Appendix

Note: All references to figures or tables identified by Arabic numerals point to the corresponding figures or tables in the main text.

A. More Implementation Details

In our reconstruction and calibration, apart from the Temporal Information Block proposed by us, the partitioning of the remaining network components remains consistent with PTQD [1] and Q-Diffusion [4] (*i.e.*, Residual Bottleneck Blocks, Attention Blocks, and the remaining layers). Specifically, for the reconstruction of the Residual Bottleneck Block, we freeze the quantization parameters of the `embedding layer`, and these parameters are only tuned in the reconstruction within the Temporal Information Block.

Additionally, the quantization settings are kept consistent with Q-Diffusion and PTQD.

B. Activation Range Variations in Finite Set

We analyze activation value ranges across all time steps in sampling data-unrelated components, *e.g.*, `time embed` and `embedding layers` for diffusion models. In Fig. I, it is evident that activation ranges vary notably among different time steps within these components. This observation suggests that the activation ranges within the same layer undergo considerable changes with varying time steps. Fortunately, the activations in the Time Information Block belong to a finite set, providing us the opportunity to conduct an accurate calibration for each time step.

C. Inappropriate Calibration Target

In this part, we further conduct experiments to provide the clues that the inappropriate reconstruction target also results in an inappropriate calibration. In the previous works, they calibrate the `embedding layers` along with the corresponding Residual Bottleneck Blocks. On the contrary, we freeze the quantized parameters of the `embedding layers` during the calibration process with a simple Min-max [7] initialization, which separates the calibration of `embedding layers` as alone. The experimental results in Tab. I demonstrate that without calibrating these layers inside the Residual Bottleneck Block can achieve better results. This confirms that the inappropriate calibration target leads to the suboptimal tuning of the quantization parameters.

D. Unconditional Image Generation on CIFAR-10

In this section, we conduct more experiments for unconditional image generation on CIFAR-10 32×32. As shown in Fig. II, our methods still achieve comprehensive improvements in terms of IS and FID compared to the existing SOTA. However, due to the lower resolution and simplicity of the images in this dataset, existing methods show minimal performance degradation, so the results we obtain may not be as pronounced.

Table I. FID and sFID on LSUN-Bedrooms 256×256 [11] for LDM-4. Prev represents BRECQ, the same as Tab. 1. Freeze denotes our trial here.

Methods	**Bits (W/A)**	FID↓	sFID↓
Full Prec.	32/32	2.98	7.09
Prev	8/8	7.51	12.54
Freeze	8/8	**6.87** (-0.64)	**10.12** (-2.42)
Prev	4/8	9.36	22.73
Freeze	4/8	**8.06** (-1.30)	**18.47** (-4.26)

Table II. Quantization results for unconditional image generation with DDIM on CIFAR-10 32×32.

Methods	**Bits (W/A)**	**CIFAR-10** 32×32	
		IS↑	FID↓
Full Prec.	32/32	9.04	4.23
PTQ4DM* [8]	4/32	9.02	5.65
Q-Diffusion† [4]	4/32	8.78	5.08
TDQ [9]	4/32	-	-
TFMQ-DM (Ours)	4/32	**9.14** (+0.12)	**4.73** (-0.35)
PTQ4DM [8]	8/8	9.02	19.59
Q-Diffusion† [4]	8/8	8.89	4.78
TDQ [9]	8/8	8.85	5.99
TFMQ-DM (Ours)	8/8	**9.07** (+0.05)	**4.24** (-0.54)
PTQ4DM* [8]	4/8	8.93	5.14
Q-Diffusion† [4]	4/8	9.12	4.98
TDQ [9]	4/8	-	-
TFMQ-DM (Ours)	4/8	**9.13** (+0.01)	**4.78** (-0.20)

E. Additional Effect of TIAR

As shown in Fig. 5, both of our proposed methods for LDM-4 on LSUN-Bedrooms 256×256 significantly reduce temporal feature errors, thereby alleviating temporal feature disturbance to a great extent. In this section, we conduct a detailed analysis of the cosine similarity between the outputs of the i^{th} Residual Bottleneck Blocks before and after quantization. We compare the results obtained with our TIAR and PTQD under w4a8 quantization, where $i = 8$ and $T = 200$ (the same as the settings in Fig. 5). From Fig. II, it can be observed that our approach significantly reduces output errors of the Residual Bottleneck Block compared to PTQD. However, it is essential to note that the error at this point involves the accumulated errors from multiple denoising iterations in diffusion models. Since Fig. 5 is not subject to the impact of accumulated errors, the trends of the lines in the two graphs may exhibit slight differences.

