Supplementary Material
Thinking in Granularity: Dynamic Quantization for Image Super-Resolution by Intriguing Multi-Granularity Clues

Mingshen Wang[1], Zhao Zhang[1,3*] Feng Li[1*], Ke Xu[2], Kang Miao[1], Meng Wang[1]

[1] Hefei University of Technology [2] Anhui University
[3]Yunnan Key Laboratory of Software Engineering, Yunan, China

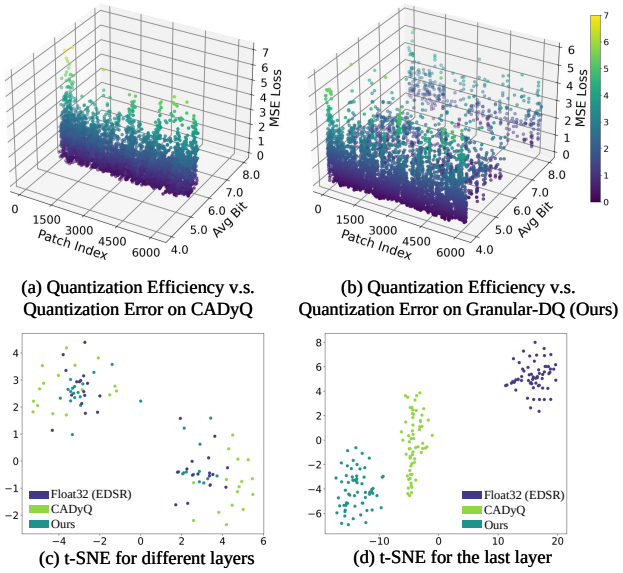
Appendix

The supplementary material mainly includes the following contents:

- The motivation of Granular-DQ.
- Ablation studies involve the thresholds (quantile t) number in E2B, and the combination with different weight quantization patterns, *i.e.* PAMS (Li et al. 2020) and QuantSR (Qin et al. 2024).
- More implementation details on the transformer-based baseline models including SwinIR-light (Liang et al. 2021) and HAT-S (Chen et al. 2023).
- More experimental results consist of $\times 2$ SR performance and additional qualitative visualization.
- Limitations in Granular-DQ.

Motivation

Recent advances (Tu et al. 2023; Hong et al. 2022; Tian et al. 2023; Lee, Yoo, and Jung 2024; Hong and Lee 2024) have demonstrated the benefits of considering the quantization sensitivity of layers and image contents in SR quantization. Taking CADyQ (Hong et al. 2022) for example, it applies a trainable bit selector to determine the proper bit-width and quantization level for each layer and a given local image patch based on the feature gradient magnitude. In our analysis, we compute the average bit-width and the quantization error measured by MSE between the reconstructions of the quantized model (via CADyQ) and the original high-precision model (EDSR) on the Test2K dataset. Figure S1 (a) reveals that the majority of patches fall within a 6-bit to 8-bit range, accompanied by a relatively elevated MSE. Furthermore, we present t-SNE maps for various quantized layers and the final layer in Figure S1 (c)-(d). Firstly, it is evident that the distribution of different layers quantized by CADyQ is markedly more scattered than that of the original model, as depicted in Figure S1 (c). Secondly, on the final layer, the features from the CADyQ-quantized model exhibit a distinct vertical pattern, which is notably at odds with the structure of the original model's feature points (Figure S1 (d)). In our investigation, CABM actually exhibited



(a) Quantization Efficiency v.s. Quantization Error on CADyQ

(b) Quantization Efficiency v.s. Quantization Error on Granular-DQ (Ours)

(c) t-SNE for different layers

(d) t-SNE for the last layer

Figure S1: Analysis of the quantization efficiency, quantization error, and feature distribution in t-SNE on CADyQ and our Granular-DQ. (a) and (b) illustrate the quantization efficiency v.s. quantization error trade-off; (c) and (d) visualize the feature distribution of two resultant models and compare with the corresponding original one (Float32: EDSR).

similar findings, although it fine-tunes the CADyQ model based on edge scores. These results indicate that: 1) Simply relying on image edge information is suboptimal for the trade-off between quantization efficiency and error; 2) The bit allocation for each layer in response to varying patches can introduce disturbances to the inter-layer relations within original models leading to disparities in the representations.

