

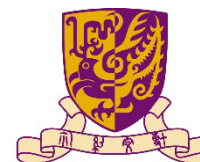
PassionSR: Post-Training Quantization with Adaptive Scale in One-Step Diffusion based Image Super-Resolution

Libo Zhu¹, Jianze Yang¹, 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,



ETH zürich



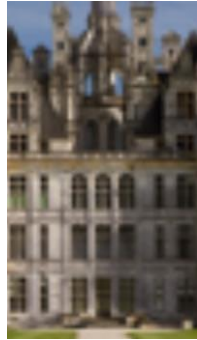
香港中文大學
The Chinese University of Hong Kong



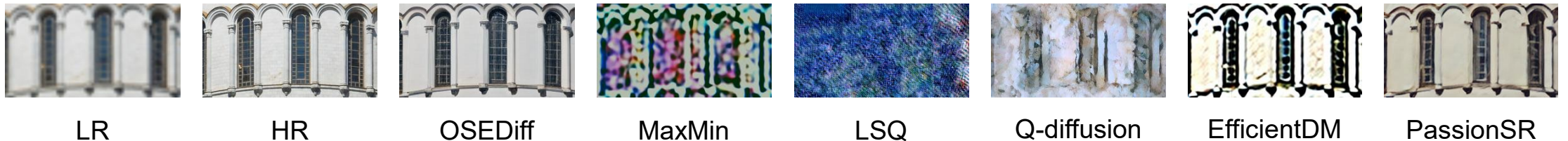
MAX PLANCK INSTITUTE
FOR INFORMATICS

Motivation

- **Diffusion Models** excel in Real-World SR tasks but face problems when deployed in resource-constrained environment.
- Steps of Diffusion SR models have been reduced to 1 but they still have large storage and computation consumptions.
- Low-bit quantization reduces computation extremely with severe performance degradation.
- We propose **PassionSR**, a Post-training Quantization method for one-step diffusion model in SR.

			
LR(X4) Step / Bits Param.(M)/Ops(G)	DiffBIR 50 / 32-bit 1,618 / 49,056	OSEDiff 1 / 32-bit 1,303 / 4,523	PassionSR 1 / 8-bit 238 / 1,060

