QuEST: Low-bit Diffusion Model Quantization via Efficient Selective Finetuning

Supplementary Material

The supplementary material is organized as follows: Sec. 6 provides comparison with TFMQ-DM; Sec. 7 provides comparison on the low-resolution dataset; Sec. 8 provides the proof and detailed analysis for Theorem 3.2; Sec. 9 presents additional examples of the imbalanced distributions across different models; Sec. 10 highlights the importance of the large values in activations; Sec. 11 offers further generated examples from our method across varying bit-widths; and Sec. 12 discusses limitations and broader considerations.

6. More Baseline Comparisons

We further compare with TFMQ [16] below:

Bedroom	W8A8	W4A8
TFMQ-DM	3.14	3.68
QuEST	**3.03**	**3.26**
ImageNet	W8A8	W4A8
TFMQ-DM	10.79	10.29
QuEST	**10.43**	**8.48**

Table 9. Comparing TFMQ.

We also supplement the metrics for Table 3:

W8A8	sFID ↓	IS ↑
QDiffusion	8.19	2.25
PTQD	9.89	2.25
EfficientDM	N/A	N/A
Ours	**6.86**	**2.27**
W4A4	sFID ↓	IS ↑
QDiffusion	N/A	N/A
PTQD	N/A	N/A
EfficientDM	15.15	**2.27**
Ours	**7.82**	2.26

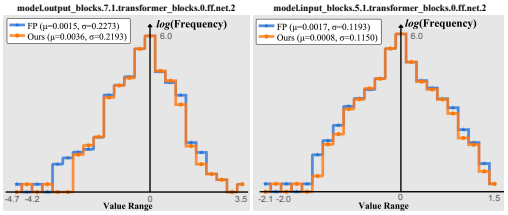
Table 10. Additional metrics on LSUN-Bedrooms. "N/A" represents generation failure.

7. Low-resolution dataset comparison

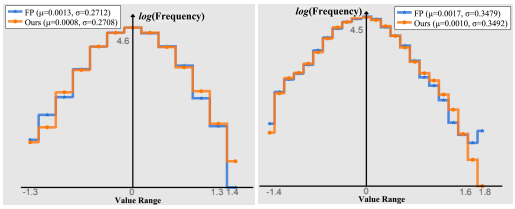
We further include experiments on CIFAR10 in Tab. 11.

	W8A8	W4A4
Q-Diffusion	3.75	N/A
EfficientDM	3.75	10.48
QuEST	**3.71**	**9.37**

Table 11. FID comparison on CIFAR10.



(a) Activation Distribution on Conditional LDM4 (ImageNet 256 × 256)



(b) Activation Distribution on Unconditional LDM4 (LSUN-Bedrooms 256 × 256)

Figure 5. Illustrations of imbalanced activation distributions on conditional LDM4 (ImageNet 256×256) and unconditional LDM4 (LSUN-Bedrooms 256×256).

8. Proof for Theorem 3.2

We provide the detailed proof for Theorem 3.2 here. The notations are consistent with the ones in the main paper.

Since the perturbation Δ is too large for accurate Taylor expansion, we can resolve it by introducing a new perturbation $\epsilon = \Delta/K$, where we divide Δ by a constant K so that ϵ is small enough for approximation. Then, Eq. (8) is rewritten as follows:

$$\begin{aligned}
&\mathbb{E}[L(z_{n,t}+\Delta;\mathbf{w})] - \mathbb{E}[L(z_{n,t};\mathbf{w})] \\
&= \sum_{i=1}^{K} \left(\mathbb{E}[L(z_{n,t}+\frac{i}{K}\Delta;\mathbf{w})] - \mathbb{E}[L(z_{n,t}+\frac{i-1}{K}\Delta;\mathbf{w})] \right) \\
&\approx \sum_{i=1}^{K} \left(\epsilon^T \overline{\mathbf{g}}^{(z_{n,t}+(i-1)\epsilon)} + \frac{1}{2}\epsilon^T \overline{\mathbf{H}}^{(z_{n,t}+(i-1)\epsilon)} \epsilon \right), \qquad (11)
\end{aligned}$$

where the approximation step follows Taylor expansion and only the first two main components are kept. The first term in Eq. (11) cannot be ignored because samples such as $z_{n,t}+(i-1)\epsilon$ may not be included in the learned distribution of the model. The second term can still be minimized by reconstruction since only the difference between quantized model output and ground-truth matters. In the following, we temporarily exclude the second term for simplicity since it can always be minimized through aligning the activation outputs.

