



PassionSR: Post-Training Quantization with Adaptive Scale in One-Step Diffusion based Image Super-Resolution

Libo Zhu¹, Jianze Li¹, Haotong Qin², Wenbo Li³, Yulun Zhang¹, Yong Guo⁴, and Xiaokang Yang¹

¹Shanghai Jiao Tong University, ²ETH Zürich, ³Chinese University of Hong Kong, Max Planck Institute for Informatics, Co., Ltd,

Introduction

Diffusion models achieve strong performance in real-world super-resolution tasks but remain costly in storage and computation. To address this, we propose **PassionSR**, a post-training quantization method tailored for one-step diffusion models in super-resolution.

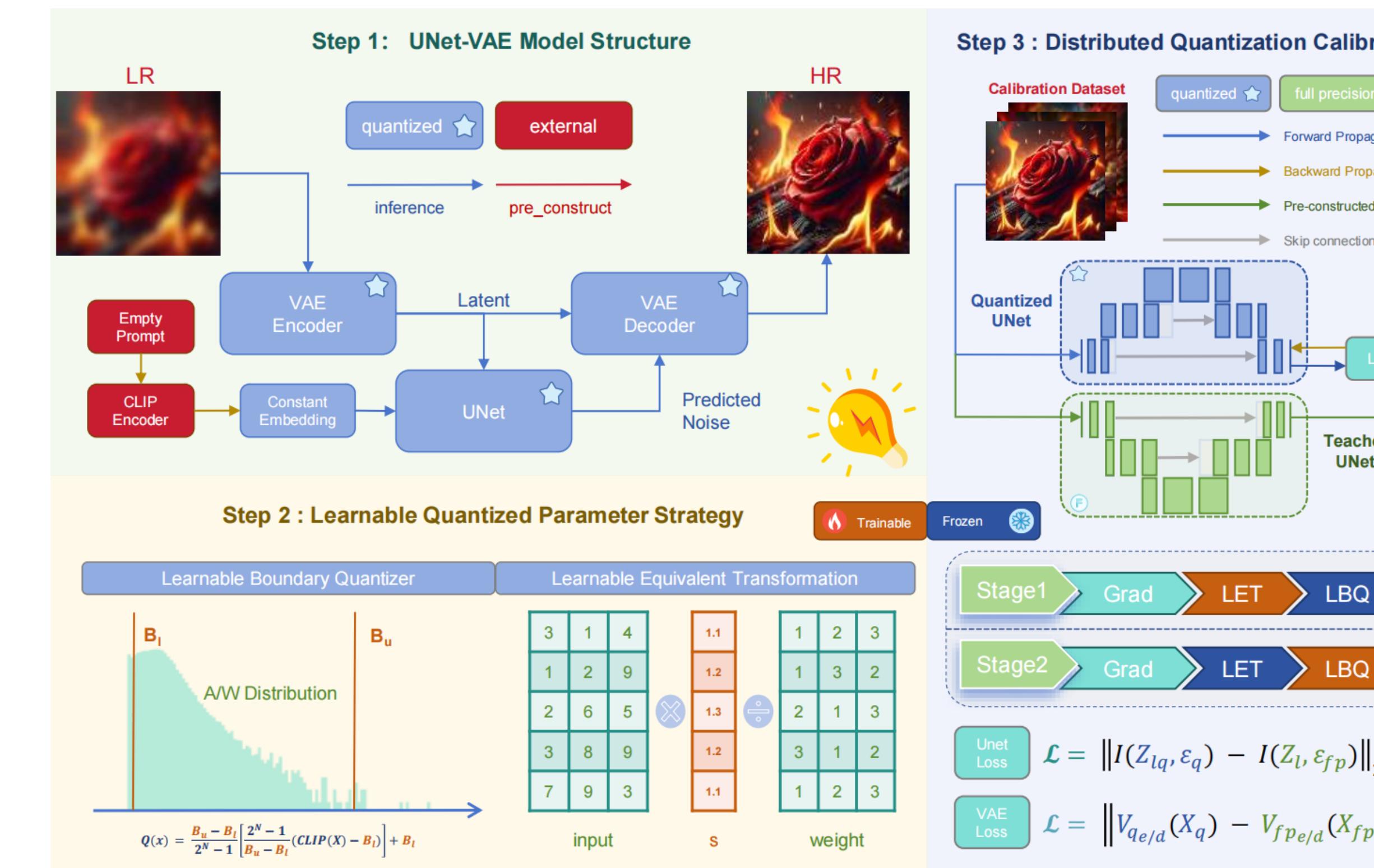
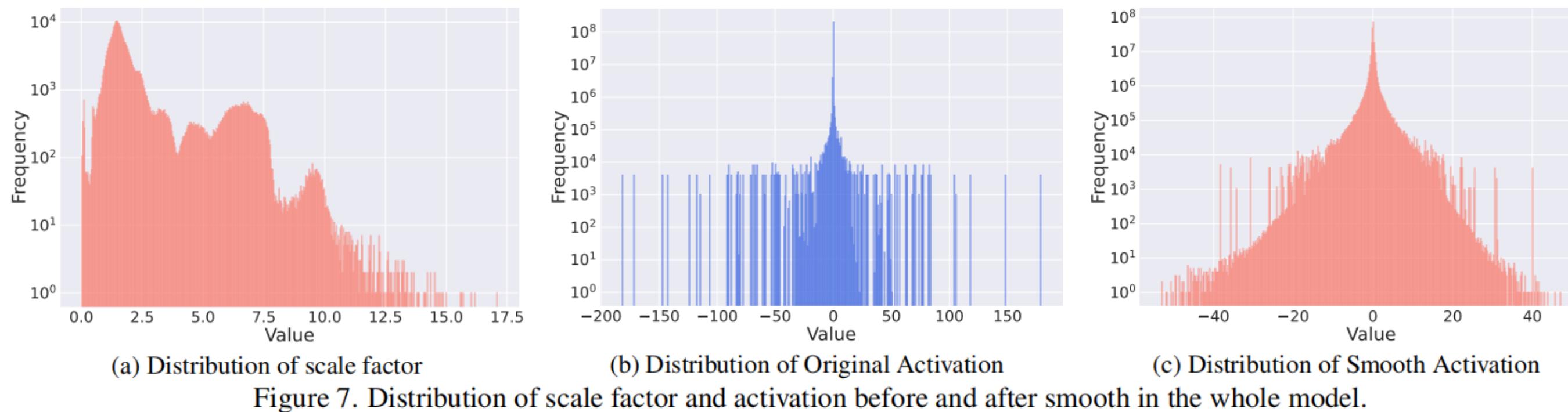


