG-AI [2] is included for completeness, though note that direct comparison is not straightforward as it is optimized for perceptual metrics rather than distortion metrics. Additional RD curves at full resolution for CLIC Validation, CLIC Test, Kodak, and Tecnick, evaluated on both PSNR and MS-SSIM, are provided in the Supplemental Material.

operating point requires an independently trained model, and (ii) variable-rate (VBR) baselines, which support multiple rate points within a single model.

Comparison with Single-Rate Models: MixCompress variants consistently match or surpass their corresponding single-rate checkpoint across all benchmarks despite being trained as a single unified model for the same total duration, where each rate point effectively sees only 1/6th of the training iterations compared to a dedicated single-rate checkpoint. On CLIC Valid, CLIC Test, and Tecnick datasets, MixCompress-LALIC-MoD achieves -25.73% , -28.71% , and -28.88% BD-rate (PSNR) reductions, exceeding the individually optimized baseline by a significant margin. Even the homogeneous MixCompress-LALIC-MoE variant improves upon single-rate LALIC on CLIC and Tecnick. On the lower-resolution Kodak dataset, MixCompress-LALIC-MoD surpasses the single-rate LALIC anchor while MixCompress-LALIC-MoE closely matches it. Notably, MixCompress-LALIC-MoD outperforms all other single-rate architectures in our comparison, including FAT [22] and LIC-TCM [24], across every benchmark while only requiring a single trained checkpoint rather than six.

Comparison with Variable-Rate Models: We re-implement two representative dense-modulation VBR strategies, QRAF [29] and Conditional Convolutions [9] on the LALIC backbone using identical training data, schedule, and evaluation protocol. Both dense VBR baselines incur substantial performance degradation. In contrast, MixCompress-LALIC-MoE effectively closes this variable-

Table 1: Rate-distortion comparison in terms of BD-Rate(PSNR)%. Results are computed using VTM-23.1 as the anchor. The **best**, **second-best**, and **third-best** results for each dataset are highlighted. Variable bit-rate (VBR) methods indicated by ✓

Method	VBR	BD-Rate (PSNR)			
		Kodak	CLIC-V	CLIC-T	Tecnick
VTM-23.1 [18]	✓	0.00	0.00	0.00	0.00
Hyper-prior [5]	✗	27.09	34.27	35.38	35.88
Minnen18 [25]	✗	6.89	2.95	0.98	3.30
Cheng20-Parallel [7]	✗	-1.77	-4.119	-5.45	-3.52
ELIC [15]	✗	-8.95	-12.22	-14.35	-15.07
TCM [24]	✗	-16.06	-18.63	-19.70	-20.41
FAT [22]	✗	-17.12	-19.07	-20.04	-23.07
LALIC [13]	✗	-18.81	-23.96	-26.43	-26.03
LALIC-QRAF	✓	-9.32	-13.22	-15.24	-14.36
LALIC-CondConv	✓	-15.12	-19.22	-22.05	-21.60
MixCompress-LALIC-MoE	✓	-18.81	-24.16	-26.80	-27.05
MixCompress-LALIC-MoD	✓	-20.12	-25.73	-28.71	-28.88

rate gap, and MixCompress-LALIC-MoD surpasses the single-rate counterpart entirely. The pattern is consistent across all four benchmarks with dense modulation methods sacrifice between 3.7-12.0 percentage points of BD-rate relative to the single-rate LALIC. We attribute MixCompress’s performance to the gradient conflict mitigation enabled by sparse structural routing, validated empirically in Table. 1. We note that earlier gain-based VBR methods [10] have been shown to underperform QRAF [29] under comparable settings; we therefore treat QRAF as the stronger dense-modulation reference and omit [10] from our comparison for brevity.



Fig. 6: Visual Comparison best viewed zoomed in. Here we cropped out a section of an image from the CLIC validation split. LALIC-QRAF at the highest quality is unable to reconstruct the cable (denoted by the red line in the original) present in the original image while the single-rate baseline LALIC as well as our multi-rate versions are capable at the same rate-point.