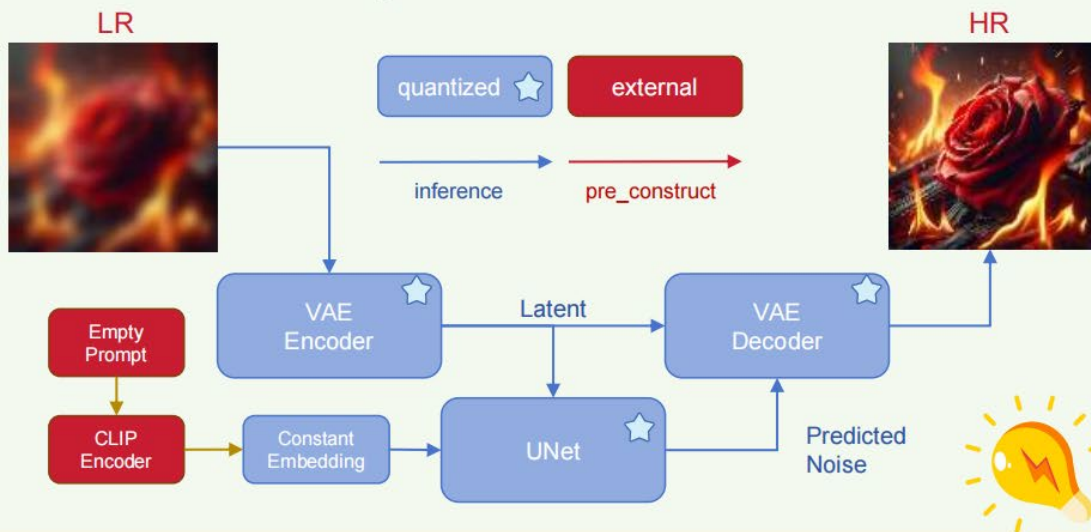
W6A6 comparison



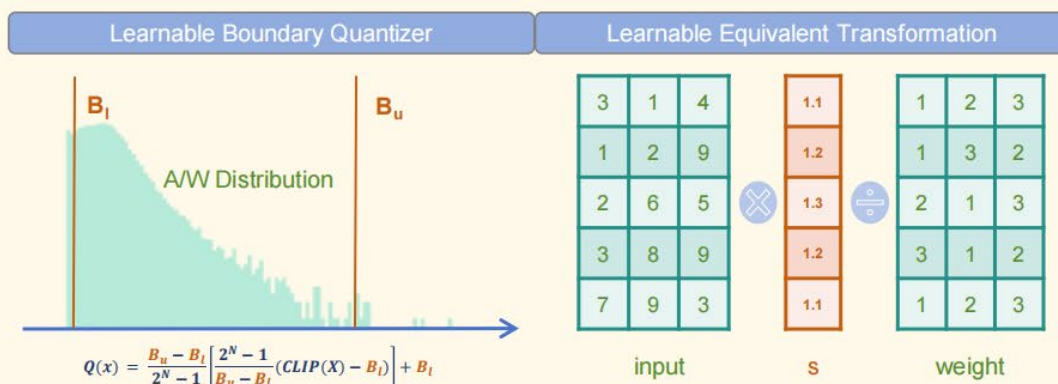
Method-Overview



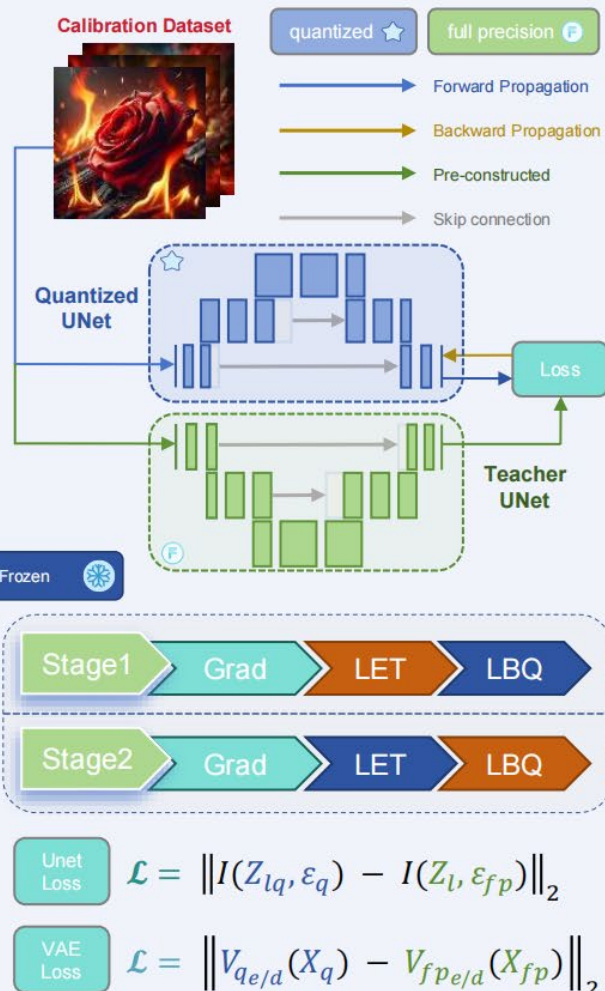
Step 1: UNet-VAE Model Structure



Step 2: Learnable Quantized Parameter Strategy



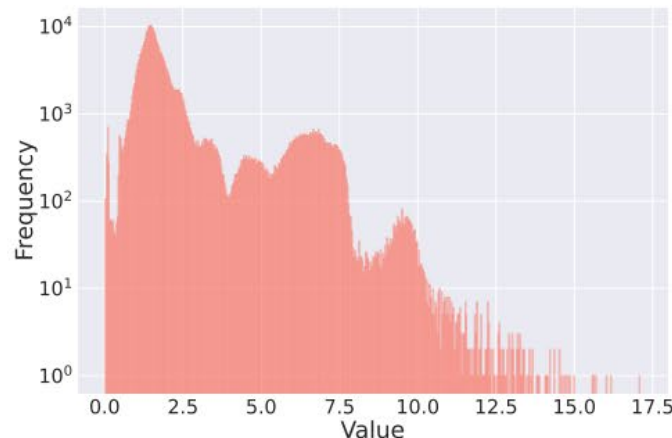
Step 3: Distributed Quantization Calibration



