This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Accurate Post-training Quantization for Diffusion Models

Changyuan Wang[1], Ziwei Wang[3], Xiuwei Xu[2], Yansong Tang[1,*], Jie Zhou[2], Jiwen Lu[2]
[1]Shenzhen Key Laboratory of Ubiquitous Data Enabling, Shenzhen International Graduate School, Tsinghua University, China [2]Department of Automation, Tsinghua University, China
[3]Carnegie Mellon University
{wangchan22@mails.,xxw21@mails.,tang.yansong@sz.,jzhou@,lujiwen@}tsinghua.edu.cn;
ziweiwa2@andrew.cmu.edu

Abstract

In this paper, we propose an accurate post-training quantization framework of diffusion models (APQ-DM) for efficient image generation. Conventional quantization frameworks learn shared quantization functions for tensor discretization regardless of the generation timesteps in diffusion models, while the activation distribution differs significantly across various timesteps. Meanwhile, the calibration images are acquired in random timesteps which fail to provide sufficient information for generalizable quantization function learning. Both issues cause sizable quantization errors with obvious image generation performance degradation. On the contrary, we design distribution-aware quantization functions for activation discretization in different timesteps and search the optimal timesteps for informative calibration image generation, so that our quantized diffusion model can reduce the discretization errors with negligible computational overhead. Specifically, we partition various timestep quantization functions into different groups according to the importance weights, which are optimized by differentiable search algorithms. We also extend structural risk minimization principle for informative calibration image generation to enhance the generalization ability in the deployment of quantized diffusion model. Extensive experimental results show that our method outperforms the state-of-the-art post-training quantization of diffusion model by a sizable margin with similar computational cost[1].

1. Introduction

Denoising diffusion generative models [11, 32] have achieved outstanding performance in a wide variety of computer vision tasks such as image edition [2, 28], style transformation [34, 38], image super-resolution [18, 30] and many others. Compared with the generative adversarial networks (GAN), diffusion models obtained recovered contents with better quality and diversity on most downstream tasks. How-

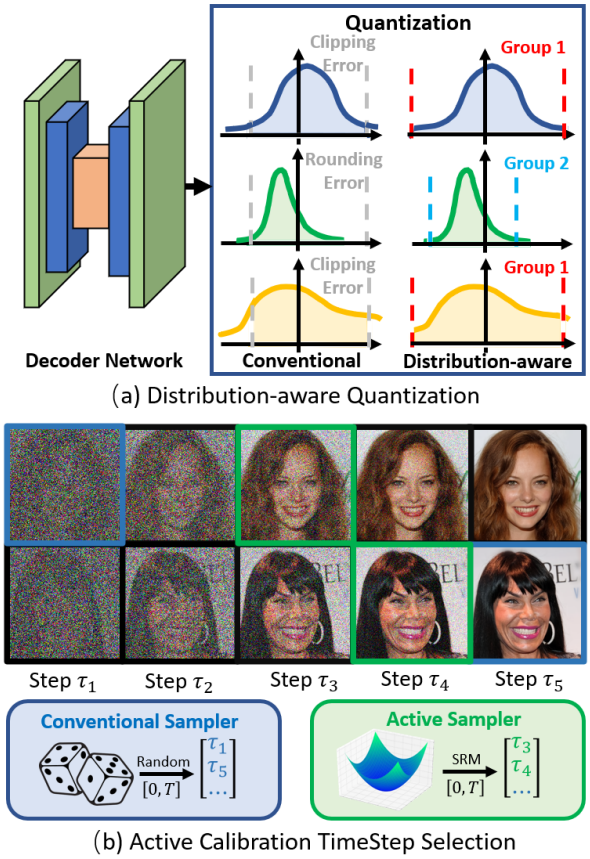


Figure 1. (a) Existing methods leverage shared quantization for activation discretization across different timesteps with significant quantization errors, while we divide timesteps into groups with specific rounding functions for each partition. (b) Conventional methods construct calibration set by randomly image selecting with ineffective supervision, while we actively sample timesteps based on the structural risk minimization (SRM) principle.