Based on the above analysis, this work aims to design a dynamic quantization approach for diverse image contents while maintaining the representation ability of the original model. To this end, we rethink the image characteristics related to image quality from the granularity and information density. As we know, the fine-granularity representations reveal the texture complexity of local regions, while coarse ones express structural semantics of the overall scene. Besides, according to Shannon's Second Theorem (Shan-

*Corresponding authors.

t^*	b^*	Set14			Urban100		
		FAB	PSNR	SSIM	FAB	PSNR	SSIM
$[0.5]$	$[4, 8]$	6.57	28.57	0.780	5.75	25.97	0.781
$[0.5, 0.9]$	$[4, 5, 8]$	**5.54**	**28.58**	**0.781**	**4.97**	**26.01**	**0.784**
$[0.4, 0.6, 0.9]$	$[4, 5, 6, 8]$	6.07	**28.58**	0.779	5.41	25.93	0.781
$[0.4, 0.6, 0.9]$	$[4, 5, 7, 8]$	6.21	28.54	0.780	5.61	25.93	0.782

Table S1: Ablation study on the influence of a different number of thresholds (quantile, denoted by t^*) and corresponding bit configuration (denoted by b^*) in E2B with EDSR.

non 1948), the entropy statistic reflects the average information density and the complexity of pixel distributions given patches, which is directly correlated to the image quality. Therefore, we propose Granular-DQ, a markedly different method that fully explores the granularity and entropy statistic of images to quantization adaption. Granular-DQ contains two sequential steps: 1) granularity-aware bit allocation for all the patches and 2) entropy-based fine-grained bit-width adaption for the patches less quantized by 1). In this way, we can see that the bit-width allocation by Graular-DQ is sparser than CADyQ, where a majority of patches are lower than 5bit with only a few patches at high bit-width (Figure S1 (b)). Moreover, the feature distribution of the layers quantized by our method is closer to that of the original model (Figure S1(c)-(d)). These validate that our Granular-DQ enables low-bit and layer-invariant quantization.

Ablation Study

Impact of the Threshold Number in E2B. We further experimentally investigated the effect of different numbers of thresholds in E2B and their corresponding candidate bit configuration. Firstly, we assume that there is only one quantile for all the input patches, which means the entropy statistic $\mathbf{H}$ is divided into two subintervals. As shown in Table S1, when we adjust the bit-widths of patches using 4/8bit, the model performs worst on both Set14 and Urban100 datasets. Similarly, when we incorporate three thresholds of t with $[0.4, 0.6, 0.9]$ to divide $\mathbf{H}$ into four subintervals, it can be seen that whether using the bit configurations of $[4, 5, 6, 8]$ or $[4, 5, 7, 8]$, the model cannot obtain satisfied quantization efficiency. In contrast, the model with two thresholds $[0.5, 0.9]$ and corresponding candidate bit-widths of $[4, 5, 8]$ achieved the best trade-off on both datasets, making it our final choice.

Influence of Different Quantization Patterns. To investigate the compatibility of our method on different quantization patterns, we conduct experiments by combining Granular-DQ with PAMS (Li et al. 2020) and QuantSR (Qin et al. 2024), where the results on Urban100 are reported in Table S2. Notably, different from existing dynamic methods (Hong et al. 2022; Tian et al. 2023), our Granular-DQ does not require the pre-trained models of PAMS or QuantSR. We can see that Granular-DQ+PAMS gets 0.07dB PSNR gains with 0.4 FAB reduction for EDSR compared to CADyQ+PAMS. When applying the QuantSR scheme on Granular-DQ, the model can achieve the best trade-off between FAB and PSNR/SSIM for both EDSR and IDN models, where even the latter surpasses the original model by 0.26dB in PSNR.

