FracQuant: Rank-Adaptive Image Super-resolution Network Quantization via Fractal Complexity Assessment

Chia-Ming Lee [1 2 3] **Yu-Fan Lin** [2 3] **Wen-Hsin Tsai** [3] **Ming-Ching Chang** [2] **Xin Li** [2] **Chih-Chung Hsu** [1 3]

Abstract

Deploying image super-resolution (SR) models on edge devices necessitates aggressive quantization. However, current content-aware methods rely on gradient-based complexity estimation which often fails in both smooth and heavily textured regions. These failures lead to unstable bit assignments and denormalization overflow at extreme low-bit settings. The fundamental issue is that gradient magnitudes do not convey information about the intrinsic rank of local image representations, which is the primary quantity governing regional compressibility. We propose FracQuant, a rank-adaptive SR quantization framework that estimates local intrinsic rank through fractal complexity assessment. Fractal dimension provides a geometry-driven and activation-insensitive proxy for rank, yielding a complexity signal whose raw coefficient of variation is up to 30 times lower than gradient-based alternatives. By integrating geometric self-similarity assessment and Dynamic Hybrid soft-hard Labeling (DHL), FracQuant achieves principled rank-adaptive bit allocation within a single-pass post-training quantization (PTQ) pipeline. Experimental results demonstrate that FracQuant matches or exceeds the performance of quantization-aware training (QAT) methods under standard W4A4 and W8A6 settings without requiring ground-truth supervision. Furthermore, FracQuant is the first method to remain numerically stable under W2A2 quantization on SR networks, a configuration where previous methods consistently overflow.

[1]National Yang Ming Chiao Tung University [2]University at Albany, State of New York [3]National Cheng Kung University. Correspondence to: Chih-Chung Hsu <cchsu@gs.ncku.edu.tw>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 267, 2026. Copyright 2026 by the author(s).

1. Introduction

Deep neural networks have revolutionized image super-resolution (SR) (Dong et al., 2014), achieving remarkable advancements in reconstruction quality. However, this progress comes at a steep cost: models such as EDSR (Lim et al., 2017) and RDN (Zhang et al., 2018) demand more than 40 GFLOPs and 10 MB of memory for a single 256×256 input, far exceeding the capabilities of resource-constrained edge devices. Network quantization directly addresses this by reducing memory demand by as much as 75% through lowering the precision of weights and activations, without altering the network architecture. To further improve efficiency, lightweight SR approaches have emerged via NAS (Chu et al., 2021), model compression (Sun et al., 2024; Hou & Kung, 2022), network pruning (Jiang et al., 2021), and knowledge distillation (Gao et al., 2018).

A key observation underlying our work is that natural image patches reside on *low-dimensional manifolds*. Patch matrices extracted from locally coherent image regions exhibit strong low-rank structure — a direct consequence of spatial self-similarity that has driven advances in matrix completion (Candès & Recht, 2009), tensor decomposition (Kolda & Bader, 2009), and parameter-efficient fine-tuning for large models (Hu et al., 2022). This intrinsic low-rank prior suggests a principled criterion for adaptive bit allocation: regions whose local patch representations are inherently low-rank require fewer bits to faithfully encode, while structurally complex, high-rank regions demand more. As empirically validated in Fig. 2, box-counting fractal dimension $\mathcal{D}_{\text{box}}$ exhibits a moderate Spearman correlation ($\rho = 0.549$) with local effective rank, motivating it as a geometry-driven rank proxy. Despite its theoretical appeal, this low-rank perspective has been entirely absent from content-aware quantization.

However, quantizing SR networks with minimal information loss presents unique challenges due to their predominantly variant activation distributions during inference, often leading to severe quantization errors. Recent works (Li et al., 2020; Qin et al., 2023; Hong & Lee, 2023; Zhong et al., 2022; Hong et al., 2022b; Liu et al., 2021) have attempted to address these challenges through various approaches. Some methods employ more fine-grained quantization func-

Method	Adaptive	Low-bit	Data	GT	BS	Iterations	Time
EDSR (Lim et al., 2017)	-	-	800	✓	16	150,000	30 hrs
PAMS (Li et al., 2020)	✗	✗	800	✓	16	15,000	2 hrs
DDTB (Zhong et al., 2022)	✗	✓	800	✓	16	30,000	4 hrs
PTQ4SR (Tu et al., 2023b)	✗	✗	100	✗	2	500	124 sec
ODM (Tu et al., 2023b)	✗	✓	800	✗	8	600,000	2.4 hrs
CADyQ (Hong et al., 2022a)	✓	✗	800	✓	16	200,000	40 hrs
CABM (Tian et al., 2023)	✓	✗	800	✓	16	350,000	70 hrs
AdaBM (Hong & Lee, 2024)	✓	✗	100	✗	2	500	**71 sec**
Granular-DQ (Wang et al., 2025)	✓	✗	100	✗	2	500	286 sec
FracQuant (Ours)	✓	✓	100	✗	2	500	**75 sec**

Table 1. **Existing methods for quantization on SR.** Adaptive denotes whether bit-widths are adaptively allocated, GT denotes the requirement for ground-truth HR images, and BS denotes the batch size. Metrics are reported for quantizing EDSR (×4).

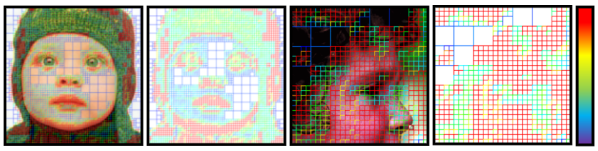

Figure 1. **Rank-adaptive bit mapping via fractal priors.** Fractal dimension serves as a geometry-driven proxy for local intrinsic rank, enabling principled bit allocation: low-rank, self-similar regions are assigned fewer bits, while high-rank, structurally complex regions receive finer precision.

tions (Li et al., 2020; Zhong et al., 2022), while others focus on dynamically updating quantization ranges (Hong et al., 2022b). While promising, these methods primarily focus on the quantization function itself rather than adaptively allocating computational resources based on image content complexity.

This limitation has led to the development of content-aware quantization approaches (Hong et al., 2022a; Tian et al., 2023; Hong & Lee, 2024), inspired by content-aware accelerating SR methods (Kong et al., 2021; Yu et al., 2021; Wang et al., 2022). Such methods fall into two categories: Post-Training Quantization (PTQ) (Hong & Lee, 2024) and Quantization-Aware Training (QAT) (Hong et al., 2022a; Tian et al., 2023). The recently proposed AdaBM (Hong & Lee, 2024) addresses the efficiency challenge through on-the-fly bit mapping, achieving results comparable to QAT methods within a compact calibration budget. Granular-DQ (Wang et al., 2025) further proposes patch-wise, layer-invariant quantization via multi-granularity clues, yet its complexity signal remains purely data-driven without exploiting intrinsic geometric structure (see Table 1).

However, all of these methods derive complexity from spatial gradient magnitudes, which carry no information about the *intrinsic rank* of local representations — the very quantity that governs regional compressibility. In smooth patches, near-zero gradients produce ill-conditioned complexity estimates, causing the bit-mapping function to assign inconsistent bit-widths and ultimately triggering denormalization overflow at 2-bit precision. In heavily textured regions, gradient saturation causes the opposite failure: overestimation

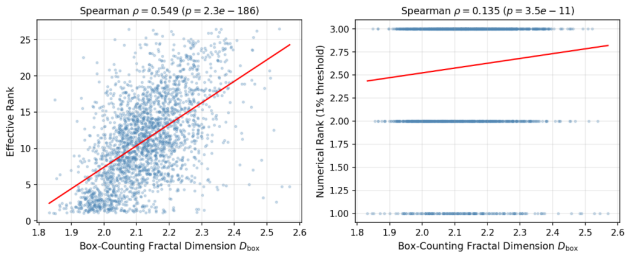

Figure 2. **Correlation between $\mathcal{D}_{\text{box}}$ and effective rank.** $\mathcal{D}_{\text{box}}$ exhibits a moderate Spearman correlation with global effective rank ($\rho = 0.549$, left) and a weak one with numerical rank at 1% threshold ($\rho = 0.135$, right); we adopt it as a geometry-driven rank proxy.

of complexity and wasteful bit allocation. Together, these failure modes make gradient-based methods fundamentally incompatible with W2A2 quantization, as confirmed experimentally in Table 1.