Visual Comparison: In Fig. 6 a localized impact of gradient conflict on high-frequency structural elements can be seen. At comparable high bit-rates, the dense QRAF [29] fails to reconstruct fine details, such as the thin cable crossing the building. This over-smoothing effect is a result of competing objectives where low bit-rate gradients favor this behavior while high bit-rate gradients

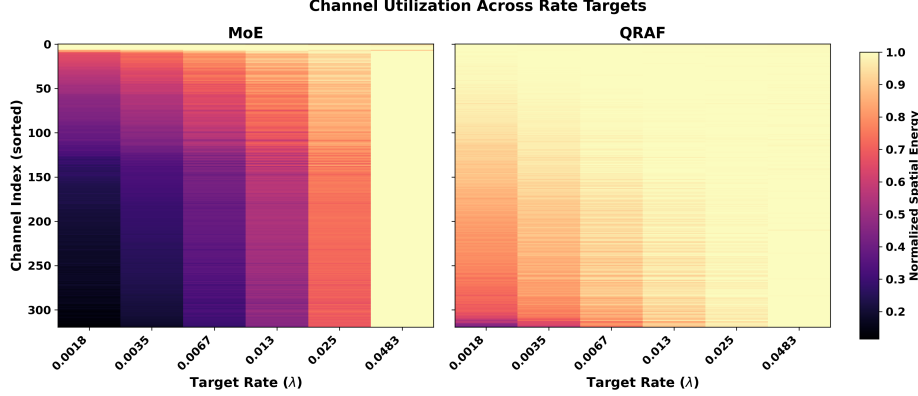


Fig. 7: Channel utilization heatmaps across different rates λ . The dense baseline QRAF [29](right) forces highly overlapping, entangled parameter usage across all rate points. In contrast, MixCompress(left) dynamically reallocates capacity, sharing channels between adjacent rates for smoother transitions between rate-points while decoupling the parameters for low and high bit-rates to minimize gradient conflict.

would want to preserve the cable. By structurally isolating these gradients both MixCompress-MoE and MixCompress-MoD are able to successfully preserve details.

MoE vs. MoD: MoD variants consistently outperform MoE, suggesting that depth-adaptive experts provide advantages beyond parameter-only specialization by adapting representational capacity to bit-rate demand. Cross-backbone generalization to LIC-TCM [24], visual comparisons, and continuous rate interpolation at unseen λ values are validated in the Supplemental Material.

4.5 Latent Analysis

Gradient Conflict Mitigation: To empirically validate our hypothesis that dense conditioning induces optimization interference, we analyze the layer-wise gradient cosine similarity between two extreme rate points (λ_{low} and λ_{high}) in Fig. 4. The dense baselines QRAF [29] and Cond-Conv [9] both exhibit long negative tails in their distributions, indicating severe gradient conflict where updates for low-rate abstraction will directly degrade parameters for high fidelity reconstruction. By structurally isolating these trajectories, MixCompress successfully shifts the dynamics towards orthogonality, heavily increasing the gradient similarity around zero and mitigating destructive interference.

Capacity Allocation: The gradient conflict mitigation is physically realized in the channel utilization strategy. As visualized in Fig. 7, a method like QRAF forces a dense, entangled parameter utilization across all target rates. In contrast, MixCompress is able to more gradually transition between rate points and more effectively decouples active parameters from low and high rate points. At the high bit-rates, QRAF exhibits signs of behavior of low bit-rate models which can be observed in the overly smoothed images. The severe under-performance of

Table 2: We report various metrics measured during inference. Note that for MoE/MoD models we report both the active number of parameters/total number of parameters. The memory usage is including all experts loaded into memory.

Method	Enc(s)	Dec(s)	Mem(G)	FLOPs(G)	Params(M)
LALIC [13]	0.234	0.184	0.875	283.39	66.13
MixCompress-MoE	0.252	0.187	1.037	311.99	77.86/88.56
MixCompress-MoD	0.257	0.190	1.262	392.02	120.22/156.55

LALIC-QRAF is attributable to two compounding failure modes specific to this backbone. QRAF adapts bit-rate by scaling the latent exclusively at the quantization boundary, meaning the sole rate-differentiating signal is quantization noise variance. However, Bi-RWKV’s LayerNorm [4] normalizes per-token variance at each block, effectively squashing this signal before it can influence downstream representations. As a result, both the entropy model and synthesis transform become rate-blind, learning a single compromise across all operating points. This incompatibility is further compounded by the gradient interference inherent to dense modulation, which remains unresolved. MixCompress avoids both failure modes by routing latents through structurally distinct pathways, producing representations that differ in direction rather than magnitude. This is further corroborated by the RD-curves (Fig. 5) where we can see QRAF has a severely limited range in both PSNR and BPP. We provide qualitative visual reconstructions demonstrating MixCompress’s superior preservation of high-frequency at higher bit-rates in the Supplemental Material.

4.6 Computational and Deployment Efficiency

We provide the computational requirements and inference latency of our framework in Table 2. Because our routers and CAT modulations are only conditioned on the rate embedding λ_e , the expert choices and modulation parameters only need to be computed once. As a result, those network parameters do not need to be kept in memory during inference. We report the optimized cached results and we can see there is only a slight increase in Encoding and Decoding time. This is significantly smaller than maintaining an ensemble of a single rate model such as LALIC [13] would require around 397 million parameters in storage. In contrast, MixCompress-MoD our largest model, covers the same rate-distortion curve with 156 million parameters with only 120 million active per pass.