Methods	Urban100		
	FAB↓	PSNR↑	SSIM↑
EDSR	32.00	26.03	0.784
PAMS	8.00	26.01	0.784
CADyQ+PAMS	6.09	25.94	0.782
Granular-DQ+PAMS	5.69	25.95	0.782
Granular-DQ+QuantSR	**4.97**	**26.01**	**0.784**
IDN	32.00	25.42	0.763
PAMS	8.00	25.56	0.768
CADyQ+PAMS	5.78	25.65	0.771
Granular-DQ+PAMS	4.73	25.62	0.770
Granular-DQ+QuantSR	**4.18**	**25.68**	**0.772**

Table S2: Investigation of the compatibility of our Granular-DQ with different quantization patterns. We observe the $\times 4$ SR results on Urban100 based on EDSR and IDN.

Implementation Details on Transformer-based Baselines

For the transformer-based models, the linear layers of the MLPs in both SwinIR-light (Liang et al. 2021) and HAT-S (Chen et al. 2023) are all quantized using the QuantSR scheme (Qin et al. 2024). Surprisingly, despite all our efforts, we still encountered the gradient explosion issue when implementing the quantization scheme for CADyQ (Hong et al. 2022) and CABM (Tian et al. 2023) in HAT-S. As a result, these two retained full precision for channel attention in the experiments. During the training phase, we randomly cropped the LR image into 64×64 with a total batch size of 16 for all scale factors, following the settings of the original model. The learning rate was initially set to 2×10^{-4} and halved after 250K iterations.

Comparison with the State-of-the-Art

Quantitative Comparison for $\times$2 SR. We further conduct experiments for $\times 2$ SR, where the quantitative results are illustrated in Table S3. Obviously, Granular-DQ demonstrates competitive trade-offs in terms of FAB and PSNR/SSIM compared to other quantization methods across all CNN models. Additionally, for SwinIR-light and HAT-S, Granular-DQ also achieves remarkably superior reconstruction accuracy to full precision than others while maintaining the lowest FAB.

More Qualitative Comparison. In Figure S2 and S3, we provide more $\times 4$ SR visual results produced by recent state-of-the-art methods and our Granular-DQ. Regardless of whether the models are CNN-based or transformer-based, Granular-DQ consistently achieves superior reconstruction details at the lowest FAB compared to other quantization methods in most instances. In each case, minimal discrepancies can be observed between Granular-DQ and its corresponding full-precision model. These findings further validate that Granular-DQ ensures an optimal trade-off between reconstruction accuracy and quantization efficiency.