To address these limitations, we propose **FracQuant**, a rank-adaptive SR quantization framework that estimates the intrinsic rank of local image representations via fractal complexity assessment, enabling principled rank-adaptive bit allocation. Natural images exhibit recursive, hierarchical, and self-similar geometric structures across scales; fractal dimension quantifies the rate at which the effective rank of local patch matrices grows with scale, providing a geometry-driven, activation-insensitive complexity proxy, as illustrated in Fig. 1 and supported in Fig. 2. Our key contributions are:

- A **fractal complexity assessment** module via box-counting and multi-scale feature extraction that approximates the *intrinsic rank* of local image representations, yielding a 30× more stable raw complexity signal than gradient-based methods (Fig. 2).
- A **geometric self-similarity assessment** module that quantifies the low-rank structure of patch similarity matrices across scales and orientations, refining bit allocation for self-similar and recurring patterns (Fig. 1).
- A **Dynamic Hybrid soft-hard Labeling (DHL)** mechanism that stabilizes bit assignment near decision boundaries, enabling the *first* numerically stable W2A2 quantization on SR networks, a setting where all prior methods including AdaBM (Hong & Lee, 2024) and Granular-DQ (Wang et al., 2025) overflow (Table 1).

2. Related Works

Single Image Super-Resolution. Deep learning has driven rapid progress in SISR, from early CNN baselines (Dong et al., 2014) to near-photorealistic architectures (Lim et al., 2017; Zhang et al., 2018). Despite their quality, models such as EDSR require over 40 GFLOPs for a 256 × 256 input, motivating lightweight designs via NAS (Chu et al.,

2021), pruning (Jiang et al., 2021), distillation (Gao et al., 2018), and quantization. Transformer-based methods improve reconstruction quality by modeling long-range dependencies, with SwinIR (Liang et al., 2021) establishing a strong baseline and DRCT (Hsu et al., 2024) further enriching representations via dense feature reuse while reducing model parameters. However, quadratic attention complexity remains a deployment bottleneck, motivating efficient attention and model compression.

Network Quantization for Image SR. Quantization maps high-precision weights and activations to low-bit representations via PTQ (Choukroun et al., 2019; Jacob et al., 2018) or QAT (Bengio et al., 2013). SR networks pose unique challenges due to strongly asymmetric activation distributions (Zhong et al., 2022; Hong et al., 2022b; Li et al., 2020). DDTB (Zhong et al., 2022) addresses this with dynamic threshold adaptation, DAQ (Hong et al., 2022b) with distribution-aware range updates, and PTQ4SR (Tu et al., 2023b) with an SR-specific calibration protocol.

Content-Aware Adaptive Quantization. Routing patches by complexity has proven effective for SR acceleration (Kong et al., 2021; Yu et al., 2021; Wang et al., 2022). CADyQ (Hong et al., 2022a) and CABM (Tian et al., 2023) extend this to QAT by assigning higher bit-widths to texturally complex regions, while AdaBM (Hong & Lee, 2024) achieves comparable quality via PTQ by decoupling layer-wise and image-wise sensitivity. Granular-DQ (Wang et al., 2025) further proposes patch-wise, layer-invariant quantization via multi-granularity and entropy-based clues, yet its complexity signal remains purely data-driven without exploiting intrinsic geometric structure. All of these methods derive complexity from spatial gradients, which suffer from high variance in smooth and saturated regions and overflow at W2A2. Our work directly targets this limitation.

Fractal Analysis in Image Processing. Fractals (Mandelbrot, 1967; 1985), characterized by self-similarity across scales, have driven advances in image compression (Barnsley & Hurd, 1993; Fisher, 1995; Wu et al., 2020; Bouda et al., 2016) via IFS and box-counting, and in SR reconstruction (He et al., 2021; Xu et al., 2013) via multi-scale self-similarity. Recent generative models (Li et al., 2025; Argilaga, 2023; Tu et al., 2023a) further show that fractal hierarchical priors yield more realistic textures. We are the first to use fractal dimension as a *complexity metric* for adaptive bit-width allocation in SR quantization.

Low-Rank Representations and Compression. Low-rank factorizations underpin compression across domains, from Tucker/CP decompositions for network weights (Kolda & Bader, 2009; Denton et al., 2014) to LoRA (Hu et al., 2022) for parameter-efficient fine-tuning. Matrix completion theory (Candès & Recht, 2009) formalizes exact recovery of low-rank matrices from far fewer observations, echoing the self-similarity structure that fractal dimension captures. FracQuant operationalizes this connection: smooth, self-similar patches yield low-rank representations and require fewer bits, while high-rank regions demand more. By using fractal dimension as a proxy for local intrinsic rank, we achieve principled rank-adaptive bit allocation within the SR quantization pipeline.

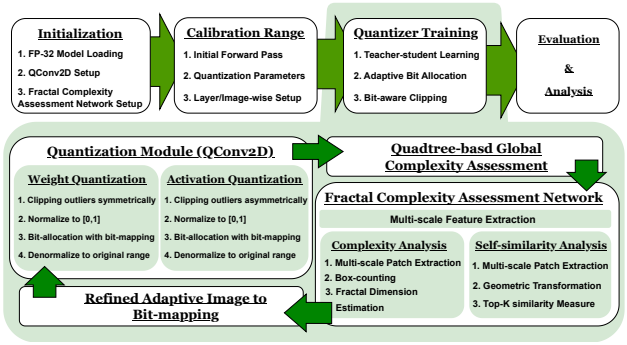

Figure 3. **End-to-end FracQuant quantization pipeline.** From a full-precision model, FracQuant proceeds through four stages: *initialization* (FP-32 loading, QConv2D and fractal-assessment setup), *calibration* (a forward pass with per-layer and per-image quantization-parameter estimation), *quantizer training* (teacher–student learning with adaptive bit allocation and bit-aware clipping), and *evaluation*. Within each QConv2D, the fractal complexity and geometric self-similarity assessments drive the refined adaptive image-to-bit mapping, while a quadtree-based global assessment keeps bit allocation content-aware.

3. Proposed Method

As illustrated in Fig. 3, FracQuant augments a frozen full-precision SR network by attaching a fractal complexity assessment network to each quantized convolution, proceeding through initialization, calibration, and quantizer training before evaluation. The framework centers on a per-image complexity signal that drives spatially adaptive bit allocation by fusing a fractal complexity score, a geometric self-similarity estimate (Sec. 3.3), and a quadtree-weighted gradient term into a refined adaptive image-to-bit mapping. The following sections detail each component in turn.

3.1. Fractal Complexity Assessment Network

Natural images exhibit rich structural patterns at multi-scales with variations in orientation and position. Traditional gradient-based complexity measures (Hong et al., 2022a; Tian et al., 2023; Hong & Lee, 2024) often fail to capture hierarchical and recursive fractal patterns (Mandelbrot, 1967; Li et al., 2025). This limitation becomes particularly apparent when quantizing SR networks, where precise representation of texture details is crucial for high-quality reconstruction.