*Corresponding author.
[1]Code is available at https://github.com/ChangyuanWang17/APQ-DM

ever, diffusion models usually require hundreds of noise evaluation steps to generate high-quality images from Gaussian noises by neural networks with millions of parameters, and the numerous forward passes in network inference result in heavy computation burden. Therefore, designing lightweight denoising process for diffusion models is highly demanded for flexible deployment in practical applications with limited resources such as mobile phones and robots.

To accelerate the generation process of diffusion models, recent studies made significant efforts in decreasing the sampling times of image denoising process [3, 14, 33] and reducing the network complexity in noise estimation of each step [19, 20, 31]. The former removes redundant steps in the reverse process of diffusion, and the latter extends the network compression techniques to noise estimation such as pruning [12, 40] and quantization [13, 17] for acceleration. We focus on the latter by quantizing the noise estimation networks with integer arithmetic inference. Due to the intractability of the training data and the unbearable training cost of diffusion models to fully optimize quantized network parameters, the post-training quantization framework for the pre-trained full-precision decoders is leveraged that only learns the rounding function parameters. Nevertheless, conventional data-free post-training quantization methods [5, 42] learn a shared layer-wise rounding function for all generation timesteps where the activation distribution varies obviously in diffusion models, and the calibration images are generated in random timesteps which fails to provide sufficient information to acquire generalizable quantization function. Consequently, both the inaccurate quantization functions and uninformative calibration images lead to significant quantization errors in noise estimation process, which degrades the synthesis performance by a sizable margin.

In this paper, we present an accurate post-training quantization framework for diffusion models in order to achieve efficient image generation. Different from existing methods that leverage shared layer-wise quantization functions for all timesteps and synthesizing calibration images in random timesteps for training, we partition timesteps into different groups to impose specific rounding functions for each group and sampling the optimal timesteps to generate informative calibrate images for quantization parameter learning. Therefore, the significant quantization errors of noise estimation in diffusion model deployment can be reduced with only negligible computation overhead. More specifically, we employ a differentiable search strategy to acquire the optimal group assignment for different generation timesteps, and learns individual rounding functions for each group with minimized discretization errors. For the differentiable search, the activations quantized by discretization functions in different groups are summed with learnable importance weights. We also generalize the structural risk minimization (SRM) principle for timestep selection to generate informative calibration images, where the entropy of rounding function weights in differentiable search and the sampling times of the timestep are considered as the criteria based on our formulation. Figure 1 demonstrates the comparison between our method and conventional data-free post-training quantization framework for diffusion models. Extensive experimental results on unconditioned synthesis and conditional image generation across various network architectures clearly demonstrate that our method sizably increases the quality of the generated images with only negligible computational complexity. Our contributions can be summarized as follows:

- We propose an accurate and efficient post-training quantization framework for pre-trained diffusion models that preserve the generation performance in image generation with 6-bit weights and activations.
- We present the distribution-aware quantization and activate timestep selection function to significantly reduce quantization errors across generation timesteps and search the representative calibration images according to structural risk minimization principle, so that the rounding functions can be optimized with more informative supervision.
- We conduct extensive experiments on a wide variety of datasets for image generation, and the results clearly demonstrate the superiority of the presented method.

2. Related Work

Efficient diffusion models: Diffusion models achieve more satisfying quality and diversity in image generation compared with GANs, while the generation efficiency is significantly decreased due to the iterative noise evaluation process with long timesteps. The denoising diffusion probabilistic model (DDPM) [11] leverages a forward pass for noise perturbation and a reverse process for image denoising. Existing methods mainly focus on leveraging a shorter sampling path without sizable performance degradation, which can be divided into two categories including convergence speedup and sampling path selection. Convergence speedup methods aim to discretize the stochastic differential equations (SDE) or the ordinary differential equations (ODE) with minimized discretization errors. Song *et al.* [32] modeled the diffusion model with a non-Markov process that considers the original images for noise perturbation, where the convergence for image generation speeds up sizably. Bao *et al.* [3] formulated an analytic form of variance and KL divergence based on a pre-trained score-based model that simultaneously enhanced the log-likelihood and the generation speed. Sampling path selection usually chooses partial timesteps in the denoising process regarding the learning objectives. Watson *et al.* [35] searched the best K sampling timesteps for noise evaluation via dynamic programming, where the goal was to maximize the evidence lower bound (ELBO) in the reverse process. Due to the inconsistent performance between the training ELBO and the generation quality, they further presented