Methods	Scale	Urban100			Test2K			Test4K		
		FAB↓	PSNR↑	SSIM↑	FAB↓	PSNR↑	SSIM↑	FAB↓	PSNR↑	SSIM↑
SRResNet	×2	32.00	32.11	0.928	32.00	32.81	0.930	32.00	34.53	0.944
PAMS	×2	8.00	31.96	0.927	8.00	32.72	0.928	8.00	34.33	0.943
CADyQ	×2	6.46	31.58	0.923	6.10	32.61	0.926	6.02	34.19	0.942
CABM	×2	5.46	31.54	0.923	5.33	32.55	0.925	5.23	34.16	0.942
Granular-DQ (Ours)	×2	**4.11**	**31.94**	**0.927**	**4.17**	**32.52**	**0.925**	**4.12**	**34.52**	**0.944**
EDSR	×2	32.00	31.97	0.927	32.00	32.75	0.928	32.00	34.38	0.943
PAMS	×2	8.00	31.96	0.927	8.00	32.72	0.928	8.00	34.33	0.943
CADyQ	×2	6.15	31.95	0.927	5.68	32.70	0.928	5.59	34.30	0.943
CABM	×2	5.59	31.92	0.927	5.39	32.74	0.927	5.31	34.33	0.943
Granular-DQ (Ours)	×2	**4.60**	**32.01**	**0.928**	**4.40**	**32.57**	**0.925**	**4.27**	**34.42**	**0.944**
IDN	×2	32.00	31.29	0.920	32.00	32.42	0.924	32.00	34.02	0.940
PAMS	×2	8.00	31.39	0.921	8.00	32.46	0.925	8.00	34.05	0.941
CADyQ	×2	5.22	31.54	0.923	4.67	32.51	0.925	4.57	34.10	0.941
CABM	×2	4.21	31.40	0.921	4.19	32.50	0.925	4.19	34.10	0.941
Granular-DQ (Ours)	×2	**4.01**	**31.63**	**0.924**	**4.05**	**32.36**	**0.922**	**4.05**	**34.35**	**0.942**
SwinIR-light	×2	32.00	32.71	0.934	32.00	32.81	0.928	32.00	34.81	0.946
PAMS	×2	8.00	32.40	0.931	8.00	32.68	0.927	8.00	34.68	0.945
CADyQ	×2	5.29	31.88	0.926	5.07	32.50	0.924	5.06	34.48	0.943
CABM	×2	5.14	31.93	0.927	4.98	32.52	0.925	4.97	34.50	0.944
Granular-DQ (Ours)	×2	**4.76**	**32.54**	**0.932**	**4.73**	**32.73**	**0.927**	**4.12**	**34.52**	**0.944**
HAT-S	×2	32.00	34.19	0.945	32.00	33.28	0.934	32.00	35.30	0.950
PAMS	×2	8.00	33.63	0.941	8.00	33.12	0.932	8.00	35.12	0.949
CADyQ	×2	5.43	33.13	0.938	5.32	32.95	0.930	5.22	34.95	0.947
CABM	×2	5.34	33.09	0.937	5.26	32.94	0.930	5.18	34.95	0.947
Granular-DQ (Ours)	×2	**4.80**	**33.71**	**0.942**	**4.78**	**33.12**	**0.932**	**4.77**	**35.12**	**0.949**

Table S3: Quantitative comparison (FAB, PSNR (dB)/SSIM) with full precision models, PAMS, CADyQ, CABM and our method on Urban100, Test2K, Test4K for ×2.

Limitation

While Granular-DQ effectively maintains promising SR performance with dramatic computational overhead reduction, it still has several limitations. First, the mixed-precision solution of Granular-DQ makes it require specific hardware design and operator support to achieve true compression acceleration. Second, its efficacy in accelerating processing for super-resolving large-size images is modest at best. In future work, we will design more efficient and effective quantization approaches to overcome these limitations.

References

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22367–22377.

Hong, C.; Baik, S.; Kim, H.; Nah, S.; and Lee, K. M. 2022. Cadyq: Content-aware dynamic quantization for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 367–383. Springer.

Hong, C.; and Lee, K. M. 2024. AdaBM: On-the-Fly Adaptive Bit Mapping for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2641–2650.

Lee, H.; Yoo, J.-S.; and Jung, S.-W. 2024. RefQSR: Reference-based Quantization for Image Super-Resolution Networks. *IEEE Transactions on Image Processing*, 33: 2823–2834.

Li, H.; Yan, C.; Lin, S.; Zheng, X.; Zhang, B.; Yang, F.; and Ji, R. 2020. Pams: Quantized super-resolution via parameterized max scale. In *Proceedings of the European Conference on Computer Vision*, 564–580. Springer.

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

Qin, H.; Zhang, Y.; Ding, Y.; Liu, X.; Danelljan, M.; Yu, F.; et al. 2024. QuantSR: Accurate Low-bit Quantization for Efficient Image Super-Resolution. In *Advances in Neural Information Processing Systems*.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3): 379–423.

Tian, S.; Lu, M.; Liu, J.; Guo, Y.; Chen, Y.; and Zhang, S. 2023. CABM: Content-Aware Bit Mapping for Single Image Super-Resolution Network With Large Input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1756–1765.

Tu, Z.; Hu, J.; Chen, H.; and Wang, Y. 2023. Toward Accurate Post-Training Quantization for Image Super Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5856–5865.