We propose a fractal complexity assessment framework decomposed into three modules: (i) **multi-scale feature**

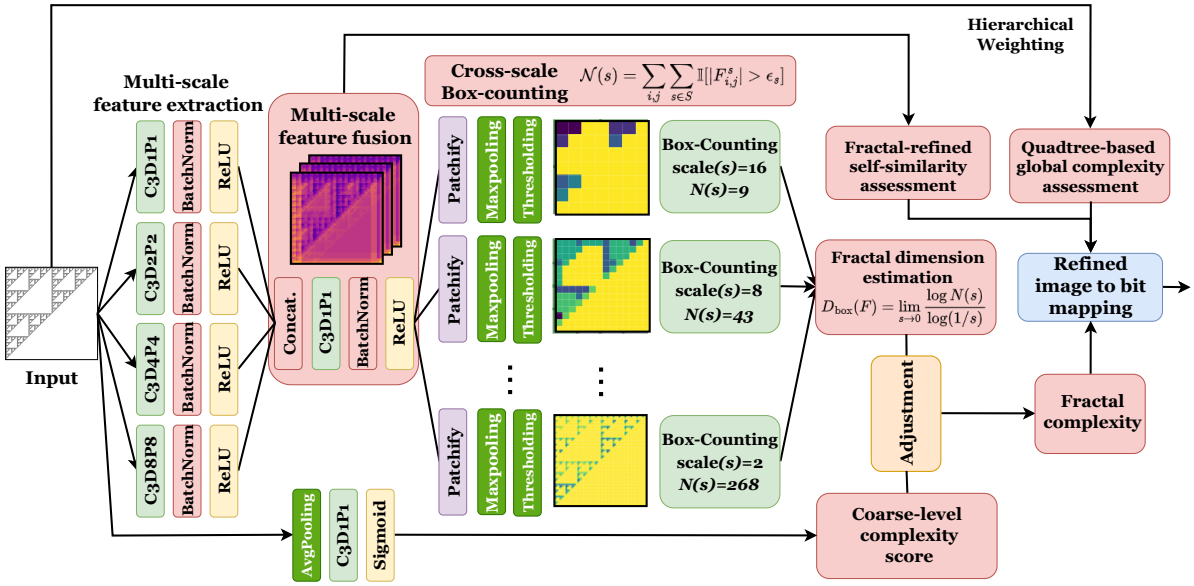

Figure 4. **Overall complexity assessment pipeline of FracQuant.** Multi-scale feature extraction and fusion are followed by cross-scale box-counting, which estimates fractal dimension $\mathcal{D}_{\text{box}}$ as a geometry-driven proxy for local intrinsic rank. The resulting fractal complexity score is combined with quadtree-weighted gradient complexity and geometric self-similarity assessment to produce a stable, rank-adaptive bit allocation signal.

extraction, (ii) **fractal dimension estimation** for capturing fine-scale textures, and (iii) **geometric self-similarity assessment** for recursive pattern analysis.

Multi-scale feature extraction. It captures the inherent self-similarity patterns across different scales, a fundamental property of fractals. Using dilated convolutions with rates 2^i:

$$F^{(i)} = \text{DilConv}(F_{\text{in}},\, r = 2^i), \quad i = 0, 1, 2, 3, \tag{1}$$

the network observes the image at multiple resolutions without increasing parameter count, detecting repeating structures invisible at a single scale. Feature fusion then preserves scale-specific information while enabling cross-scale interaction.

3.2. Fractal Dimension Estimation

Fractal Dimension and Intrinsic Rank

A set $\mathcal{F} \subset \mathbb{R}^d$ has *box-counting dimension* (Minkowski dimension) (Mandelbrot, 1967; 1985):

$$\dim_B(\mathcal{F}) = \lim_{s \to 0} \frac{\log \mathcal{N}(s)}{\log(1/s)},$$

where $\mathcal{N}(s)$ counts the minimum boxes of side s needed to cover $\mathcal{F}$. For a patch matrix $\mathbf{P} \in \mathbb{R}^{d \times N}$, effective rank scales approximately as $\text{rank}_s(\mathbf{P}) \propto s^{-\dim_B}$ (Candès & Recht, 2009): a lower fractal dimension implies a lower-dimensional subspace occupation, i.e., the patch is inherently more compressible:

$$\mathcal{D}_{\text{box}} \text{ low} \implies \text{rank}(\mathbf{P}) \text{ low} \implies \text{fewer bits required.}$$

The complexity of image patterns often manifests at different scales, making traditional single-scale analysis insufficient. We employ fractal dimension analysis via box-counting (Wu et al., 2020; Bouda et al., 2016) (Fig. 4), which partitions the feature space into boxes of varying sizes to provide a comprehensive multi-scale assessment. The estimated fractal dimension $\mathcal{D}_{\text{box}}$ quantifies the relationship between measurement scale and pattern coverage:

$$\mathcal{D}_{\text{box}}(I_{LR}) \approx -\frac{\log \mathcal{N}(S_{\text{max}}) - \log \mathcal{N}(S_{\text{min}})}{\log(S_{\text{max}}) - \log(S_{\text{min}})}, \tag{2}$$

where $\mathcal{N}(s)$ is the number of boxes of size s required to cover the feature structure (Mandelbrot, 1967; Li et al., 2025) across scale set S. To ensure computational efficiency and rotation-invariance, we utilize dyadic scales and incorporate rotation-invariant counting:

$$\mathcal{N}(s) = \sum_{i,j} \sum_{s \in S} \mathbb{I}[|F^s_{i,j}| > \epsilon_s], \tag{3}$$

where ϵ_s is a scale-dependent threshold after max-pooling, and $F^s_{i,j}$ is the feature map partitioned by scale s at coordinate (i, j).

Coarse-level complexity assessment. While $\mathcal{D}_{\text{box}}$ captures geometric self-similarity structure, it does not encode global channel-wise feature statistics. A complementary coarse-level score addresses this:

$$\mathcal{C}_{\text{base}} = \sigma(W_p \cdot \text{BN}(W_c \cdot \text{AvgPool}(F_{\text{fused}}) + b_c) + b_p), \tag{4}$$

where average pooling compresses spatial dimensions into

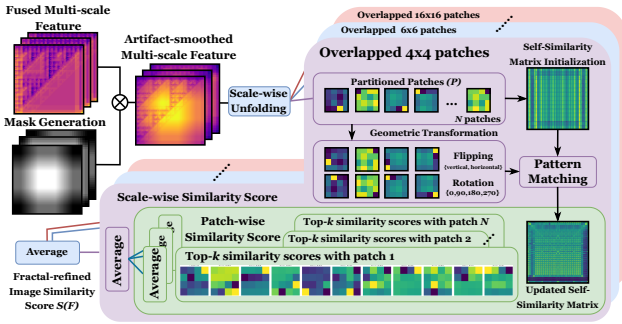

Figure 5. **Illustration of fractal-refined self-similarity assessment.** By rotating partitioned patches containing multi-scale information, a more comprehensive assessment is achieved compared to searching for self-similarity in the original image.

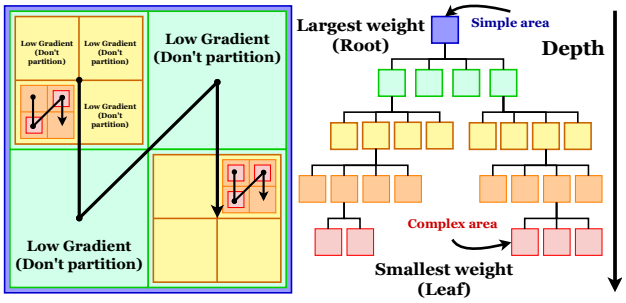

Figure 6. **Quadtree-based global complexity assessment.** Root nodes represent smooth, low-gradient regions and receive larger weights, while leaf nodes capture fine-grained textured regions with smaller weights. This hierarchical weighting prevents homogeneous regions from dominating the complexity estimate while preserving sensitivity to high-frequency details.

a channel-wise summary. The final fractal complexity integrates both signals:

$$\mathcal{C}_{\text{fractal}}(I_{LR}) = \mathcal{C}_{\text{base}} \cdot (1 + \mathcal{D}_{\text{box}}(I_{LR})), \tag{5}$$

where $\mathcal{D}_{\text{box}}$ acts as a geometry-driven multiplicative modulator, amplifying the base score in proportion to local structural complexity.

3.3. Geometric Self-similarity Assessment

Natural images contain recurring patterns across scales and orientations, a property directly linked to low-rank structure: self-similar patches yield nearly rank-deficient patch matrices, implying fewer bits suffice for faithful encoding (Candès & Recht, 2009). We exploit this via a hierarchical self-similarity framework that evaluates pattern repetition under geometric transformations (Fig. 5):

$$T(P) = \mathcal{G}(P, \theta, \sigma), \tag{6}$$

where $\mathcal{G}_\theta$ rotates by $\theta \in \{0, 90, 180, 270\}$ or flips horizontally or vertically, and σ controls patch-scale. Self-similarity is then measured via:

$$\mathcal{S}(I_{LR}) = \frac{1}{Z} \sum_{i,j} \text{Top-}\mathcal{K}_{k \neq i}(\text{cos-sim}(P_i, T(P_j))), \tag{7}$$

where patches P_i are extracted across multiple scales and patch sizes, and Z ensures consistent comparison between feature maps. A higher $\mathcal{S}$ indicates stronger self-similarity, lower intrinsic rank, and thus greater compressibility.