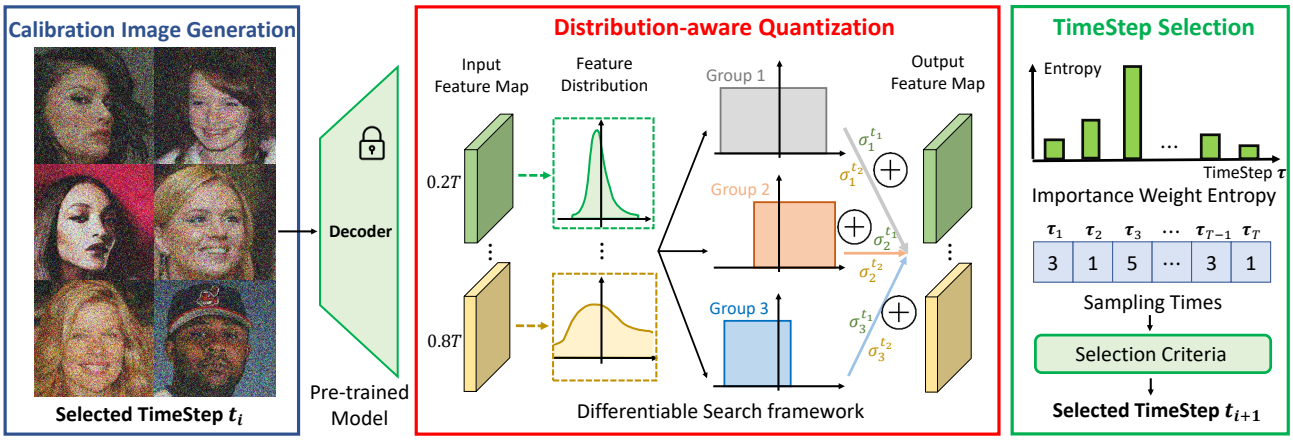


Figure 2. The overall pipeline of our method. The calibration images are generated according to the selected timesteps, and activations in the pre-trained diffusion models are parallelly quantized by rounding functions of all groups. The output feature maps are acquired by adding the quantized value with the importance weights, where the quantization parameters and the importance weights are jointly optimized. The importance weight entropy and the sampling times are considered in the timestep selection criteria to decide the optimal timestep for calibration image generation in the next round.

Kernel Inception Distance (KID) [36] as the optimization objective to differentiably search the sampling timesteps. In this paper, we aim to reduce the complexity of single-step denoising process by quantization, which is orthogonal to the acceleration techniques of sampling path shortening.