Contribution

- **PassionSR** is the first **low-bit (6/8 bit) PTQ** model for one-step diffusion super-resolution.
- It adopts a simplified effective **UNet-VAE** model architecture as full-precision version.
- Two key techniques, **LET** and **DQC**, are introduced to enhance quantization performance, stability and efficiency.
- PassionSR delivers perceptual quality **close to full precision** and **outperforms** existing quantization methods.



Methods



➤ Linear Equivalent Transformation :

The **input matrix X** shaped as $\mathbb{R}^{N \times C_{in}}$, utilizes **weight matrix $W \in \mathbb{R}^{C_{in} \times C_{out}}$** and **bias matrix $B \in \mathbb{R}^{1 \times C_{out}}$** to calculate output matrix $Y \in \mathbb{R}^{N \times C_{out}}$.

To introduce **learnable equivalent transformation**, we use **learnable scale factor $s \in \mathbb{R}^{1 \times C_{in}}$** and **learnable bias $\delta \in \mathbb{R}^{1 \times C_{in}}$** to transform input X :

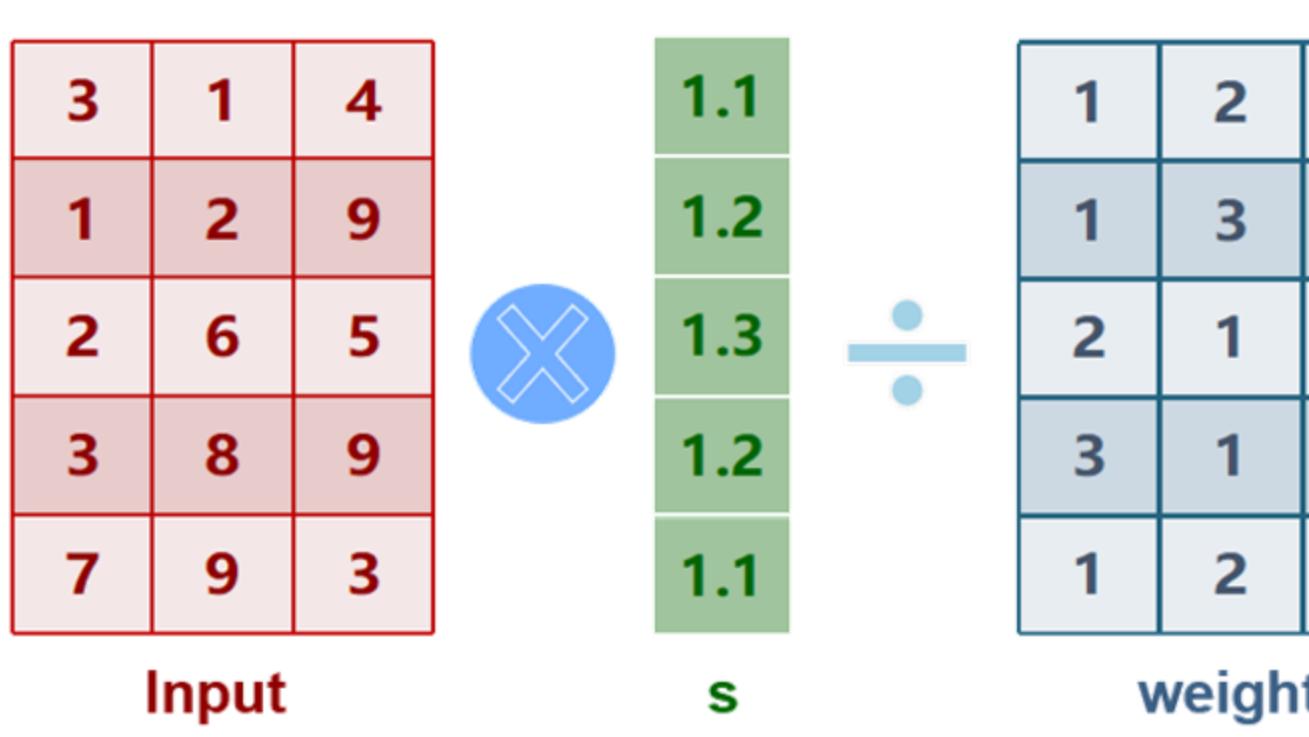
$$\tilde{W} = s \odot W, \tilde{X} = (X - \delta) \oslash s, \tilde{B} = B + \delta W$$

where \odot, \oslash represent element-wise multiplication and division.

During quantization process, the output of full-precision operation and low-bit quantization,

$$\begin{cases} Y_q = Q_a(\tilde{X})Q_w(\tilde{W}) + Q_b(\tilde{B}) \\ Y_{fp} = \tilde{X}\tilde{W} + \tilde{B} = XW + B = Y \end{cases}$$

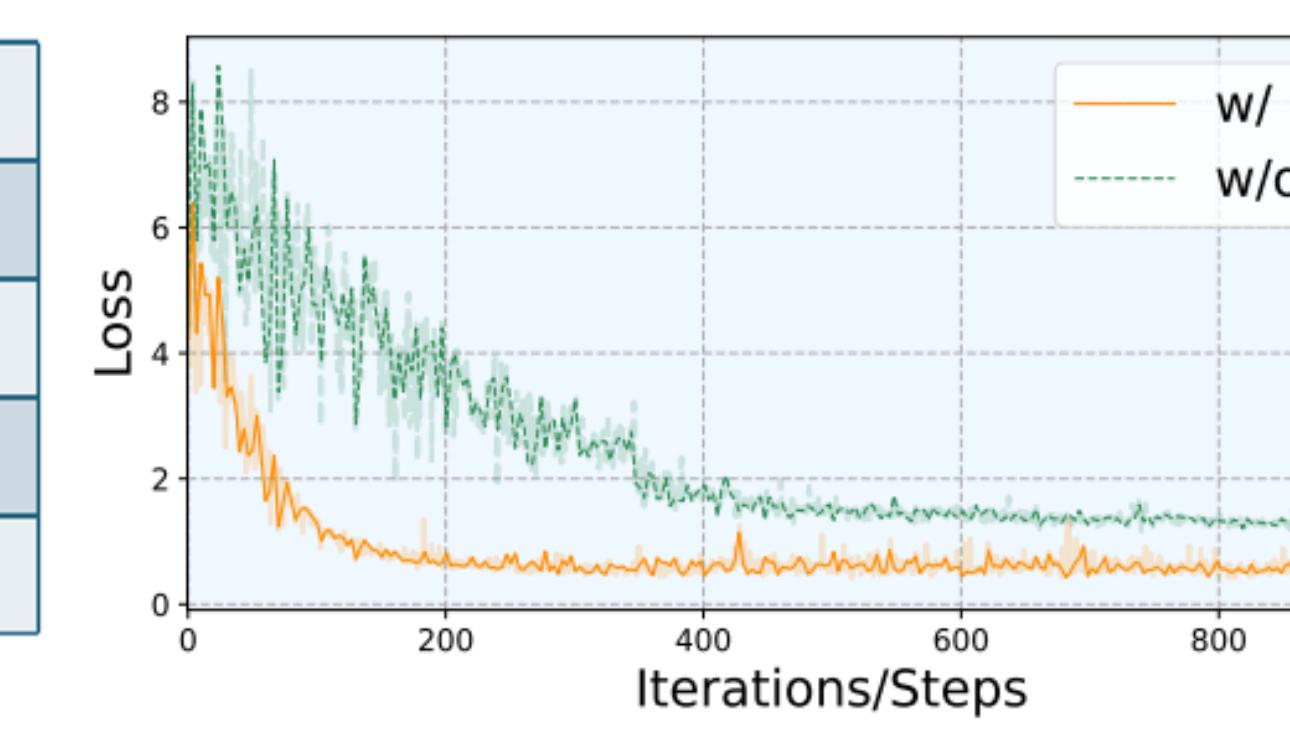
where Q_a, Q_w, Q_b represent the quantization operation on activation, weight and bias.