Low-rank view of self-similarity. The patch similarity matrix underlying Eq. (7) is $\mathbf{S} = \mathbf{P}^\top \mathbf{P} \in \mathbb{R}^{N \times N}$. When image content is self-similar, patches are nearly linearly dependent, so $\text{rank}(\mathbf{S}) \ll N$ (Candès & Recht, 2009). Thus $\mathcal{S}(I_{LR})$ provides an *algebraic* estimate of intrinsic rank, complementing the *geometric* estimate from $\mathcal{D}_{\text{box}}$: both signals agree that high self-similarity implies low rank implies fewer required bits.

3.4. Refined Adaptive Image-to-Bit Mapping

Complexity-based image-to-bit mapping. We propose a hybrid complexity measure integrating gradient and fractal signals to overcome the instability of gradient-only estimation in smooth and saturated regions:

$$\mathcal{C}(I_{LR}) = \alpha \cdot \mathcal{C}_{\text{quad}}(I_{LR}) + (1 - \alpha) \cdot \mathcal{C}_{\text{frac}}(I_{LR}), \tag{8}$$

where $\mathcal{C}_{\text{quad}}$ applies a Quadtree-based weighting to gradient complexity (Fig. 6):

$$\mathcal{C}_{\text{quad}} = \frac{\sum_{i=1}^{N} \mathcal{C}_{\text{grad}}(r_i) \cdot w(d(r_i))}{\sum_{i=1}^{N} w(d(r_i))}, \tag{9}$$

where N is the total number of partitioned regions, $\mathcal{C}_{\text{grad}}(r_i)$ is the complexity score of region r_i, $d(r_i)$ is its depth in the Quadtree hierarchy, and $w(d) = \frac{1}{2^d}$ assigns higher importance to larger, smoother regions near the root. We set $\alpha = 0.1$ (Sec. 4.1), reflecting that fractal complexity is the dominant signal.

Dynamic Hybrid soft-hard Labeling (DHL). The discrete nature of conventional bit allocation leads to suboptimal performance when complexity values fall near threshold boundaries. DHL combines a differentiable soft assignment with a discrete hard assignment during training, enabling nuanced bit decisions while preserving inference efficiency. The soft assignment is:

$$b_{\text{soft}} = \tanh\left(\frac{\mathcal{C}(I_{LR}) - (u'_{\texttt{i2b}} + l'_{\texttt{i2b}})/2}{(u'_{\texttt{i2b}} - l'_{\texttt{i2b}})/2}\right), \tag{10}$$

where $l'_{\texttt{i2b}}$ and $u'_{\texttt{i2b}}$ are learnable parameters defining the complexity range. The discrete assignment is:

$$b_{\text{hard}} = \begin{cases} 0.0, & \mathcal{C}(I_{LR}) < l'_{\texttt{i2b}} \\ +0.5, & l'_{\texttt{i2b}} \leq \mathcal{C}(I_{LR}) \leq u'_{\texttt{i2b}} \\ +1.0, & \mathcal{C}(I_{LR}) > u'_{\texttt{i2b}} \end{cases} \tag{11}$$

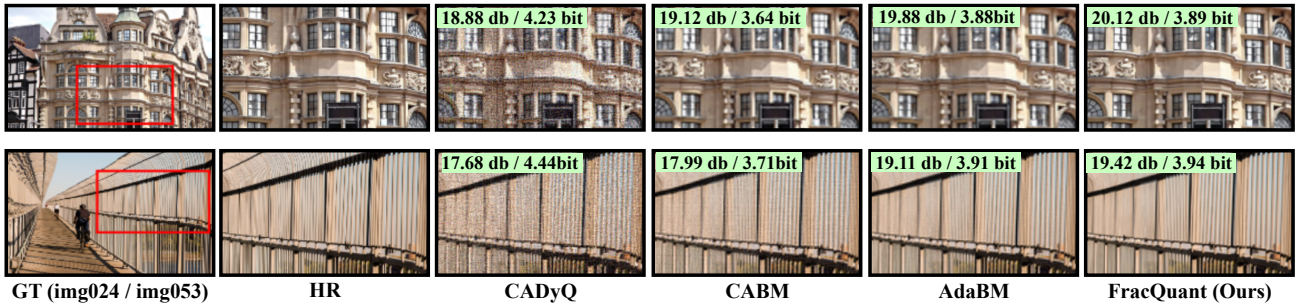

Figure 7. **Qualitative results on Urban100 with W4A4 EDSR.** FracQuant retains sharper edges and finer structural details than gradient-based methods, suggesting that rank-adaptive bit allocation better preserves high-frequency structures under aggressive quantization.

A straight-through estimator (Bengio et al., 2013) reconciles both for end-to-end training:

$$b'_I = b_{\text{soft}} - \mathcal{E}(b_{\text{soft}}) + \mathcal{E}(b_{\text{hard}}), \tag{12}$$

where $\mathcal{E}(\cdot)$ denotes the gradient detachment operation. DHL further incorporates the self-similarity score from Sec. 3.3 to modulate the final bit assignment:

$$b^*_I = \texttt{I2B}(I_{LR}) = b_I\big(1 + \gamma\mathcal{S}(I_{LR})\big)\big(1 - \eta\,\mathcal{E}(\mathcal{S}(I_{LR}))\big), \tag{13}$$

with $\gamma = 0.4$ and $\eta = 0.5$ set empirically. Structurally complex, high-rank regions receive proportionally larger bit allocations. The entropy suppression term $\eta\,\mathcal{E}(\mathcal{S})$ simultaneously attenuates this increment for self-similar regions, whose repeating patterns imply low intrinsic rank despite apparent visual complexity, realizing rank-proportional bit assignment within the SR quantization pipeline.

Loss function for optimization. We freeze the pre-trained full-precision (FP) network weights $\mathcal{P}$ and optimize quantized network $\mathcal{Q}$ parameters using L1-loss and structured knowledge transfer (SKT) loss (Tu et al., 2023b; Li et al., 2020) for teacher-student learning, and the bit-allocation loss (Hong & Lee, 2024) for regularization. The reconstruction losses include:

$$\mathcal{L}_{\text{L1}} = \tfrac{1}{N}\textstyle\sum_i^N \|\mathcal{P}(I^i_{LR}) - \mathcal{Q}(I^i_{LR})\|_1, \tag{14}$$

$$\mathcal{L}_{\text{SKT}} = \frac{1}{N\cdot K}\sum_i^N\sum_k^K \left\| \frac{F^{i,k}_{\mathcal{P}}}{\|F^{i,k}_{\mathcal{P}}\|_2} - \frac{F^{i,k}_{\mathcal{Q}}}{\|F^{i,k}_{\mathcal{Q}}\|_2} \right\|^2, \tag{15}$$

$$\mathcal{L}_{\text{Bit}} = \max\Big(\tfrac{1}{N\cdot K}\textstyle\sum_i^N\sum_k^K b^{i,k} - b_{\text{tar}},\ 0\Big), \tag{16}$$

where $F^{i,k}_{\mathcal{P}}$ and $F^{i,k}_{\mathcal{Q}}$ are the k-th layer outputs of $\mathcal{P}$ and $\mathcal{Q}$, N is the batch size, and b_{tar} is the target bit-width. The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{SKT}}\mathcal{L}_{\text{SKT}} + \lambda_{\text{Bit}}\mathcal{L}_{\text{Bit}}$, with $\lambda_{\text{SKT}} = 10$ and $\lambda_{\text{Bit}} = 50$ set empirically. For clipping ranges, we follow AdaBM (Hong & Lee, 2024) and use $\mathcal{L}^{\text{clip}}_{\text{total}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{SKT}}\mathcal{L}_{\text{SKT}}$.