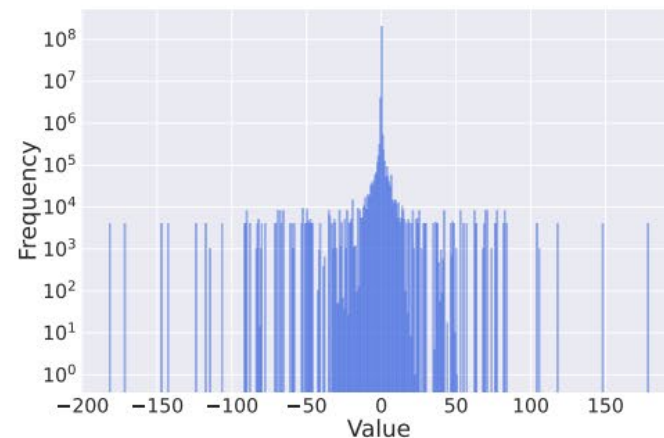
Overall

- **Step 1:** We simplify the structure of OSediff by removing the components **DAPE** and **CLIP Encoder** to obtain PassionSR-FP.
- **Step 2:** We use a quantizer with two key trainable parts, including a **learnable boundary quantizer** and a **learnable equivalence transform**.
- **Step 3:** We design a **distributed calibration strategy** and a **special loss function** to accelerate the convergence of the quantization calibration.

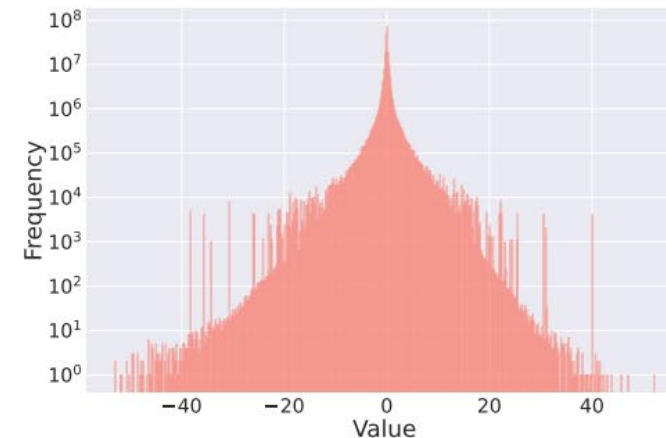
Method-LET



(a) Distribution of scale factor



(b) Distribution of Original Activation



(c) Distribution of Smooth Activation

LET

3	1	4
1	2	9
2	6	5
3	8	9
7	9	3

Input



1.1
1.2
1.3
1.2
1.1

s



1	2	3
1	3	2
2	1	3
3	1	2
1	2	3

weight

Linear Quantization 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.