Network quantization: Network quantization has aroused extensive interest in computer vision due to the significant reduction in storage and computational cost, as the full-precision variables are substituted by quantized values and the multiply-add (MAC) operations are replaced by integer arithmetics. Quantization-aware training (QAT) [6, 24] finetunes the quantized network with training dataset of full-precision models. Due to the inaccessibility of the full training set and the extremely high training cost, post-training quantization (PTQ) [10, 21, 25, 26] that optimizes the rounding functions with a small calibration set is more practical in realistic applications. Choukroun *et al.* [7] minimized the l_2 distance between the quantized and full-precision tensors to avoid obvious task performance degradation, and Zhao *et al.* [41] duplicated the channels with outliers and halved the value so that the clipping loss could be reduced without increasing the rounding errors. Liu *et al.* [25] preserved the relative ranking orders of the self-attention in vision transformers to prevent information loss in post-training quantization, and explored mixed-precision quantization strategy according to the nuclear norm of attention map and features. Zero-shot PTQ further extends the limits that efficiently quantize neural networks without any real image data. Cai *et al.* [5] optimized the pixel values of the generated images to enforce the statistics of sample batches to mimic the batch normalization (BN) layers in the full-precision networks. Li *et al.* [22] further extended PTQ framework to transformer architectures by diversifying the self-attention of different patches with patch similarity metrics. As Shang *et al.* [31] and Li *et al.* [20] observed, different activation distribution across timesteps and the effectiveness change of calibration images acquired in various timesteps amplify the quantization errors in existing methods. To avoid the overfitting of the step-wise quantization caused by limited calibration samples, we present the distribution-aware quantization for diffusion models across timesteps with significantly reduced learnable parameters. Meanwhile, different from [31] that manually assigned the timestep index for calibration generation, we generalize the structural risk minimization principle to discover the optimal timesteps.

3. Approach

In this section, we first introduce the preliminaries of post-training quantization for diffusion models and then detail the distribution-aware quantization across generation timesteps with the differentiable search framework. Finally, we demonstrate the timestep selection for calibration image generation according to the structural risk minimization principle.

3.1. Network Quantization

Diffusion models leverage a forward pass to impose noise on images and a reverse pass to transform Gaussian noise into an image for generation. Denoting the real data as $\boldsymbol{x}_0$ and the latent image at the t_{th} step as $\boldsymbol{x}_t$, the probability of the forward process can be represented as follows:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}), \tag{1}$$

where β_t means the variance schedule at the t_{th} step that indicates imposed Gaussian noise to the latent image. When the total number of forward steps denoted as T becomes large enough, the latent image $\boldsymbol{x}_T$ can be regarded as the

standard Gaussian noise. We leverage an approximated condition distribution $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ to generate the latent image in reverse process due to the intractability of true distribution $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, where the approximated distribution is parameterized by neural networks with weight θ:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}|\boldsymbol{\mu}_{\theta,t}(\boldsymbol{x}_t), \boldsymbol{\Sigma}_{\theta,t}(\boldsymbol{x}_t)). \quad (2)$$

The training process of diffusion model aims to minimize the negative log-likelihood with the evidence lower bound optimization in variational inference:

$$L_{VLB} = \mathbb{E}_{q(\boldsymbol{x}_{0:T})}[\log \frac{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}{p_\theta(\boldsymbol{x}_{0:T})}] \geqslant -\mathbb{E}_{q(\boldsymbol{x}_0)} \log \boldsymbol{p}_\theta(\boldsymbol{x}_0). \quad (3)$$

In the practical applications, iterative noise estimation process is implemented with the diffusion model for content generation, and the heavy computational cost of the reverse phase disables the deployment in resource-constrained devices such as mobile phones and robots. To accelerate the denoising process for each reverse step, post-training quantization leverages a small calibration set to learn the rounding function parameters for weights and activations of the decoder, where the quantization function can be represented as follows:

$$\hat{x} = s \cdot \Phi([x/s], z_{min}, z_{max}), \quad (4)$$

where x and $\hat{x}$ represent real-valued and quantized matrix respectively. $[\cdot]$ means the rounding function to the nearest integer and Φ is the clipping operation that regularizes the element into the range from z_{min} to z_{max}. The quantization scaling parameter s indicates the interval between adjacent rounding points. As empirically demonstrated in [31], the activation distribution varies significantly across different timesteps during the reverse process, and the shared rounding functions usually cause severe quantization errors for image generation. Moreover, randomly selecting the timestep to generate latent images for calibration set construction fails to provide sufficient information for generalizable quantization function learning.

3.2. Distribution-aware Quantization for Diverse Activation Distribution

Since the activation distribution changes significantly across timesteps, discretizing the full-precision intermediate features in similar data distribution with the same quantization functions can reduce the quantization errors. We first describe the distribution-aware quantization scheme and then illustrate the differentiable group assignment of timesteps.