4. Experiments

Implementation details and Experiment Settings. We calibrate on 100 LR images randomly sampled from the DIV2K training set (Agustsson & Timofte, 2017) for one epoch (batch size 16). Activation ranges are initialized with MinMax (Jacob et al., 2018) and weights with OMSE (Choukroun et al., 2019). After freezing SR weights, we fine-tune clipping ranges and mapping parameters for 15 epochs (batch size 2) via Adam (Kingma & Ba, 2015) with initial learning rates of 0.01 for clipping and layer-wise factors and 0.1 for `I2B` parameters, decayed by 0.9 per epoch. Following common practice (Hong et al., 2022a; Tian et al., 2023; Zhong et al., 2022; Li et al., 2020), we apply adaptive quantization to activations and static quantization to weights in the SR body. Fractal analysis uses patch sizes $\{4, 6, 8, 12, 16\}$ and box-counting scales $\{2, 3, 4, 6, 8, 12, 16\}$. All experiments run on a single NVIDIA 2080Ti GPU. We report PSNR, SSIM, and Feature Average Bit-width (FAB, mean bit-width across all images).

Baselines. We compare against two categories: **(1) Adaptive QAT methods** (CADyQ (Hong et al., 2022a) and CABM (Tian et al., 2023)), which require ground-truth HR images and 40–70 hours of retraining; and **(2) PTQ methods** (AdaBM (Hong & Lee, 2024), Granular-DQ (Wang et al., 2025), PTQ4SR (Tu et al., 2023b), MinMax (Jacob et al., 2018), and Percentile). We evaluate EDSR (Lim et al., 2017), SRResNet (Ledig et al., 2017), RDN (Zhang et al., 2018), and SwinIR-light (Liang et al., 2021) at $4\times$ upscaling on Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2010), BSD100 (Martin et al., 2001), Urban100 (Huang et al., 2015), Test2K, and Test4K (Kong et al., 2021). Large inputs are split into 96×96 patches following (Hong et al., 2022a; Tian et al., 2023; Hong & Lee, 2024).

Comparison with Adaptive Quantization Methods. As shown in Table 2, FracQuant achieves **stable W2A2 quantization**, a feat unattainable by all prior methods, while matching or exceeding QAT-level quality without ground-truth

Model	QAT	GT	Process Time	W / A	Set5		Urban100		Test2K		Test4K	
					FAB	PSNR / SSIM	FAB	PSNR / SSIM	FAB	PSNR / SSIM	FAB	PSNR / SSIM
Bicubic (× 4)	-	-	-	- / -	-	26.89 / 0.789	-	21.24 / 0.639	-	24.97 / 0.715	-	26.1 / 0.764
EDSR	-	-	-	32 / 32	32.0	32.10 / 0.890	32.0	26.04 / 0.784	32.0	27.71 / 0.782	32.0	28.80 / 0.814
EDSR-CADyQ	✓	✓	40 hrs	8 / 6MP	6.5	32.04 / 0.890	6.6	25.98 / 0.784	6.1	27.69 / 0.781	6.0	28.79 / 0.814
EDSR-CABM	✓	✓	70 hrs	8 / 6MP	5.8	31.97 / 0.884	5.8	25.90 / 0.782	5.6	27.65 / 0.781	5.5	28.73 / 0.812
EDSR-AdaBM	✗	✗	**71 sec**	8 / 6MP	5.7	32.05 / 0.888	5.7	25.96 / 0.782	5.3	27.65 / 0.779	5.2	28.71 / 0.809
EDSR-FracQuant (**Ours**)	✗	✗	75 sec	8 / 6MP	5.8	**32.06 / 0.889**	**5.7**	**26.12 / 0.801**	**5.3**	**27.70 / 0.780**	**5.2**	**28.82 / 0.817**
EDSR-CADyQ	✓	✓	40 hrs	4 / 4MP	4.3	31.24 / 0.864	4.9	25.12 / 0.753	4.9	27.43 / 0.771	4.9	28.49 / 0.803
EDSR-CABM	✓	✓	70 hrs	4 / 4MP	5.8	31.17 / 0.852	4.4	24.98 / 0.746	4.4	27.33 / 0.767	4.4	28.36 / 0.798
EDSR-AdaBM	✗	✗	**71 sec**	4 / 4MP	3.8	31.32 / 0.864	4.3	25.49 / 0.759	3.9	27.40 / 0.758	3.8	28.39 / 0.784
EDSR-Granular-DQ	✗	✗	296 sec	4 / 4MP	3.8	31.35 / 0.867	4.9	25.48 / 0.784	4.3	27.48 / 0.774	4.4	28.58 / 0.800
EDSR-FracQuant (**Ours**)	✗	✗	75 sec	4 / 4MP	**3.9**	**31.47 / 0.875**	**3.9**	**25.58 / 0.764**	**3.8**	**27.56 / 0.784**	**3.7**	**28.66 / 0.814**
EDSR-CADyQ	✓	✓	40 hrs	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
EDSR-CABM	✓	✓	70 hrs	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
EDSR-AdaBM	✗	✗	**71 sec**	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
EDSR-Granular-DQ	✗	✗	296 sec	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
EDSR-FracQuant (**Ours**)	✗	✗	75 sec	2 / 2MP	**1.9**	**28.78 / 0.812**	**1.9**	**24.43 / 0.734**	**1.9**	**26.53 / 0.748**	**1.9**	**27.85 / 0.778**
SRResNet	-	-	-	32 / 32	32.0	25.90 / 0.782	32.0	25.86 / 0.779	32.0	27.64 / 0.781	32.0	28.72 / 0.813
SRResNet-CADyQ	✓	✓	53 hrs	8 / 6MP	6.2	31.89 / 0.887	6.4	25.89 / 0.779	6.2	27.65 / 0.780	6.2	28.73 / 0.812
SRResNet-CABM	✓	✓	93 hrs	8 / 6MP	5.8	31.87 / 0.885	5.8	25.79 / 0.776	5.6	27.60 / 0.779	5.6	28.67 / 0.811
SRResNet-AdaBM	✗	✗	**92 sec**	8 / 6MP	5.8	31.89 / 0.887	5.6	25.72 / 0.773	5.2	27.55 / 0.777	5.1	28.62 / 0.809
SRResNet-FracQuant (**Ours**)	✗	✗	98 sec	8 / 6MP	**5.8**	**31.92 / 0.889**	**5.6**	**26.01 / 0.782**	**5.2**	**27.68 / 0.781**	**5.1**	**28.74 / 0.813**
SRResNet-CADyQ	✓	✓	53 hrs	4 / 4MP	4.2	31.08 / 0.842	4.1	25.39 / 0.761	4.1	27.38 / 0.771	4.1	28.46 / 0.804
SRResNet-CABM	✓	✓	93 hrs	4 / 4MP	3.7	31.17 / 0.858	3.8	25.42 / 0.764	3.8	27.46 / 0.774	3.6	28.52 / 0.806
SRResNet-AdaBM	✗	✗	92 sec	4 / 4MP	3.8	31.17 / 0.864	4.2	25.32 / 0.757	3.9	27.31 / 0.766	3.9	28.25 / 0.782
SRResNet-Granular-DQ	✗	✗	314 sec	4 / 4MP	3.9	31.21 / 0.866	4.00	25.98 / 0.783	4.0	27.55 / 0.773	4.0	28.93 / 0.820
SRResNet-FracQuant (**Ours**)	✗	✗	**98 sec**	4 / 4MP	**3.9**	**31.23 / 0.879**	**3.9**	**25.70 / 0.772**	**3.8**	**27.61 / 0.774**	**3.7**	**28.69 / 0.807**
SRResNet-CADyQ	✓	✓	53 hrs	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
SRResNet-CABM	✓	✓	93 hrs	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
SRResNet-AdaBM	✗	✗	**92 sec**	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
SRResNet-Granular-DQ	✗	✗	314 sec	2 / 2MP	✗	✗	✗	✗	✗	✗	✗	✗
SRResNet-FracQuant (**Ours**)	✗	✗	**98 sec**	2 / 2MP	**1.9**	**27.12 / 0.810**	**1.9**	**24.14 / 0.728**	**1.9**	**26.36 / 0.741**	**1.9**	**27.12 / 0.771**
SwinIR-light	-	-	-	32 / 32	32.0	26.07 / 0.791	32.0	26.46 / 0.798	32.0	27.72 / 0.779	32.0	29.14 / 0.825
SwinIR-CADyQ	✓	✓	–	8 / 6MP	5.89	25.98 / 0.789	5.15	25.87 / 0.779	5.01	27.54 / 0.772	5.01	28.92 / 0.819
SwinIR-CABM	✓	✓	–	8 / 6MP	5.76	25.91 / 0.787	5.34	25.88 / 0.780	4.92	27.62 / 0.774	4.91	29.02 / 0.821
SwinIR-Granular-DQ	✗	✗	–	8 / 6MP	5.70	25.94 / 0.789	4.79	26.42 / 0.796	4.74	27.67 / 0.778	4.76	29.11 / 0.824
SwinIR-FracQuant (**Ours**)	✗	✗	–	8 / 6MP	**5.61**	**26.01 / 0.790**	**4.71**	**26.58 / 0.801**	**4.68**	**27.89 / 0.792**	**4.63**	**29.38 / 0.851**
SwinIR-CADyQ	✓	✓	–	4 / 4MP	5.21	25.43 / 0.778	5.15	25.48 / 0.773	5.01	27.38 / 0.768	5.01	28.76 / 0.814
SwinIR-CABM	✓	✓	–	4 / 4MP	5.18	25.39 / 0.776	5.34	25.52 / 0.774	4.92	27.44 / 0.769	4.91	28.87 / 0.816
SwinIR-Granular-DQ	✗	✗	–	4 / 4MP	5.11	25.54 / 0.783	5.26	25.48 / 0.784	5.01	27.84 / 0.786	4.94	29.32 / 0.826
SwinIR-FracQuant (**Ours**)	✗	✗	–	4 / 4MP	**5.02**	**25.68 / 0.797**	**5.02**	**25.71 / 0.798**	**4.92**	**28.06 / 0.801**	**4.84**	**29.41 / 0.854**