5 Ablation Studies

We conduct controlled ablations on the LALIC backbone to systematically evaluate each design component of MixCompress. All variants are trained under identical settings and evaluated using the R-D protocol in Sec. 4. Table 3 reports BD-rate differences relative to the VTM-23.1 [18] anchor.

Table 3: Component-wise ablation of MixCompress. All models trained with variable rate control with 4 experts. Additional encoder only results in the Supplemental.

Configuration	Kodak	CLIC-V	CLIC-T	Tecnick
VTM-23.1 [18]	0	0	0	0
Variable-rate (Cond-Conv)	-15.15	-19.27	-22.09	-21.53
+ AuxT [21]	-17.06	-20.76	-23.37	-23.50
+ CAT	-17.89	-21.92	-24.23	-24.41
+ MoE (N=4, K=2)	-18.81	-24.16	-26.80	-27.05
+ MoD (N=4, K=2)	-20.12	-25.73	-28.71	-28.88

We begin with a unified variable-rate model implemented via conditional convolution (Cond-Conv), where convolutional weights are modulated by the rate embedding λ_e . Although this dense modulation strategy leads to clear performance improvement over VTM-23.1 on all benchmark datasets, this simple weight modulation is insufficient to preserve the performance of independently trained single-rate models from Table. 1.

Introducing AuxT shortcuts partially mitigates this gap, reducing the BD-rate loss relative to the single-rate baseline and indicating that frequency-domain shortcuts help stabilize parameter sharing across rates. Replacing AuxT with our rate-aware CAT transforms further improves performance, showing that explicit rate-conditioned subband modulation is more effective than static auxiliary transforms for variable-rate modeling.

Next, we replace the Cond-Conv layers with our MoE blocks. This narrows the performance gap to the single-rate models on Kodak and surpasses them on CLIC and Tecnick (Table. 1). This depicts that sparse expert specialization is more effective than dense weight modulation for variable-rate LIC. Finally, utilizing our MoD blocks consistently yields the strongest BD-Rate performance across all datasets. By enabling routing to control both parameter selection and representational capacity, MixCompress-MoD effectively adapts model capacity across rates and outperforms the independently trained single-rate baselines.

6 Conclusion

We propose MixCompress, a unified variable-rate learned image compression framework that overcomes the gradient interference inherent to dense parameter modulation. By integrating sparsely gated Mixture-of-Experts (MoE) and Mixture-of-Depth (MoD) modules, alongside Conditional Auxiliary Transforms (CAT), our architecture dynamically partitions network capacity based on the target rate λ . This structural specialization successfully isolates competing rate-distortion objectives allowing for more efficient variable-rate models. Extensive evaluations on Kodak, CLIC, and Tecnick datasets demonstrate that MixCompress consistently matches and often surpasses individually trained single-rate baselines. By demonstrating that conditional structural allocation is fundamentally more effective than dense weight modulation, MixCompress establishes a new Pareto frontier for practical, multi-rate neural image compression.

References

1. Workshop and Challenge on Learned Image Compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2021) [10](#)
2. Alshina, E., Ascenso, J., Ebrahimi, T.: Jpeg ai: The first international standard for image coding based on an end-to-end learning-based approach. *IEEE MultiMedia* **31**(4), 60–69 (2024). <https://doi.org/10.1109/MMUL.2024.3485255> [11](#)
3. Asuni, N., Giachetti, A.: TESTIMAGES: a Large-scale Archive for Testing Visual Devices and Basic Image Processing Algorithms. In: Giachetti, A. (ed.) STAG. pp. 63–70. Eurographics (2014) [10](#)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016) [14](#)
5. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-End Optimized Image Compression. In: International Conference on Learning Representations (ICLR) (2017) [1](#), [3](#), [4](#), [12](#)
6. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational Image Compression With a Scale Hyperprior. In: International Conference on Learning Representations (ICLR) (2018) [1](#), [3](#), [4](#)
7. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned Image Compression with Discretized Gaussian Mixture Likelihoods and Attention Modules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7939–7948 (2020) [1](#), [12](#)
8. Chi, Z., Dong, L., Huang, S., Dai, D., Ma, S., Patra, B., Singhal, S., Bajaj, P., Song, X., Mao, X.L., et al.: On the Representation Collapse of Sparse Mixture of Experts. *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 34600–34613 (2022) [7](#)
9. Choi, Y., El-Khamy, M., Lee, J.: Variable Rate Deep Image Compression with a Conditional Autoencoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) [2](#), [3](#), [5](#), [8](#), [11](#), [13](#)
10. Cui, Z., et al.: Asymmetric Gained Deep Image Compression with Continuous Rate Adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) [2](#), [3](#), [12](#)
11. Eastman Kodak Company: Kodak Lossless True Color Image Suite (PhotoCD PCD0992). Online (1993), image dataset [10](#)
12. Fedus, W., Zoph, B., Shazeer, N.: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* **23**(120), 1–39 (2022) [3](#)
13. Feng, D., Cheng, Z., Wang, S., Wu, R., Hu, H., Lu, G., Song, L.: Linear Attention Modeling for Learned Image Compression. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 7623–7632 (2025) [1](#), [2](#), [3](#), [9](#), [10](#), [12](#), [14](#)
14. Fleig, E., Geistert, J., Bochinski, E., Jongbloed, R., Sikora, T.: Edge-Aware Autoencoder Design for Real-Time Mixture-of-Experts Image Compression. In: 2023 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–5. IEEE (2023) [3](#)
15. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: 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 (CVPR). pp. 5718–5727 (2022) [1](#), [4](#), [12](#)