Shared quantization functions may cause large clipping errors for widely distributed activations and rounding errors for narrowly distributed ones. Directly assigning specific rounding functions for network activations in each timestep leads to overfitting in optimization because of the limited calibration samples, and quantizing activations in the timestep where optimal quantization range is similar with the same rounding functions can achieve better trade-off between the quantization accuracy and rounding function generalizability. Assuming partitioning all T timesteps into C groups, the quantization strategy can be written in the following:

$$\hat{x} = s_{g(i)} \cdot \Phi(\left[x/s_{g(i)}\right], z_{min}^{g(i)}, z_{max}^{g(i)}), \quad (5)$$

where $g(i)$ represents the assigned group for activations in the i_{th} timestep. Meanwhile, $s_{g(i)}$, $z_{min}^{g(i)}$ and $z_{max}^{g(i)}$ respectively stand for scale parameters, the lower bound and the upper bound of quantization for activations in the i_{th} timestep. Assigning the optimal group indexes for different timesteps is critical in distribution-aware quantization to reduce the quantization errors without obvious computation overhead. Since enumerating assignment permutation is NP-hard to find the optimal solution, we extend the differentiable search framework to efficiently partition timesteps with minimal quantization errors. In the differentiable search, the latent images are quantized by all quantization functions, whose output values are summed with learnable importance weights to form the input for the next layer in the diffusion model:

$$\hat{x} = \sum_{g=1}^{G} \sigma_g s_g \cdot \Phi([x/s_g], z_{min}^g, z_{max}^g), \quad (6)$$

where σ_g means the importance weight of the quantization function for the c_{th} group with the normalization $\sum_{g=1}^{G} \sigma_g = 1$. When the training process completes, the rounding function with the largest importance weight is selected to be the search results for group-wise quantization. Despite the noise estimation loss of diffusion models, we also enforce the importance weights to approach zero or one by minimizing the entropy to avoid discretization errors in rounding function selection. The overall optimization objective J is written as follows, where we denote $\boldsymbol{\epsilon}_\theta(\sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, t)$ as $\boldsymbol{\epsilon}_\theta$ for simplicity:

$$\min_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}} J = J_d + \lambda J_e = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta||_2^2 + \lambda \sum_{g=1}^{G} -\sigma_g^t \log \sigma_g^t, \quad (7)$$

where J_d and J_e respectively represent the simplified variational lower bound of diffusion model objective and the discretization minimization loss in differentiable search, and the hyperparameter λ balances the importance of different terms. σ_g^t demonstrates the importance weights of the quantization function for the g_{th} group in the t_{th} timestep. The noise $\boldsymbol{\epsilon}$ from standard Gaussian distribution is approximated by the predicted noise $\boldsymbol{\epsilon}_\theta$ in the optimization objective. The diffusion parameter $\overline{\alpha}_t = \prod_{i-1}^{t} 1 - \beta_i$ controls the strength of noise in diffusion. We jointly update the parameters in quantization functions and the importance weights until convergence or achieving the maximal iteration steps, and the discretized hypernetwork is directly employed to generate images efficiently.

3.3. TimeStep Selection for Calibration Generation

The pipeline of post-training quantization for iterative reverse process in diffusion models differs significantly from that in conventional vision models. Leveraging latent images in all timesteps leads to unbearable training cost for quantization function learning, and latent images in adjacent timesteps can only offer redundant information for parameter optimization. On the contrary, randomly select part of the timesteps usually fails to provide sufficient supervision that is representative to demonstrate the real distribution of the latent images. Therefore, it is desirable to actively sample the timesteps to generate latent images for quantization parameter learning with effective guidance. We generate representative samples by structural risk minimization, which minimize the distance between selected and real distribution. We first introduce the extension of SRM principle to active timestep selection, and then formulate the selection criteria that can be feasibly computed.