Table 2. **Comparisons with content-aware adaptive bit-mapping methods on SR**, evaluated with EDSR and SRResNet of scale 4. CADyQ (Hong et al., 2022a) and CABM (Tian et al., 2023) require extensive QAT process with ground-truth (GT) HR images, which result in long process time. All previous methods, including AdaBM (Hong & Lee, 2024) and Granular-DQ (Wang et al., 2025) fail to ultra-low bit quantization due to unstable gradient-base complexity measurement.

HR images or hours of retraining. At W4A4, FracQuant consistently outperforms AdaBM across all benchmarks and backbones within 75–98 seconds of calibration, with the most pronounced gains on Test2K and Test4K, indicating that fractal-based complexity captures high-frequency structural information more faithfully than gradient-based proxies. Although Granular-DQ matches or exceeds FracQuant on some W4A4 settings, its patch-wise sequential inference adds 280–314 seconds of overhead and overflows at W2A2; FracQuant's W4A4 edge is therefore mainly its single-pass calibration speed. The combination of W2A2 capability and sub-100-second calibration makes FracQuant attractive for resource-constrained edge deployment, where memory footprint is a key constraint.

Comparison with Static PTQ Methods. Table 3 compares FracQuant against non-adaptive PTQ baselines (MinMax, Percentile, PTQ4SR (Tu et al., 2023b)) on EDSR and RDN at W4A4. FracQuant surpasses PTQ4SR by **+0.96 dB PSNR** on Set5 with EDSR while achieving a lower FAB (3.9 vs. 4.0), demonstrating that rank-adaptive bit allocation provides genuine efficiency gains beyond fixed-bit quantization. The advantage is consistent across all benchmarks and both backbones: on Urban100, FracQuant outperforms PTQ4SR by **+0.66 dB** with EDSR and **+0.12 dB** with RDN, indicating that fractal-based complexity estimation better captures the structural diversity of urban scenes. Fine-tuning further amplifies these gains, as FracQuant with fine-tuning consistently achieves the best PSNR/SSIM/FAB trade-off across all settings.

W2A2 stability. Fig. 8 reveals the root cause of prior method failures at W2A2. Gradient-based complexity scores exhibit high variance across image regions: in smooth patches, near-zero spatial gradients produce ill-conditioned estimates, causing the I2B mapping to assign inconsistent

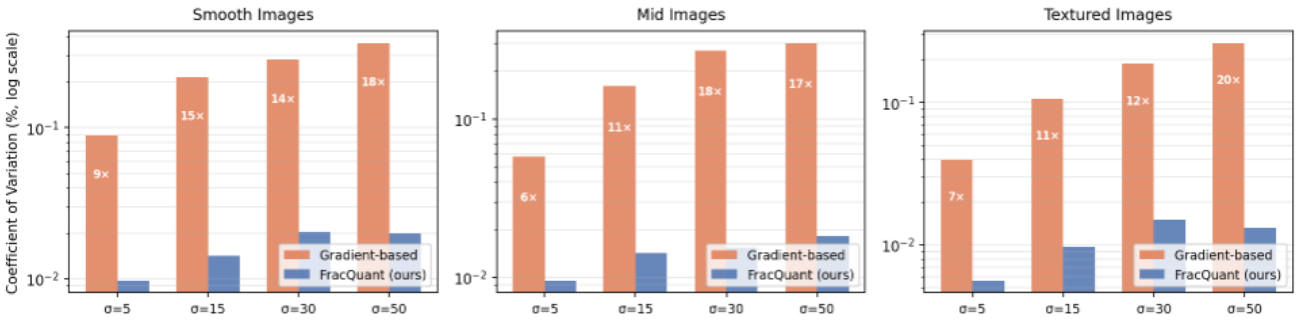

Figure 8. **Stability comparison between gradient-based methods and FracQuant.** Gradient-based scores exhibit high variance in smooth and textured regions, causing overflow at W2A2. FracQuant's fractal dimension provides a geometry-driven proxy with 30× lower raw variation, enabling the first stable W2A2 quantization for SR networks.

Model	FT	W / A	Set5		Set14		BSD100		Urban100	
			FAB	PSNR / SSIM	FAB	PSNR / SSIM	FAB	PSNR / SSIM	FAB	PSNR / SSIM
Bicubic	-	- / -	-	26.89 / 0.810	-	23.96 / 0.688	-	24.32 / 0.656	-	21.24 / 0.639
EDSR-MinMax	✗	4 / 4	4.0	26.83 / 0.624	4.0	25.04 / 0.546	4.0	24.57 / 0.503	4.0	23.12 / 0.536
EDSR-Percentile	✗	4 / 4	4.0	24.03 / 0.776	4.0	23.95 / 0.712	4.0	24.42 / 0.687	4.0	21.62 / 0.677
EDSR-MinMax+FT	✓	4 / 4	4.0	28.97 / 0.821	4.0	26.47 / 0.721	4.0	26.24 / 0.687	4.0	23.46 / 0.674
EDSR-Percentile+FT	✓	4 / 4	4.0	27.01 / 0.819	4.0	25.71 / 0.736	4.0	25.69 / 0.707	4.0	23.18 / 0.707
EDSR-PTQ4SR	✓	4 / 4	4.0	30.51 / 0.836	4.0	27.62 / 0.735	4.0	26.88 / 0.693	4.0	24.92 / 0.721
EDSR-FracQuant (**Ours**)	✓	4 / 4	**3.9**	**31.47 / 0.875**	**3.7**	**28.22 / 0.773**	**3.5**	**27.46 / 0.744**	**3.9**	**25.58 / 0.764**
RDN-MinMax	✗	4 / 4	4.0	25.91 / 0.632	4.0	24.22 / 0.549	4.0	24.29 / 0.530	4.0	22.24 / 0.523
RDN-Percentile	✗	4 / 4	4.0	18.83 / 0.771	4.0	18.28 / 0.662	4.0	19.83 / 0.646	4.0	16.77 / 0.625
RDN-MinMax+FT	✓	4 / 4	4.0	28.50 / 0.810	4.0	26.15 / 0.703	4.0	26.00 / 0.673	4.0	23.35 / 0.673
RDN-Percentile+FT	✓	4 / 4	4.0	21.24 / 0.798	4.0	20.68 / 0.690	4.0	21.85 / 0.668	4.0	18.64 / 0.652
RDN-PTQ4SR	✓	4 / 4	4.0	28.32 / 0.813	4.0	26.11 / 0.709	4.0	25.82 / 0.671	4.0	23.31 / 0.668
RDN-FracQuant (**Ours**)	✓	4 / 4	**3.9**	**28.74 / 0.816**	**3.8**	**26.35 / 0.715**	**3.6**	**26.24 / 0.684**	**3.8**	**23.43 / 0.671**