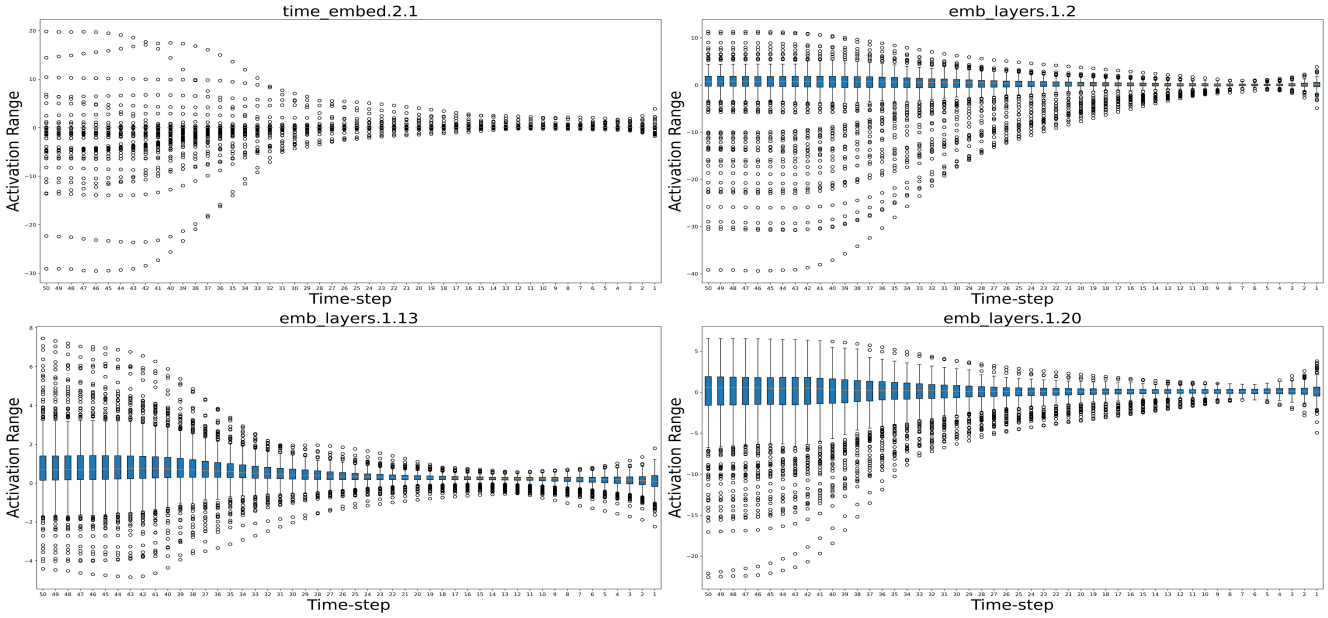


Figure I. Activation ranges within sampling data-unrelated components for LDM-4 on LSUN-Bedrooms 256×256 with 50 denoising steps. We randomly select 4 linear or convolutional layers' activations in these components to demonstrate the range variation.

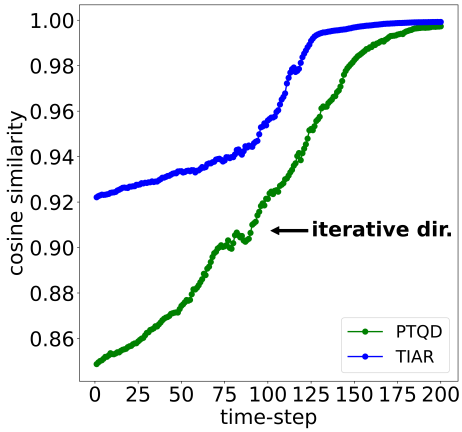


Figure II. Cosine similarity of the Residual Bottleneck's outputs across different PTQ Methods.