Structural risk minimization principle minimizes the upper bound of the true risk on unseen data distribution, where the bound can be written as follows for a dataset containing n samples with the probability at least $1-\delta$[4]:

$$E(J(x)) \leqslant \overline{E(J(x))} + 2R_n(\mathcal{F}) + \sqrt{\frac{\ln 1/\delta}{n}}, \tag{8}$$

where $E(J(x))$ and $\overline{E(J(x))}$ respectively illustrate the true expectation of the risk J for real data distribution x and the empirical expectation of that for sampled data from x, and $R_n(\mathcal{F})$ is the Rademacher complexity over the function class $\mathcal{F}$. The SRM principle requires the data to be sampled from i.i.d. distribution, while the latent images in selected timesteps should be more informative and representative. Therefore, we rewrite the SRM principle in the following way, where the detailed formulation is in the supplementary:

$$E(J) \leqslant \overline{E_S(J)} + MMD(p(X), p(X_s)) + \mathcal{R}_0, \tag{9}$$

where we omit the data distribution x for simplicity. $\overline{E_S(J)}$ denotes the empirical risk of the latent images of selected timesteps for noise estimation, and $\mathcal{R}_0$ demonstrates the complexity of the diffusion model in the reverse process. X and X_s stand for the distribution of latent images generated in all timesteps and the selected ones. The maximal mean discrepancy between two distributions $p(X)$ and $p(X_s)$ is represented as $MMD(p(X), p(X_s))$, which demonstrates the generalization ability of the calibration sets for quantization learning. The first criteria acquired by worst-case empirical risk of sampled latent images is formulated as follows for timesteps selection :

$$\min_{t,\boldsymbol{x}_t} \overline{E_S(J)} = \sum_{\boldsymbol{x}_t \in \mathcal{S}} J_d + \lambda J_e, \tag{10}$$

where $\mathcal{S}$ represents the images selected in the calibration set, and J is the optimization objective defined in (7). This formula aims to train the highly quantified diffusion model with the constructed calibration sets. Since the original latent $\boldsymbol{x}_T$ can be regarded as the standard Gaussian noise without bias, the objective J_d does not affect the worst-case empirical risk with different timesteps. The variance of J_e influences the worst-case empirical risk across timesteps, because the entropy of the importance weights of distribution-aware quantization functions changes with the timesteps. Therefore, the criteria s_1 from the empirical risk minimization can be transformed to selecting the timestep with the highest entropy of importance weights as $s_1 = \sum_{g=1}^{G} -\sigma_g^t \log \sigma_g^t$.

Meanwhile, The definition of maximal mean discrepancy can be written as follows, where we denote $MMD(p(X), p(X_s))$ as M for simplicity:

$$\begin{aligned} \min_t M &= \sup || \frac{1}{|U|} \sum_{\boldsymbol{x}_t \in U} \boldsymbol{\epsilon}_\theta - \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}_t \in \mathcal{S}} \boldsymbol{\epsilon}_\theta || \\ &= \frac{\varphi}{N_t + 1} \propto \frac{1}{N_t + 1}, \end{aligned} \tag{11}$$

where U means the full set containing all original latent and timesteps for calibrating image selection, and $|\cdot|$ represents the number of elements in the set. φ is a constant in timestep sampling and N_t denotes the number of sampling times for the t_{th} timestep in calibration set construction. Because the estimated noise of the samples in the full set is intractable, we utilize the number of sampling times to optimize the maximal mean discrepancy based on upper confidence bound (UCB) principle [1] which achieves exploitation-exploration trade-off when sampling. A detailed derivation of the formula (11) can be found in the supplementary material. For the timestep when we sample a large number of latent images for calibration set construction, the maximal mean discrepancy becomes low as we acquire sufficient information of the latent image distribution in this timestep. Therefore, we explore latent images in the timestep with few sampling times to further minimize the maximal mean discrepancy with high marginal benefits. The criteria s_2 from the maximal mean discrepancy is designed as $s_2 = 1/(N_t + 1)$. The overall timestep selection criteria can be written as follows:

$$\max_t s = s_1 + \eta s_2 = \sum_{g=1}^{G} -\sigma_g^t \log \sigma_g^t + \frac{\eta}{N_t + 1}, \tag{12}$$

where η is a hyperparameter to balance the importance of empirical risk and maximal mean discrepancy. The overall pipeline of our method is depicted in Figure 2. With the optimally selected timesteps, the generated calibration images can provide effective supervision for quantization function learning, which can be well generalized in deployment.