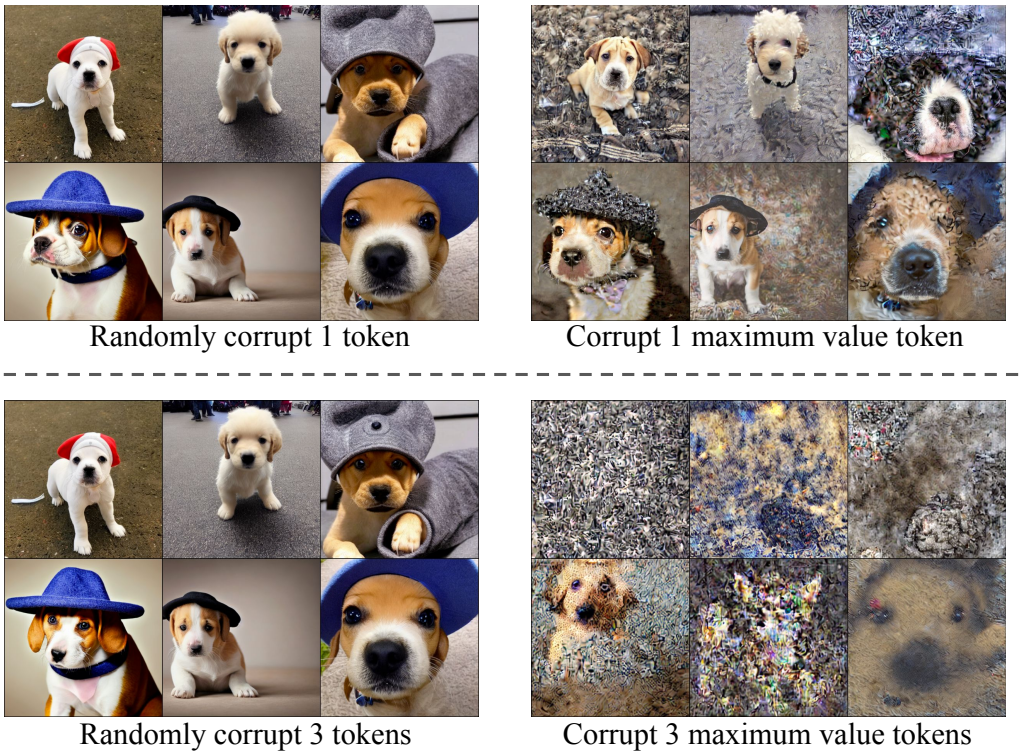


Figure 6. Comparison of different corruptions made on different tokens.

Given the objective function (MSE loss) of diffusion models, we analyze that:

$$\sum_{i=1}^{K} \epsilon^T \overline{\mathbf{g}}^{(z_{n,t}+(i-1)\epsilon)} = 2\epsilon^T \sum_{i=1}^{K} (\tilde{z}^i_{n-1,t} \cdot \mathbf{w}_n - \overline{z}_{n,t}) \approx 2\epsilon^T \sum_{i=1}^{K} (\tilde{z}^i_{n-1,t} \cdot \mathbf{w}_n - z_{\text{FP}}), \quad (12)$$

where $\mathbf{w}_n$ is the weight for layer n, $\tilde{z}^i_{n-1,t}$ is the activation of the $(n-1)$th layer in a quantized model to get $z_{n,t} + (i-1)\epsilon$. Ground-truth $\overline{z}_{n,t}$ can be approximated by the full-precision output z_{FP}. We see that $\tilde{z}^i_{n-1,t}$ and z_{FP} cannot be changed, thus to minimize Eq. (12), we need to finetune $\mathbf{w}_n$. From a general perspective, Eq. (12) also indicates that the model has not converged well to a local minimum given the perturbed inputs, thus when we finetune the model layers given the quantized inputs, we are actually training the model towards convergence over new samples and increasing its robustness.