Method-DQC

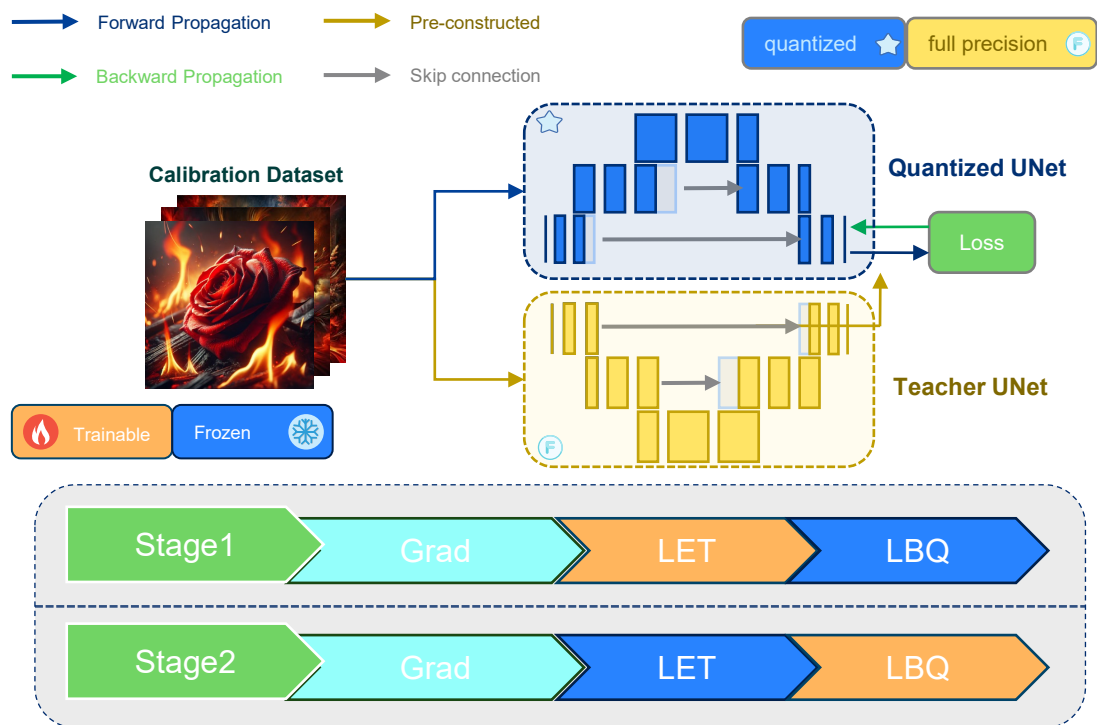


Figure 5. Loss comparison between w/ and w/o DQC

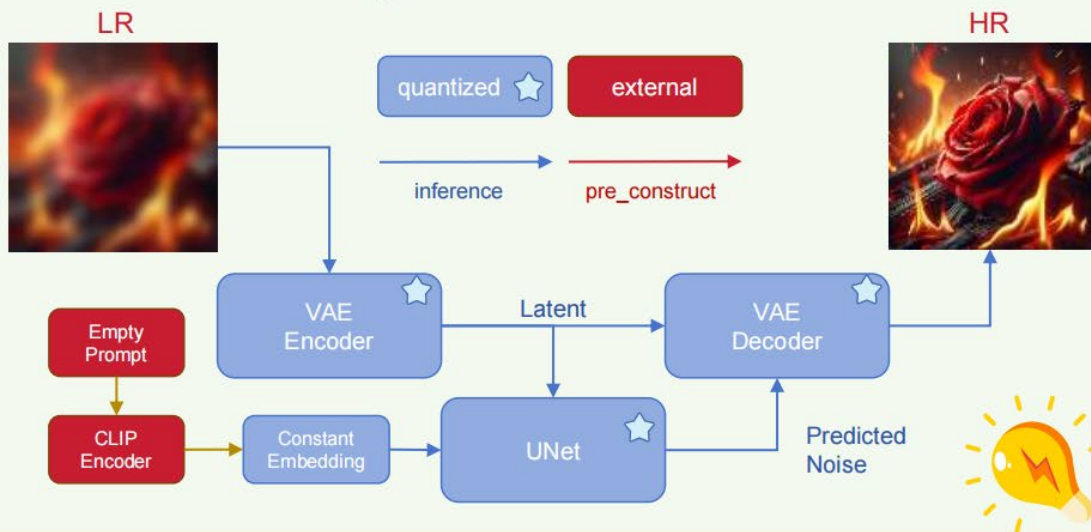
DQC

- 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 and stabilizes training.