Table 3. **Comparisons with static PTQ methods on SR.** FracQuant outperforms PTQ4SR by up to +0.96 dB on Set5 with EDSR while achieving a lower FAB, demonstrating that rank-adaptive bit allocation provides genuine efficiency gains over fixed-precision quantization.

bit-widths and ultimately triggering denormalization overflow in 2-bit activations. Across both smooth and heavily textured regions, this instability manifests as erratic per-image bit assignments that destabilize the entire calibration procedure. FracQuant's raw fractal-dimension scores show a 30× **lower coefficient of variation** across image types, owing to their scale-invariant, geometry-driven nature that remains well-conditioned even in low-texture regions where gradients vanish. Beyond the raw signal, these more consistent complexity estimates help keep the I2B mapping within a reliable range, mitigating the bit-width oscillations that cause overflow at 2-bit precision. The DHL mechanism further stabilizes the soft-to-hard transition during calibration, replacing the abrupt discrete quantization step with a smooth differentiable proxy. Together, these two components enable FracQuant to achieve **28.78 dB on Set5** and **24.43 dB on Urban100** with EDSR at W2A2 — the only stable results in this setting, though still well below full-precision quality.

Visual Comparison Fig. 7 shows qualitative results on Urban100 at W4A4 with EDSR. FracQuant produces sharper edges and more natural textures in fine-grained structures (window grids, brick patterns, text) compared to AdaBM,

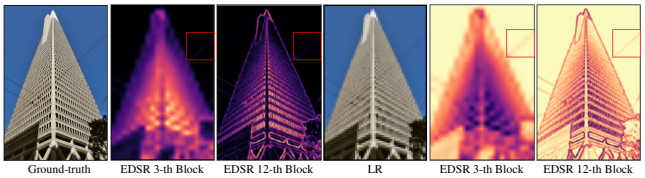

Figure 9. **Feature map comparison: AdaBM vs. FracQuant.** AdaBM's gradient-based allocation assigns insufficient bits to complex patches, yielding increasingly smoothed intermediate features at deeper layers; FracQuant's fractal-aware allocation preserves richer activation patterns across layers by leveraging fractal dimension as a proxy for local intrinsic rank, ultimately translating into sharper SR reconstruction.

consistent with the quantitative gains on urban-scene benchmarks. The feature map visualization in Fig. 9 illustrates that FracQuant preserves richer activation patterns across network layers: AdaBM's gradient-based allocation assigns insufficient bits to complex patches, yielding smoothed intermediate features, whereas FracQuant's fractal-aware allocation maintains detailed representations that translate into better SR reconstruction.

Complexity	Similarity	S	ϵ_s	PSNR	SSIM
$[2, 4, 6, 8]$	$[4, 6, 8, 12, 16]$	4	1e-6	27.839	0.7148
$[2, 3, 4, 6, 8]$	$[4, 6, 8, 12, 16]$	4	1e-6	28.787	0.8128
$[2, 3, 4, 6, 8, 10]$	$[4, 6, 8, 12, 16]$	4	1e-6	28.787	0.8128
$[2, 3, 5, 7, 9, 11]$	$[4, 6, 8, 12, 16]$	4	1e-6	25.844	0.5675
$[2, 3, 4, 6, 8, 10, 12, 14, 16]$	$[4, 6, 8, 12, 16]$	4	1e-6	28.787	0.8128
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	3	1e-6	28.689	0.7826
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	5	1e-6	27.493	0.6859
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	6	1e-6	28.275	0.7468
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	4	5e-6	27.864	0.7101
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	4	1e-5	28.309	0.7482
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	4	5e-5	27.816	0.7069
$[2, 3, 4, 6, 8, 12, 16]$	$[4, 6, 8, 12, 16]$	4	1e-6	**28.787**	**0.8128**

Table 4. **Sensitivity to fractal-assessment hyper-parameters**, evaluated on Set5 with W2A2 EDSR ($\times$4). Default setting highlighted.

M	F	S	DHL	FAB	PSNR/SSIM	FAB	PSNR/SSIM
✗	✗	✗	✗	3.8	31.02 / 0.860	3.7	25.11 / 0.736
✓	✓	✓	✗	3.9	31.44 / 0.872	3.9	25.54 / 0.759
✗	✗	✗	✓	3.8	31.14 / 0.862	3.8	25.14 / 0.738
✓	✗	✗	✓	3.8	31.17 / 0.862	3.7	23.80 / 0.662
✗	✓	✗	✓	3.7	31.35 / 0.869	3.9	25.48 / 0.754
✗	✗	✓	✓	4.0	30.86 / 0.792	4.0	24.89 / 0.714
✓	✓	✓	✓	**3.9**	**31.47 / 0.875**	**3.9**	**25.58 / 0.764**

Table 5. **Ablation study on each attribute of FracQuant** evaluated on Set5/Urban100 with W4A4 EDSR ($\times$4). M, F, S respectively denote the multi-scale feature extraction, fractal complexity assessment, and geometric self-similarity assessment module.

4.1. Ablation Study

Hyper-parameter sensitivity. Table 4 probes the fractal-assessment hyper-parameters on Set5 at W2A2, the most fragile setting. Reconstruction is robust to the *number* of box-counting scales: once the fine scales $\{2, 3, 4\}$ are included, adding coarser scales leaves PSNR essentially unchanged (28.79 dB), whereas omitting fine scales ($\{2, 4, 6, 8\}$) or using non-dyadic scales ($\{2, 3, 5, 7, 9, 11\}$) degrades it to 27.84 and 25.84 dB, consistent with our rotation-invariant dyadic design. The Top-K value and the counting threshold ϵ_s have clear optima at K=4 and ϵ_s=10^{-6}, with deviations costing up to $\sim$1.3 dB. These trends hold even under W2A2, indicating that our default configuration is a stable operating point rather than a finely tuned one.

Component-wise Importance. Table 5 ablates each component on Set5 and Urban100 using W4A4 EDSR ($\times$4). DHL alone reduces instability at complexity boundaries, already improving over the baseline. Multi-scale feature extraction (M) is essential for structurally complex scenes: without it, fractal assessment degrades significantly on Urban100, indicating that cross-scale context is a prerequisite for reliable rank estimation. Fractal complexity assessment (F) improves FAB efficiency, enabling more aggressive yet stable bit allocation, while self-similarity assessment (S) preserves fine-grained detail, with its absence causing a pronounced SSIM drop. Together, all four components are mutually reinforcing: the full model achieves the best PSNR/SSIM/FAB trade-off across both benchmarks, with a cumulative Urban100 gain of $+0.47$ dB over the baseline.

Gradient weight α. We ablate the blending coefficient α that mixes gradient and fractal complexity in Eq. (8). Setting α=0.10 (vs. α=0.05) has negligible effect at W4A4 ($\pm$0.01 dB) but yields consistent improvements at W2A2 ($+$0.05 dB on Set5, $+$0.02 dB on Urban100). At extreme compression, gradient complexity becomes more discriminative as fractal features begin to saturate, making a slightly higher gradient contribution beneficial.