F. Inference Cost of TSC

In this section, we assess the inference overhead of our TFMQ-DM method on real hardware, specifically the Intel® Xeon® Gold 6248R Processor. All floating-point and quantized operations are implemented using Intel's OpenVINO toolkit [1]. As illustrated in Table III, in comparison to the UNet quantized with the built-in w8a8 quantization method in the OpenVINO toolkit, our approach results in a memory overhead of less than 0.076%, yielding a $2.38\times$ acceleration compared to the original floating-point model. Moreover, our method introduces less than 0.5% additional latency compared to the built-in w8a8 quantization in the OpenVINO toolkit.

[1] OpenVINO toolkit

Table III. Inference analysis of Stable Diffusion with 50 denoising time-steps on Intel CPU.

Methods	**Bits (W/A)**	UNet Size (Mb)	Latency (s)	Speedup
Full Prec.	32/32	3278.81	81.01	-
OpenVINO	8/8	821.15	33.93	2.39×
TFMQ-DM	8/8	821.77	34.07	2.38×

G. Study of Sampling with Advanced Samplers

Apart from employing the DDIM sampler [10], we also utilize a variant of DDPM [2] called PLMS [5] on the CelebA-HQ 256×256 dataset [3]. This better demonstrates the superiority of our TFMQ-DM framework compared to previous works. From Tab. IV, the introduced TFMQ-DM substantially reduces FID and sFID, surpassing PTQD by margins of 12.40 and 7.09, respectively.

Table IV. Quantization results for unconditional image generation with PLMS on CelebA-HQ 256×256.

Methods	**Bits (W/A)**	**CelebA-HQ** 256×256	
		FID↓	sFID↓
Full Prec.	32/32	8.92	10.42
Q-Diffusion [4]	4/8	24.31	22.11
PTQD [1]	4/8	21.08	17.38
TFMQ-DM (Ours)	4/8	**8.68** (-12.40)	**10.29** (-7.09)

Additionally, we present experiments performed using the DPM++ solver [6] on LSUN-Churches 256×256 [11]. As illustrated in Tab. V, our framework consistently outperforms existing methods in terms of performance on this

dataset with the DPM++ solver.

Table V. Quantization results for unconditional image generation with DPM++ on LSUN-Churches 256×256.

Methods	Bits (W/A)	**LSUN-Churches** 256×256	
		FID↓	sFID↓
Full Prec.	32/32	4.12	10.55
Q-Diffusion [4]	4/8	7.80	23.24
PTQD [1]	4/8	7.45	22.74
TFMQ-DM (Ours)	4/8	**5.51** (-1.94)	**13.15** (-9.59)

H. Comparison of Visualization Results

Within this section, we present random samples derived from full-precision and w4a8 quantized diffusion models with a fixed random seed. These quantized models were created through our TFMQ-DM or previous state-of-the-art methods. The figures below illustrate the obtained results. As shown from Fig. III to Fig. VIII, our framework yields results that closely resemble those of the full-precision model, showcasing higher fidelity. Moreover, it excels in finer details, producing superior outcomes in some intricate aspects (zoom in to closely examine the relevant images).

References

[1] Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. In *NeurIPS*, 2023. 1, 2, 3

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. 2

[4] Xiuyu Li, Long Lian, Yijiang Liu, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *ICCV*, 2023. 1, 2, 3

[5] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *ICLR*, 2022. 2

[6] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, 2023. 2

[7] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization, 2021. 1

[8] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *CVPR*, 2023. 1

[9] Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. In *NeurIPS*, 2023. 1

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[11] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop, 2016. 1, 2

(a) FP

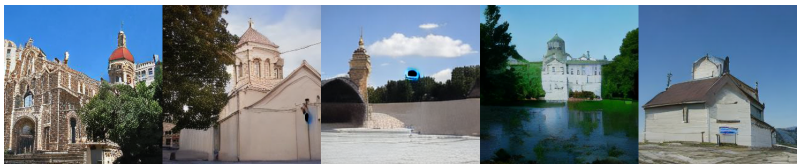
(b) Q-Diffusion (w4a8)

(c) PTQD (w4a8)

(d) TFMQ-DM (w4a8)

Figure III. Random samples from w4a8 quantized and full-precision LDM-4 on CelebA-HQ 256×256. The resolution of each sample is 256×256.