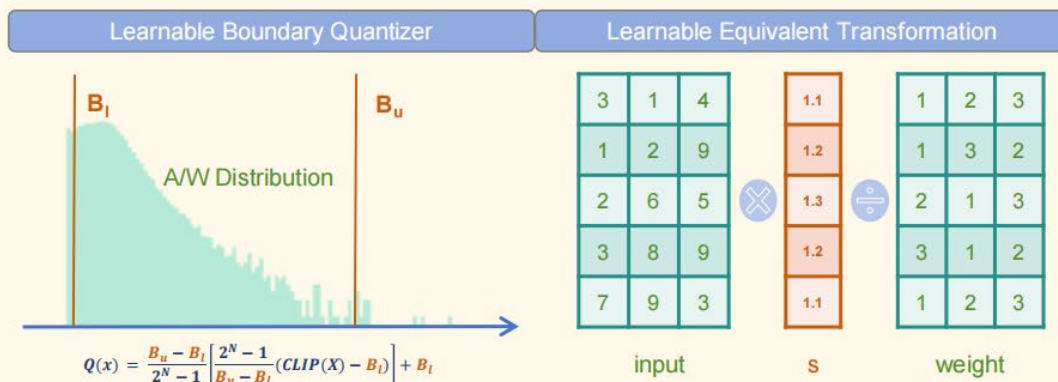
Method-Overview



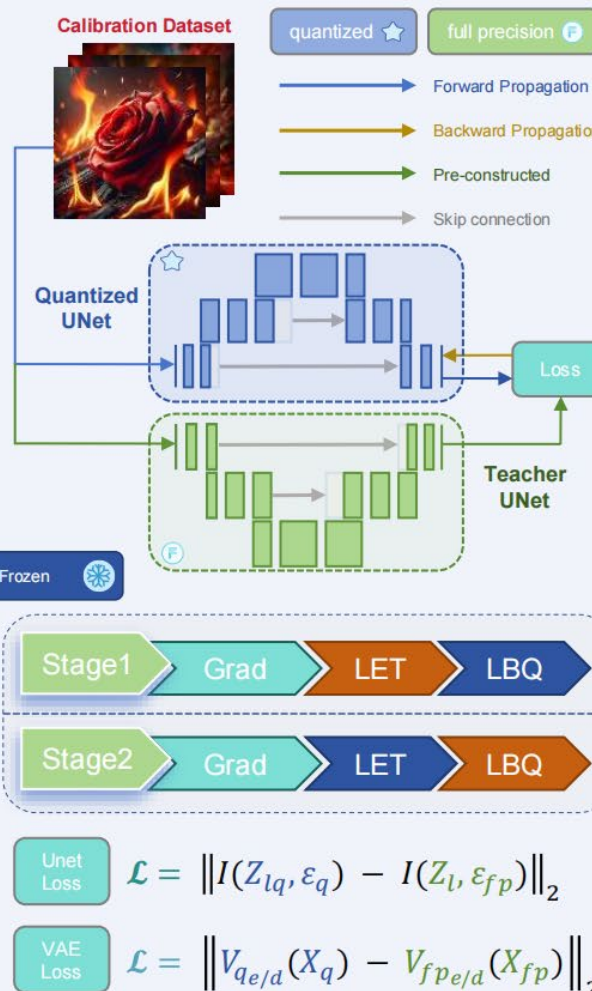
Step 1: UNet-VAE Model Structure



Step 2: Learnable Quantized Parameter Strategy



Step 3: Distributed Quantization Calibration



Overall

- **Step 1:** We simplify the structure of OSediff by removing the components **DAPE** and **CLIP Encoder** to obtain PassionSR-FP.
- **Step 2:** We use a quantizer with two key trainable parts, including a **learnable boundary quantizer** and a **learnable equivalence transform (LET)**.
- **Step 3:** We design a **distributed quantization calibration (DQC)** strategy and a **special loss function** to accelerate the convergence of the quantization calibration.