Bit allocation mapping strategy. We compare the legacy offset scheme $\{-1, -0.5, 0\}$ against a boost variant $\{0, +0.5, +1\}$ that adds bits to complex images instead of subtracting from simple ones. The boost variant yields higher peak PSNR ($+$0.22 dB Set5 at W4A4), but its FAB rises from 3.88 to 4.73. When evaluating at matched FAB ($\approx$4.07), the two schemes perform comparably (within $\pm$0.04 dB), confirming that the boost variant's apparent gains stem from consuming additional bits rather than more efficient allocation. The legacy scheme therefore offers a better compression–quality tradeoff under a fixed rate budget, validating our design choice in Eq. (11).

5. Conclusion

We presented **FracQuant**, a rank-adaptive SR quantization framework that grounds complexity estimation in fractal geometry. Fractal dimension serves as a geometry-driven, activation-insensitive proxy for local patch rank, resolving the instability of gradient-based estimation and enabling principled bit allocation for edge devices. Combined with geometric self-similarity assessment and Dynamic Hybrid soft-hard Labeling, FracQuant matches or surpasses QAT-based methods under standard W4A4/W8A6 settings without ground-truth supervision, and is the first method to achieve stable W2A2 quantization across multiple SR backbones. Beyond SR, we believe fractal dimension may offer a useful complexity signal for other dense prediction tasks where gradient-based measures are unreliable, opening a new direction for geometry-driven adaptive compression in low-level vision.

Acknowledgment

This study was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants 112-2221-E-006-157-MY3, 114-2627-M-A49-003, and 114-2218-E-035-001. We thank the National Center for NCHC of NARLabs in Taiwan for providing computational and storage resources.

References

Agustsson, E. and Timofte, R. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

Argilaga, A. Fractal informed generative adversarial networks (fi-gan): Application to the generation of x-ray ct images of a self-similar partially saturated sand. *Computers and Geotechnics*, 158:105384, 2023. ISSN 0266-352X. doi: https://doi.org/10.1016/j.compgeo.2023.105384.

Barnsley, M. F. and Hurd, L. P. *Fractal image compression*. A. K. Peters, Ltd., USA, 1993. ISBN 1568810008.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

Bouda, M., Caplan, J. S., and Saiers, J. E. Box-counting dimension revisited: Presenting an efficient method of minimizing quantization error and an assessment of the self-similarity of structural root systems. *Frontiers in Plant Science*, 7:149, February 2016. doi: 10.3389/fpls.2016.00149.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Choukroun, Y., Kravchik, E., Yang, F., and Kisilev, P. Low-bit quantization of neural networks for efficient inference. In *ICCV Workshops*, 2019.

Chu, X., Zhang, B., Ma, H., Xu, R., and Li, Q. Fast, accurate and lightweight super-resolution with neural architecture search. In *ICPR*, 2021.

Denton, E., Zaremba, W., Bruna, J., LeCun, Y., and Fergus, R. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, 2014.

Dong, C., Loy, C. C., He, K., and Tang, X. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014.

Fisher, Y. (ed.). *Fractal image compression: theory and application*. Springer-Verlag, Berlin, Heidelberg, 1995. ISBN 0387942114.

Gao, Q., Zhao, Y., Li, G., and Tong, T. Image super-resolution using knowledge distillation. In *ACCV*, 2018.

He, S., Wu, C., Liu, C., and Wang, Z. Single image super resolution enhancement method based on fractal interpolation of iteration function system. In *Journal of Physics: Conference Series*, volume 1883, pp. 012059. IOP Publishing, 2021. doi: 10.1088/1742-6596/1883/1/012059. 2nd International Conference on Computer Information and Big Data Applications, 26–28 March 2021, Wuhan, China.

Hong, C. and Lee, K. M. Overcoming distribution mismatch in quantizing image super-resolution networks. *arXiv preprint arXiv:2307.13337*, 2023.

Hong, C. and Lee, K. M. Adabm: On-the-fly adaptive bit mapping for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2641–2650, 2024. doi: 10.1109/CVPR52733.2024.00255.

Hong, C., Baik, S., Kim, H., Nah, S., and Lee, K. M. Cadyq: Content-aware dynamic quantization for image super-resolution. In *ECCV*, 2022a.

Hong, C., Kim, H., Baik, S., Oh, J., and Lee, K. M. Daq: Channel-wise distribution-aware quantization for deep image super-resolution networks. In *WACV*, 2022b.

Hou, Z. and Kung, S.-Y. Multi-dimensional dynamic model compression for efficient image super-resolution. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3492–3502, 2022. doi: 10.1109/WACV51458.2022.00355.

Hsu, C.-C., Kang, C.-M., and Lee, W.-C. DRCT: Saving image super-resolution away from information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Huang, J.-B., Singh, A., and Ahuja, N. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jiang, X., Wang, N., Xin, J., Xia, X., Yang, X., and Gao, X. Learning lightweight super-resolution networks with weight pruning. *Neural Networks*, 144:21–32, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Kong, X., Zhao, H., Qiao, Y., and Dong, C. Classsr: A general framework to accelerate super-resolution networks by data characteristic. In *CVPR*, 2021.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Li, H., Yan, C., Lin, S., Zheng, X., Zhang, B., Yang, F., and Ji, R. Pams: Quantized super-resolution via parameterized max scale. In *ECCV*, 2020.

Li, T., Sun, Q., Fan, L., and He, K. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1833–1844, 2021.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshops*, 2017.

Liu, J., Wang, Q., Zhang, D., and Shen, L. Super-resolution model quantized in multi-precision. *Electronics*, 10(17): 2176, 2021.

Mandelbrot, B. How long is the coast of britain? statistical self-similarity and fractional dimension. *Science*, 156 (3775):636–638, May 1967. doi: 10.1126/science.156.3775.636.

Mandelbrot, B. Self-affine fractals and fractal dimension. *Physica Scripta*, 32(4):257–260, 1985. doi: 10.1088/0031-8949/32/4/001.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, pp. 416–423, 2001.

Qin, H., Zhang, Y., Ding, Y., liu, Y., Liu, X., Danelljan, M., and Yu, F. Quantsr: Accurate low-bit quantization for efficient image super-resolution. In *NeurIPS*, 2023.

Sun, Y.-C., Yeo, C. Y., Chu, E., Chen, J.-C., and Liu, Y.-L. Fiper: Generalizable factorized fields for joint image compression and super-resolution. *arXiv*, 2024.

Tian, S., Lu, M., Liu, J., Guo, Y., Chen, Y., and Zhang, S. Cabm: Content-aware bit mapping for single image super-resolution network with large input. In *CVPR*, 2023.

Tu, C.-H., Chen, H.-Y., Carlyn, D., and Chao, W.-L. Learning fractals by gradient descent. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023a. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i2.25342. URL https://doi.org/10.1609/aaai.v37i2.25342.

Tu, Z., Hu, J., Chen, H., and Wang, Y. Toward accurate post-training quantization for image super resolution. In *CVPR*, 2023b.

Wang, M., Zhang, Z., Li, F., Xu, K., Miao, K., and Wang, M. Thinking in granularity: Dynamic quantization for image super-resolution by intriguing multi-granularity clues. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2025.

Wang, S., Liu, J., Chen, K., Li, X., Lu, M., and Guo, Y. Adaptive patch exiting for scalable single image super-resolution. In *ECCV*, 2022.

Wu, J., Jin, X., Mi, S., and Tang, J. An effective method to compute the box-counting dimension based on the mathematical definition and intervals. *Results in Engineering*, 6: 100106, June 2020. doi: 10.1016/j.rineng.2020.100106.

Xu, H., Zhai, G., and Yang, X. Single image super-resolution with detail enhancement based on local fractal analysis of gradient. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(10):1740–1754, 2013. doi: 10.1109/TCSVT.2013.2248305.

Yu, K., Wang, X., Dong, C., Tang, X., and Loy, C. C. Path-restore: Learning network path selection for image restoration. *IEEE TPAMI*, 2021.

Zeyde, R., Elad, M., and Protter, M. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pp. 711–730. Springer, 2010.

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., and Fu, Y. Residual dense network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhong, Y., Lin, M., Li, X., Li, K., Shen, Y., Chao, F., Wu, Y., and Ji, R. Dynamic dual trainable bounds for ultra-low precision super-resolution networks. In *ECCV*, 2022.