(a) FP



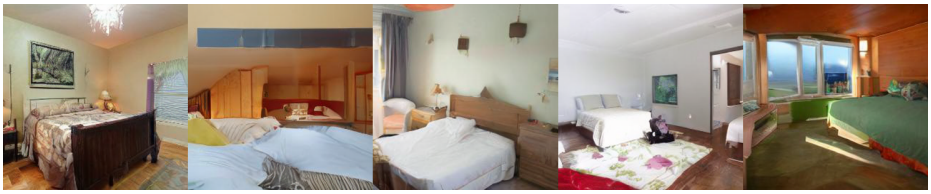
(b) Q-Diffusion (w4a8)

(c) PTQD (w4a8)

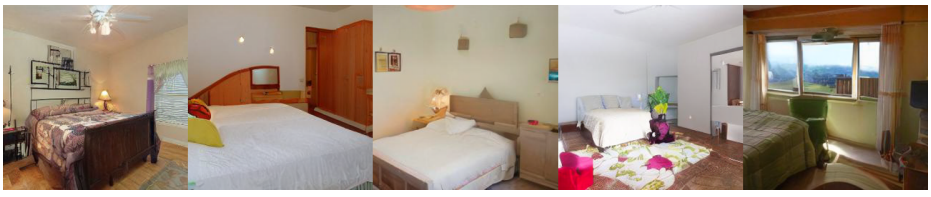
(d) TFMQ-DM (w4a8)

Figure IV. Random samples from w4a8 quantized and full-precision LDM-8 on LSUN-Churches 256×256. The resolution of each sample is 256×256.

(a) FP

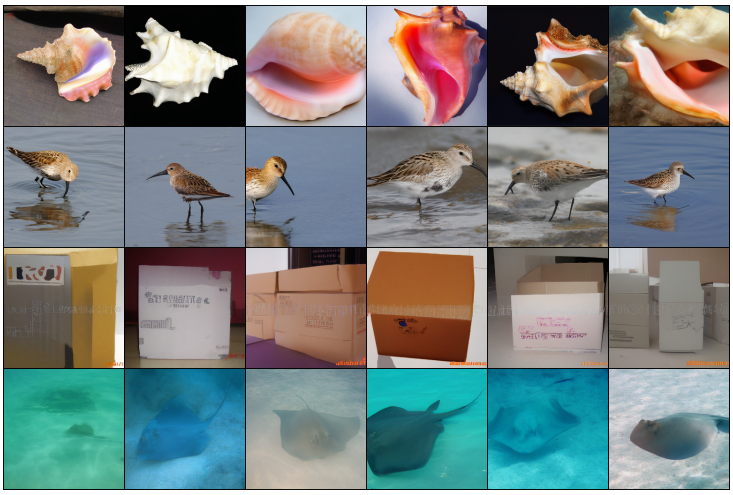


(b) PTQD (w4a8)