➤ Distributed Quantization Calibration :

Due to the discontinuous nature of the rounding function, the training process of model quantization often suffers from instability—particularly when simultaneously calibrating the boundaries of **low-bit quantization (LBQ)** and the scale factors in **learnable equivalent transformations (LET)**.

To address this, we propose a **Distributed Quantization Calibration (DQC)** strategy that splits the calibration into two sequential stages. After updating the scale and offset parameters of LET in the first stage, LBQ is re-initialized to adapt to the updated quantization vectors. This DQC strategy significantly accelerates convergence, stabilizes training process, and enhances the quantized model's performance meanwhile.



Experiments

➤ Ablation Study

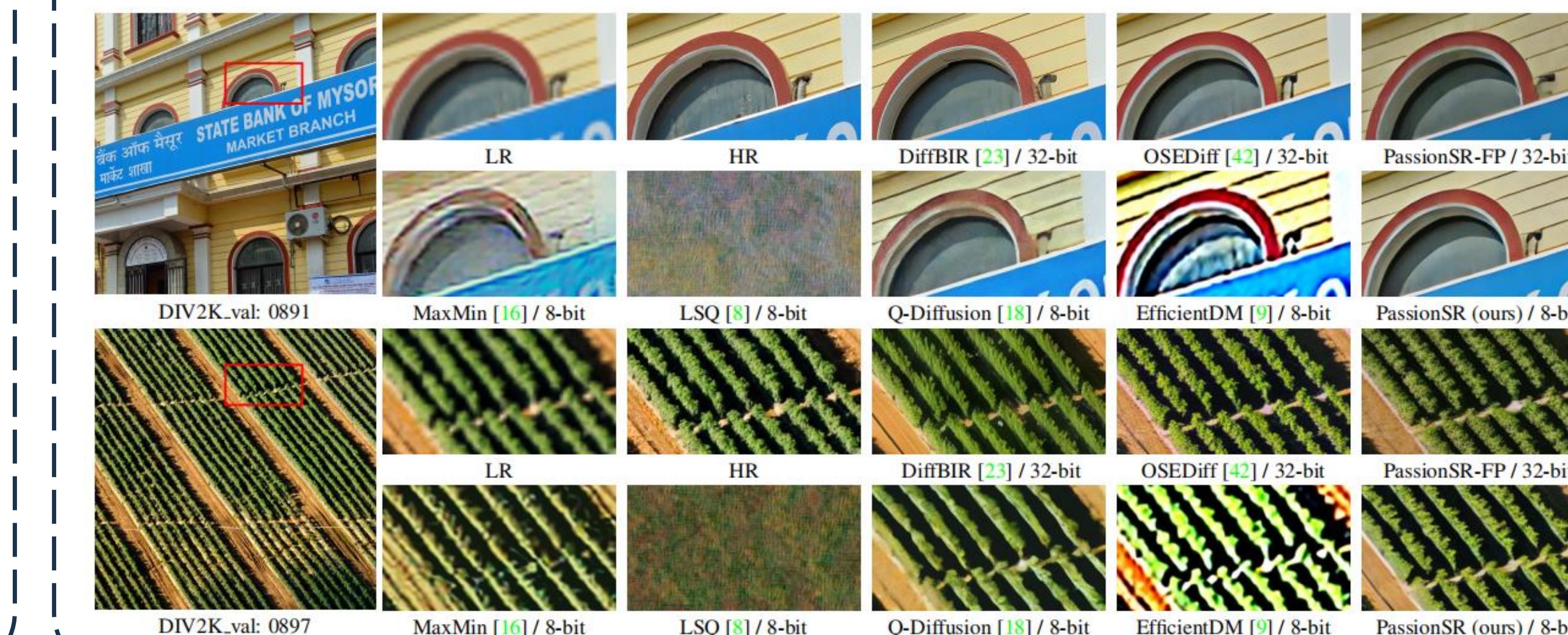
Methods	Efficiency Time (h)	GPU (GB)	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	RealSR NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑
MaxMin	0.00	0	15.55	0.2417	0.8018	0.4449	9.263	42.15	0.2791	0.4174
LBQ	2.66	40	23.15	0.6621	0.5022	0.3115	7.234	47.75	0.3071	0.4787
LBQ+LET	3.87	40	25.40	0.7529	0.3798	0.2584	6.604	44.26	0.2414	0.3224
LBQ+LET+DQC	1.07	28	24.41	0.7374	0.3427	0.2419	5.449	55.08	0.3083	0.4849