Experiments



Datasets	Bits	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIP-IQA↑
RealSR	W32A32	OSDiff [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
		Q-Diffusion [18]	24.88	0.6967	0.4993	0.2696	8.437	44.69	0.2352	0.5604
		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
	W6A6	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
		Q-Diffusion [18]	19.75	0.4727	0.6877	0.4024	7.381	56.46	0.4380	0.6439
		EfficientDM [9]	14.75	0.4386	0.5233	0.3451	7.497	42.97	0.2498	0.3740
		PassionSR (ours)	25.15	0.7196	0.4199	0.2592	8.618	44.43	0.2131	0.4612
DRealSR	W32A32	OSDiff [42]	25.57	0.7885	0.3447	0.1808	4.371	37.22	0.4794	0.7540
		PassionSR-FP	26.70	0.7978	0.3339	0.1765	4.336	37.03	0.4686	0.7520
	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
		Q-Diffusion [18]	27.14	0.7184	0.4765	0.2895	9.861	26.44	0.2284	0.5608
		EfficientDM [9]	15.55	0.4183	0.6291	0.3555	6.859	28.61	0.2468	0.4150
		PassionSR (ours)	27.41	0.8146	0.3422	0.1918	6.070	33.56	0.4286	0.7554
	W6A6	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
		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
		PassionSR (ours)	26.62	0.7984	0.4429	0.2571	8.484	26.26	0.1824	0.4358
DIV2K_val	W32A32	OSDiff [42]	24.95	0.7154	0.2325	0.1197	3.616	68.92	0.4340	0.6842
		PassionSR-FP	25.16	0.7221	0.2373	0.1185	3.573	69.27	0.4402	0.6958
	W8A8	MaxMin [12]	22.33	0.6618	0.5639	0.2731	7.563	33.68	0.1913	0.2818
		LSQ [8]	13.90	0.2537	0.9932	0.5515	9.578	48.11	0.3512	0.3246
		Q-Diffusion [18]	24.20	0.6813	0.3997	0.2400	7.955	51.95	0.2709	0.6243
		EfficientDM [9]	15.24	0.4954	0.6041	0.3374	6.856	48.78	0.2685	0.4235
		PassionSR (ours)	25.11	0.7199	0.2496	0.1277	4.424	67.92	0.3993	0.6939
	W6A6	MaxMin [12]	11.66	0.1606	0.8509	0.4966	11.30	45.47	0.2764	0.3523
		LSQ [8]	12.21	0.0858	1.0695	0.5424	8.564	52.74	0.2872	0.4692
		Q-Diffusion [18]	18.92	0.4939	0.6227	0.3718	6.162	51.50	0.3946	0.5814
		EfficientDM [9]	15.09	0.4991	0.5953	0.3292	6.900	46.01	0.2570	0.4007
		PassionSR (ours)	24.34	0.7097	0.3440	0.2075	7.039	51.19	0.2267	0.4802

Table 2. Quantitative UNet-VAE quantization experiments results. PassionSR-FP is used as full-precision backbones rather than original OSDiff. W8A8 denotes 8 bit weight and 8 bits activation quantization. The best results in the same setting are colored with red.

Quantitative

- In the **UNet-VAE quantization** experiments, PassionSR significantly outperforms existing methods in the W8A8 and W6A6 settings.
- For **8-bit quantization**, PassionSR performs close to or even better than full-precision OSDiff across the datasets, whereas the other quantization methods show a significant degradation in the quantitative evaluation.
- For **6-bit quantization**, comparative methods scored lower on structural metrics (e.g., PSNR\SSIM), but higher on the unreferenced IQA assessment. PassionSR achieved the highest reference IQA scores and relatively low non-reference IQA scores.
- Images with high quantization noise still achieved high scores on the no-reference IQA metrics, explaining why some quantization methods have poor visual quality even though they perform better on the no-reference IQA metrics.

Experiments



Figure 6. Visual comparison ($\times 4$) with high-resolution image, full-precision model's output and different quantization methods in some challenging cases at W8A8 and W6A6 UNet-VAE quantization. PassionSR gains significant visual advantages over other methods.