9. Examples of Imbalanced Activation Distributions

Apart from Fig. 2, we show that the imbalance in the activation distribution is a common phenomenon in different model structures and datasets. In Fig. 5, we show more results of activation distributions of latent diffusion models on ImageNet 256 × 256 and LSUN-Bedrooms 256 × 256.

10. Importance of large values in activations

As shown in Fig. 2, quite a few values are rather large and diversely distributed. These values pose difficulties on activation quantization, and being rather important and not negligible. To demonstrate this, we corrupt certain tokens in the activation outputs of the diffusion model and check the corresponding generated images. The corruption is done by setting the token values as all zeros. As shown in Fig. 6, we compare two settings: (1) corrupt a certain number of tokens randomly; (2) corrupt the same number of the tokens with the largest values.

We see that when corrupting randomly, generation performance is hardly effected. However, corrupting the same amount of tokens (even only one token) with the largest values leads to significantly degenerated images.

11. More generated image examples

11.1. Unconditional Image Generation

The generated images for LSUN-Bedrooms 256×256 under different bit-widths are shown in Fig. 7. Images for LSUN-Churches 256×256 are shown in Fig. 9.

(a) Full Precision

(b) W8A8

(c) W4A8

(d) W4A4

Figure 7. Unconditional image generation examples for LSUN-Bedrooms 256×256.

11.2. Class-conditional image generation

Fig. 10 shows the generated images for 3 different classes.

11.3. Text-to-image generation

Fig. 8 shows the generated images using Stable Diffusion v1.4 under different bit-width.

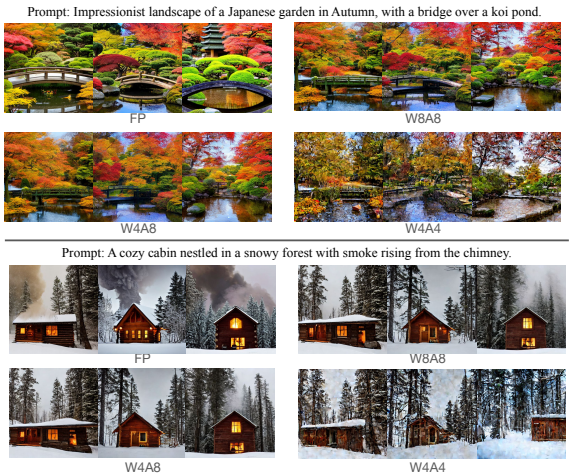


Figure 8. Text-to-image generation results on Stable Diffusion.

12. Limitations and Broader Impacts

The primary objective of this paper is to further the research in enhancing the efficiency of diffusion models. While it confronts societal consequences akin to those faced by research on generative models, it is important to recognize the potential impacts that quantized models could have on current techniques, including watermarking and safety checking. Inappropriate integration of current methodologies may result in unforeseen performance issues, a factor that deserves attention and awareness.

(a) Full Precision

(b) W8A8

(c) W4A8

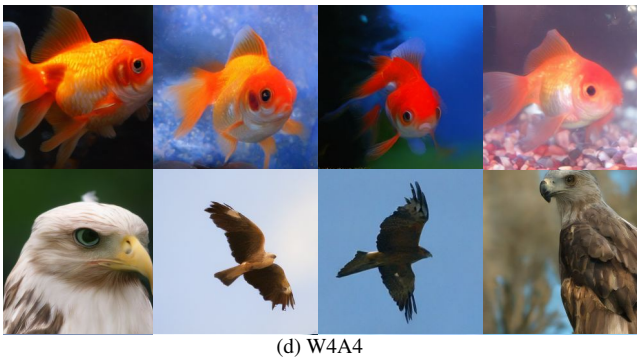
(d) W4A4

Figure 9. Unconditional image generation examples for LSUN-Churches 256×256.

(a) Full Precision

(b) W8A8

(c) W4A8



(d) W4A4

Figure 10. Conditional image generation results for ImageNet 256×256.