4. Experiments

In this section, we first introduce the implementation details of our method. We then conduct ablation studies to evaluate

Bitwidth	Group	C-Error	G-Error	IS↑	FID↓
W8A8	1	1.16	1.22	8.93	5.32
	4	0.87	0.93	8.97	4.73
	8	0.79	0.82	9.07	4.24
	16	0.75	0.86	8.98	4.29
W6A6	1	1.92	2.03	8.82	9.73
	4	1.88	1.76	8.92	7.10
	8	1.57	1.68	9.06	6.57
	16	1.52	1.74	9.24	6.77

Table 1. Effects of the number of timestep partitions in the distribution-aware quantization. C-Error and G-Error depict the quantization errors of activations in calibration and generation respectively.

Method	Bitwidth	Size of Calibration Set			
		128	256	512	1024
Random	W8A8	5.72	5.64	5.34	5.41
	W6A6	11.65	10.12	9.18	8.92
Heuristic	W8A8	5.75	5.56	5.27	5.21
	W6A6	12.26	10.19	9.03	8.83
Active	W8A8	5.99	4.46	4.49	4.24
	W6A6	12.61	11.73	7.83	6.57

Table 2. Different timestep sampling strategies for calibration set construction across various sizes of calibration images. WBAB depicts the weights and activations are quantized to B-bit.

the effectiveness of the distribution-aware quantization and the optimal timestep selection for calibration image generation. Meanwhile, we visualize the importance evolution during the differentiable search and analyze the influence of hyperparameters on generation quality. Finally, we compare our method with the state-of-the-art post-training quantization frameworks in diffusion models to show our superiority.

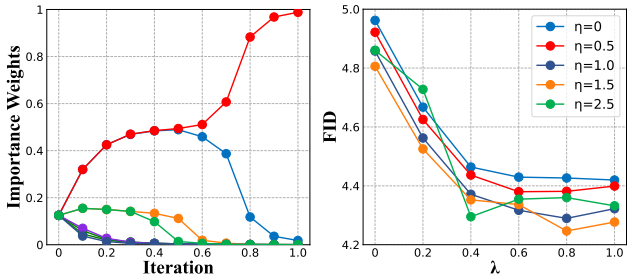
4.1. Implementation Details

We utilize the diffusion frameworks for post-training quantization including DDIM [32] and LDMs [29] with their pre-trained weights, which require 100 iterative denoising timesteps for image generation in most experiments. We set the bitwidth of quantized weight and activation to 6 and 8 to evaluate our method in different quality-efficiency trade-offs and utilized the uniform quantization scheme where the interval between adjacent rounding points was equal. For distribution-aware quantization across different timesteps, we partitioned all timesteps into eight groups in most experiments. We followed the initialization of the quantization function parameters in [20] for the baseline methods and our APQ-DM, where we minimized the l_p distance [27, 37] between the full-precision and quantized activations to optimize the value range for clipping. The hyperparameters λ in the objective of differentiable search and η in the timestep selection criteria were set to 0.8 and 1.5 respectively.

For the parameter learning in differentiable search, we generated 1024 images for hyper-network learning where the batchsize was assigned with 64 for calibration set construction. The learning rate was initialized to $3e^{-3}$ and $5e^{-3}$ for 6 and 8-bit diffusion models and ended up with $1e^{-5}$ for all bitwidth settings with 0.05 decaying strategy. The quantization function parameters and the importance weights were jointly updated for 10 epochs in the differentiable search, and the acquired distribution-aware quantization function was directly employed for image generation.