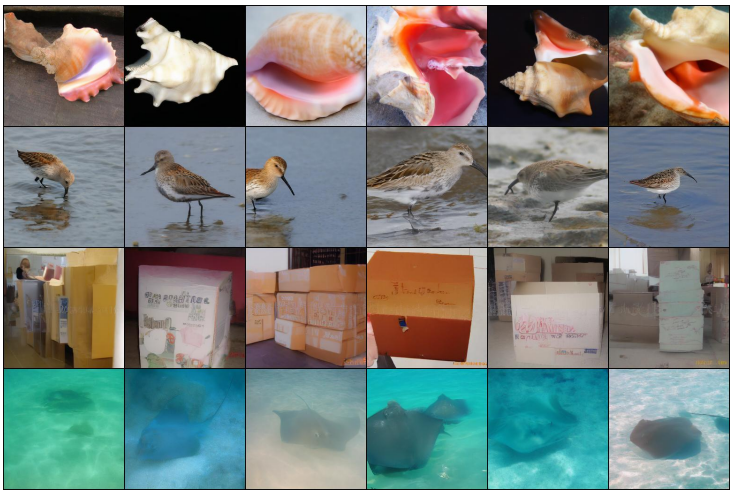


(c) TFMQ-DM (w4a8)

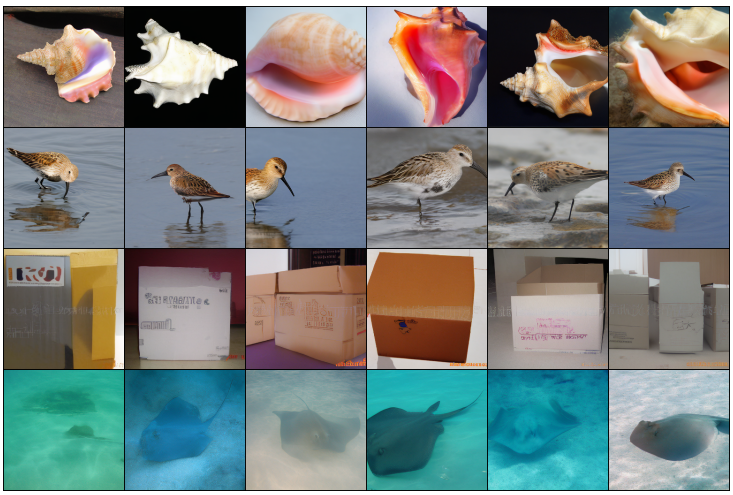
Figure V. Random samples from w4a8 quantized and full-precision LDM-4 on LSUN-Bedrooms 256×256. The resolution of each sample is 256×256.



(a) FP



(b) PTQD (w4a8)



(c) TFMQ-DM (w4a8)

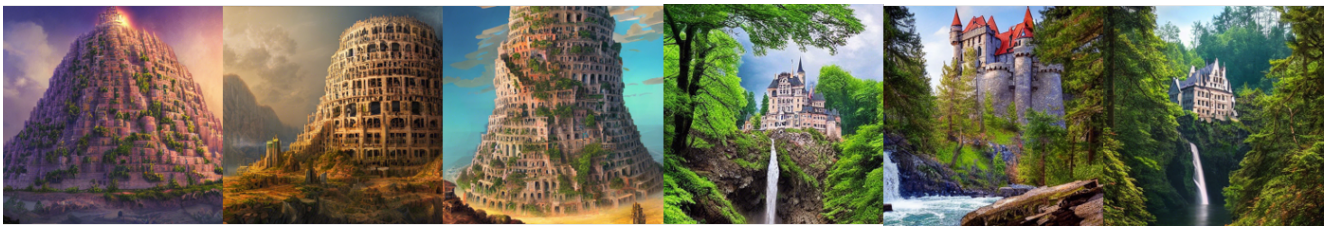
Figure VI. Random samples from w4a8 quantized and full-precision LDM-4 on ImageNet 256×256. The resolution of each sample is 256×256.

(a) FP

(b) Q-Diffusion (w4a8)

(c) TFMQ-DM (w4a8)

Figure VII. Random samples from w4a8 quantized and full-precision DDIM on CIFAR-10 32×32. The resolution of each sample is 32×32.



(a) FP

(b) Q-Diffusion (w4a8)

(c) TFMQ-DM (w4a8)

Figure VIII. Random samples from w4a8 quantized and full-precision Stable Diffusion. (Left) prompt: *A digital illustration of the Babel tower, detailed, trending in artstation, fantasy vivid colors*. (Right) prompt: *A beautiful castle beside a waterfall in the woods*. The resolution of each sample is 512×512.