Table 4. Ablation study on our proposed components: LBQ, LET, and DQC. Our ablation experiments are in the setting of W6A6 UNet-VAE quantization. We test each ablation method on RealSR and record their calibration time and GPU costs.

➤ Quantitative Results

Datasets	Bits	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑
RealSR	W32A32	OSEDiff [42]	25.27	0.7379	0.3027	0.1808	4.355	67.43	0.4766	0.6835
		PassionSR-FP	25.39	0.7460	0.2984	0.1813	4.453	67.05	0.4680	0.6796
	W8A8	MaxMin [12]	23.16	0.6875	0.5463	0.2879	7.932	32.92	0.1849	0.2363
		LSQ [8]	15.39	0.3375	0.9944	0.5427	10.08	50.11	0.3533	0.3173
	W6A6	Q-Diffusion [18]	24.88	0.6967	0.4993	0.2696	8.437	44.69	0.2352	0.5604
DRealSR	W32A32	EfficientDM [9]	14.77	0.4253	0.5478	0.3462	7.526	44.75	0.2568	0.4000
		PassionSR (ours)	25.67	0.7499	0.3140	0.1932	5.654	65.88	0.4437	0.6912
	W8A8	MaxMin [12]	15.55	0.2417	0.8018	0.4449	9.263	42.15	0.2791	0.4174
		LSQ [8]	13.73	0.1081	1.0900	0.5450	8.430	53.61	0.3036	0.4396
	W6A6	Q-Diffusion [18]	19.75	0.4727	0.6877	0.4024	7.381	56.46	0.4380	0.6439
RealSR	W32A32	PassionSR (ours)	25.15	0.7196	0.4199	0.2592	8.618	44.43	0.2131	0.4612
	W8A8	MaxMin [12]	24.97	0.7989	0.5091	0.2921	8.215	24.05	0.1846	0.3163
		LSQ [8]	14.56	0.1795	1.1661	0.592	10.19	29.07	0.4010	0.3970
	W6A6	Q-Diffusion [18]	27.14	0.7184	0.4765	0.2895	9.861	26.44	0.2284	0.5608
		PassionSR (ours)	27.41	0.8146	0.3422	0.1918	6.070	33.56	0.4286	0.7554
DRealSR	W32A32	MaxMin [12]	13.08	0.2291	0.8131	0.5077	10.51	35.83	0.2702	0.3864
		LSQ [8]	12.95	0.0934	1.1890	0.5833	8.591	26.39	0.2911	0.5600
	W8A8	Q-Diffusion [18]	21.75	0.6096	0.7008	0.4039	6.854	24.39	0.4109	0.6696
		EfficientDM [9]	15.07	0.4287	0.6127	0.357	6.690	28.37	0.2351	0.3973
	W6A6	PassionSR (ours)	26.62	0.7984	0.4429	0.2571	8.484	26.26	0.1824	0.4358

➤ Visual Results



DIV2K_val: 0891 MaxMin [16] / 8-bit LSQ [8] / 8-bit Q-Diffusion [18] / 8-bit EfficientDM [9] / 8-bit PassionSR (ours) / 8-bit

DIV2K_val: 0897 MaxMin [16] / 32-bit LSQ [8] / 32-bit Q-Diffusion [18] / 32-bit EfficientDM [9] / 32-bit PassionSR (ours) / 32-bit