16. He, D., Zheng, Y., Sun, B., Wang, Y., Qin, H.: Checkerboard Context Model for Efficient Learned Image Compression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) [1](#)
17. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive Mixtures of Local Experts. *Neural Computation* **3**(1), 79–87 (1991) [3](#)
18. JVET: VVC Test Model (VTM) Reference Software. Fraunhofer Heinrich-Hertz-Institut (2024), version 23.1. Available at: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM [10](#), [12](#), [14](#), [15](#)
19. Krasin, I., Duerig, T., Alldrin, N., Veit, A., Abu-El-Haija, S., Belongie, S., Cai, D., Feng, Z., Ferrari, V., Gomes, V., Gupta, A., Narayanan, D., Sun, C., Chechik, G., Murphy, K.: OpenImages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from <https://github.com/openimages> (2017) [9](#)
20. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=qrwe7XHTmYb> [3](#)
21. Li, H., Li, S., Dai, W., Cao, M., Kan, N., Li, C., Zou, J., Xiong, H.: On Disentangled Training for Nonlinear Transform in Learned Image Compression. In: The Thirteenth International Conference on Learning Representations (ICLR) (2025), <https://openreview.net/forum?id=U67JQNTzo> [3](#), [5](#), [6](#), [15](#)
22. Li, H., Li, S., Dai, W., Li, C., Zou, J., Xiong, H.: Frequency-Aware Transformer for Learned Image Compression. In: The Twelfth International Conference on Learning Representations (ICLR) (2024), <https://openreview.net/forum?id=HKGQDDTuvZ> [1](#), [2](#), [3](#), [11](#), [12](#)
23. Liu, B., Liu, X., Jin, X., Stone, P., Liu, Q.: Conflict-averse Gradient Descent for Multi-task Learning. *Advances in Neural Information Processing Systems (NeurIPS)* **34**, 18878–18890 (2021) [8](#)
24. Liu, J., Sun, H., Katto, J.: Learned Image Compression with Mixed Transformer-CNN Architectures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14388–14397 (2023) [1](#), [2](#), [3](#), [9](#), [10](#), [11](#), [12](#), [13](#)
25. Minnen, D., Ballé, J., Toderici, G.D.: Joint Autoregressive and Hierarchical Priors for Learned Image Compression. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2018) [1](#), [3](#), [12](#)
26. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A.S., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021) [3](#), [7](#)
27. Sener, O., Koltun, V.: Multi-task Learning as Multi-objective Optimization. *Advances in Neural Information Processing Systems (NeurIPS)* **31** (2018) [7](#), [8](#)
28. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In: International Conference on Learning Representations (ICLR) (2017) [3](#)
29. Tong, K., Wu, Y., Li, Y., Zhang, K., Zhang, L., Jin, X.: QVRF: A Quantization-Error-Aware Variable Rate Framework for Learned Image Compression. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1310–1314. IEEE (2023) [2](#), [3](#), [5](#), [8](#), [11](#), [12](#), [13](#)
30. Yang, F., Herranz, L., Van De Weijer, J., Guitián, J.A.I., López, A.M., Mozerov, M.G.: Variable Rate Deep Image Compression with Modulated Autoencoder. *IEEE Signal Processing Letters* **27**, 331–335 (2020) [2](#), [3](#)

31. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient Surgery for Multi-task Learning. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 5824–5836 (2020) 5, 7
32. Zhai, Y., Ma, Y., Tang, L., Jiang, W., Wang, R.: Deepfgs: Fine-grained scalable coding for learned image compression. In: *2025 Data Compression Conference (DCC)*. pp. 263–272 (2025). <https://doi.org/10.1109/DCC62719.2025.000343>