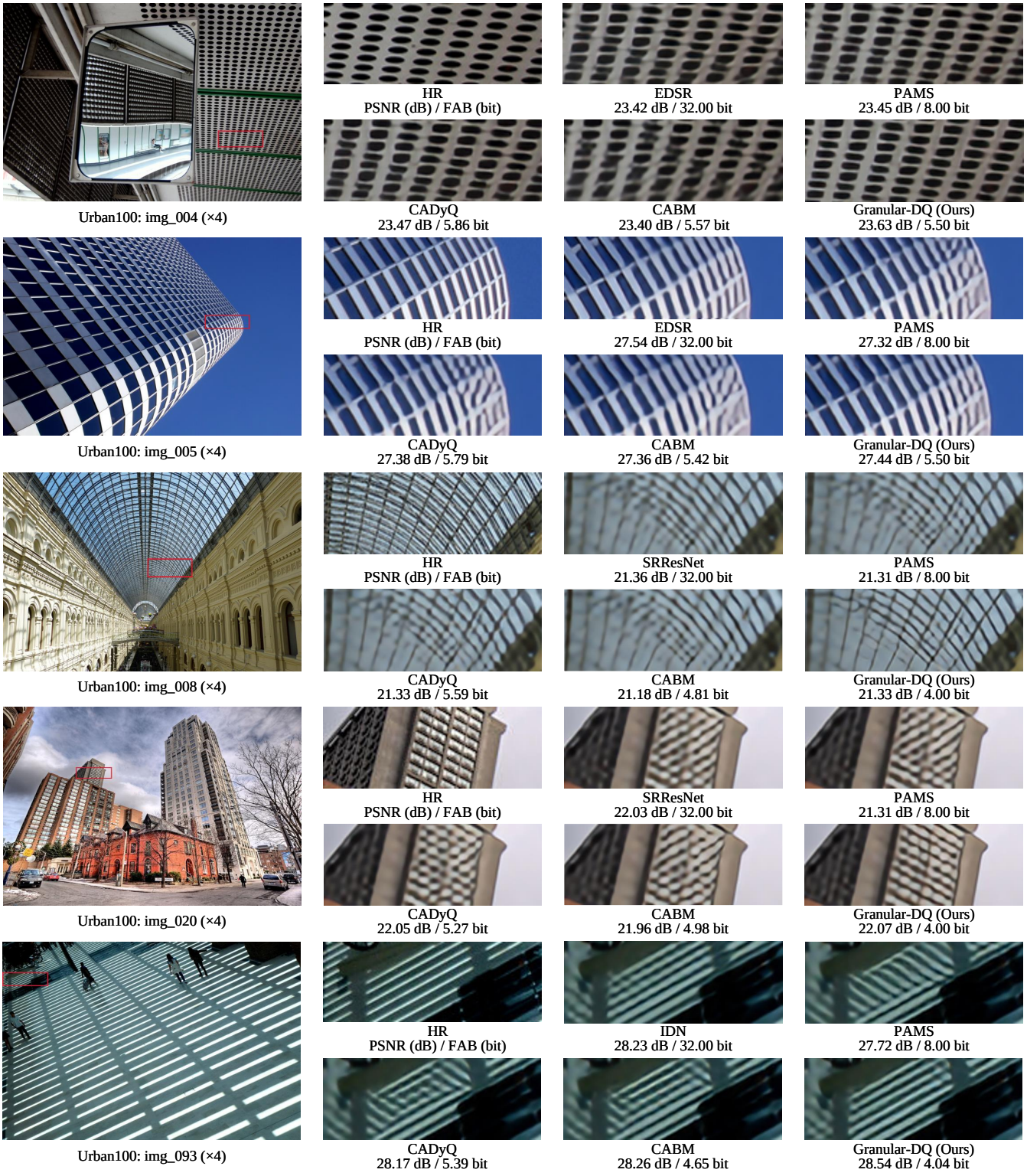


Figure S2: More visual comparison (×4) on Urban100 (×4) for different methods.

Figure S3: More visual comparison (×4) on Test2K (×4) for different methods.