Compression

Method	Bit	Params / M ($\downarrow$ Ratio)	Ops / G ($\downarrow$ Ratio)
OSDiff	W32A32	1,303 ($\downarrow 0\%$)	4,523 ($\downarrow 0\%$)
PassionSR-FP	W32A32	949 ($\downarrow 27.13\%$)	4,240 ($\downarrow 6.25\%$)
PassionSR-U	W8A8	300 ($\downarrow 76.96\%$)	3,732 ($\downarrow 17.50\%$)
	W6A6	246 ($\downarrow 81.11\%$)	3,689 ($\downarrow 18.44\%$)
PassionSR-UV	W8A8	238 ($\downarrow 81.77\%$)	1,060 ($\downarrow 76.56\%$)
	W6A6	178 ($\downarrow 86.32\%$)	795 ($\downarrow 82.42\%$)

Table 3. Compression ratio of different quantization settings. PassionSR-U refers to UNet-only quantization while PassionSR-UV refers to UNet-VAE quantization.

Visual

- A visual comparison of the UNet-VAE quantification ($\times 4$) is shown, where we have selected several challenging .
- PassionSR generates sharper results and better textures than other methods, indication its superior performance.
- Notably, in some cases PassionSR even outperforms the full-precision version PassionSR-FP.

Experiments



Methods	Efficiency						RealSR			
	Time (h)	GPU (GB)	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	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.

Ablation

- All components have benefits in terms of model performance improvement.
- LET brings huge performance improvement.
- DQC stabilizes the LET’s calibration, delivering faster convergence and lower GPU memory overheads

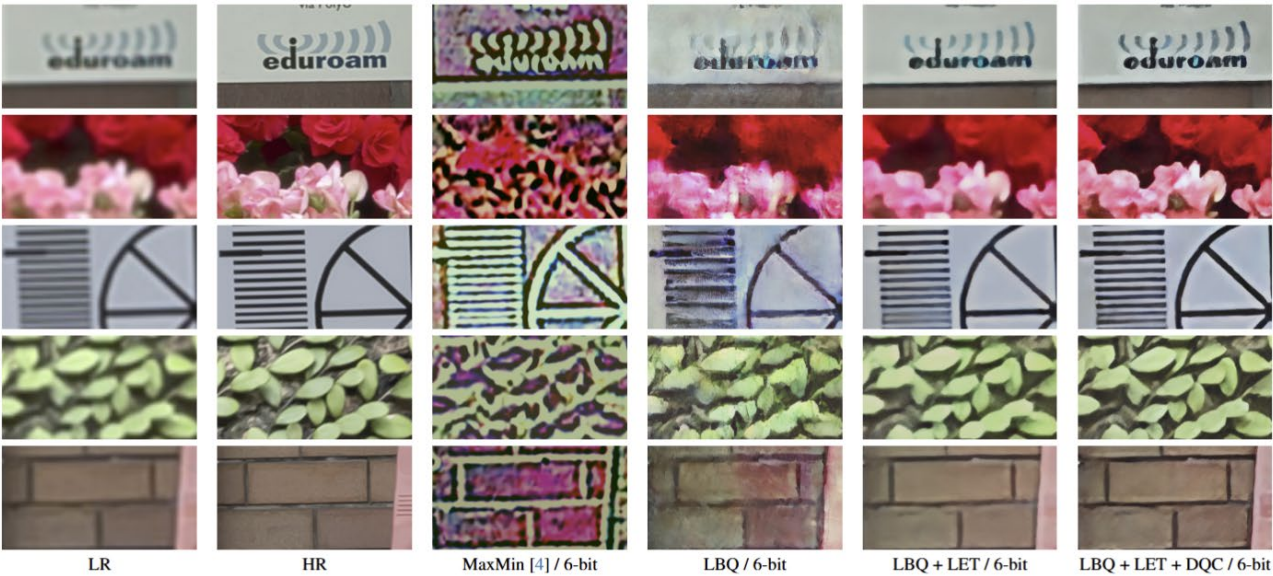


Figure 2. Visual comparison (×4) with low-resolution, high-resolution images and different quantization settings in ablation study.

Contribution

We propose **PassionSR**, a post-training quantization method for one-step diffusion-based SR model.

- **Simplified Structure:** simplified diffusion SR model only consists of the core UNet and VAE.
- **LET & DQC:** LET improves model performance largely and DQC stabilizes calibration process.
- **Performance:** Outperforms SOTA diffusion quantization methods for SR.

Poster

- Time: XXX
XXX



Project

Thanks