4.2. Ablation Study

In order to investigate the influence of the distribution-aware quantization for network activations across different



(a) (b)

Figure 3. (a) The evolution of branch importance weights during the differentiable search. (b) The generation quality w.r.t. different hyperparameters λ and η.

timesteps, we vary the number of groups with different trade-offs between quantization accuracy and rounding function generalizability. To show the effectiveness of our active timestep selection for calibration set generation, we compare our strategy with various sampling techniques. Meanwhile, we modified the hyperparameter λ and η to demonstrate the effect of the discretization loss in rounding function selection and the maximal mean discrepancy in timestep selection criteria. All experiments in the ablation study were conducted with the 32×32 cifar-10 dataset and the DDIM diffusion framework.

Performance w.r.t. the number of timestep groups: Dividing the timesteps into more groups can significantly reduce the clipping and rounding errors for differently distributed activations in quantization function learning, while may resulting in the rounding function overfitting due to the limited calibration samples and large-scale learnable parameters. Table 1 illustrates the quantization errors, Inception Score (IS) and FID score for our method that partitions the timesteps into different numbers of groups. Observing the FID and IS for different group partition settings, we conclude that dividing the timesteps into several groups outperforms both the shared quantization policy and the step-wise rounding functions. Therefore, we assign the number of groups for the timestep partition to 8 in the rest of experiments to achieve the optimal trade-off between the quantization accuracy and rounding function generalizability.

Method	Bitwidth	Cifar-10			CelebA			LSUN-Bedroom			LSUN-Church		
		IS↑	FID↓	sFID↓	IS↑	FID↓	sFID↓	IS↑	FID↓	sFID↓	IS↑	FID↓	sFID↓
Baseline	FP	9.07	4.23	4.37	2.61	6.49	13.82	2.45	6.39	9.45	2.76	10.98	16.16
LSQ	W8A8	8.74	13.78	6.93	2.29	15.02	15.99	2.13	16.95	18.85	2.58	28.49	30.95
PTQ4DM		8.82	5.69	4.42	2.43	6.44	14.15	2.23	7.48	12.42	2.76	10.98	17.28
Q-Diffusion		8.87	4.78	4.49	2.41	6.60	14.19	2.27	7.04	12.24	2.72	12.72	16.96
APQ-DM		**9.07**	**4.24**	**4.37**	**2.58**	**6.07**	**13.09**	**2.55**	**6.46**	**11.82**	**2.84**	**9.04**	**16.74**
LSQ	W6A6	8.34	35.96	37.04	1.94	78.37	85.04	1.68	122.45	126.24	1.87	131.78	140.39
PTQ4DM		8.72	11.28	6.92	2.13	24.96	20.95	2.11	16.85	19.65	2.48	32.85	36.74
Q-Diffusion		8.76	9.19	5.80	2.16	23.37	19.83	2.09	17.57	18.74	2.52	33.77	35.27
APQ-DM		**9.06**	**6.57**	**4.71**	**2.30**	**16.86**	**17.85**	**2.30**	**15.72**	**17.24**	**2.63**	**24.75**	**29.24**

Table 3. Comparisons with the state-of-the-arts data-free post-training quantization methods on unconditional image generation for DDIM diffusion models across various datasets and bitwidth setting.

Performance w.r.t. different timestep sampling strategies for calibration set construction: We compare our active timestep sampling strategy for calibration set generation with random and heuristic sampling policies [31]. Random sampling assigns an integer number drawn from uniform distribution from zero to the maximal time steps T, and heuristic sampling determines the timestep from a Gaussian distribution $\mathcal{N}(\mu, \frac{T}{2})$ where μ is less than $\frac{T}{2}$. Table 2 shows the generation quality for different timestep sampling methods across various sizes of the calibration set. Our active sampling strategy outperforms the random and heuristic sampling policies by a large margin, and the advantage is more obvious for calibration sets with small sizes because informative samples are extremely important for post-training quantization in the scenario without sufficient images.

Visualization of importance weight in differentiable search: Figure 3a depicts the evolution of importance weights during the differentiable search for group assignment, where different colors represent disparate groups. At the early stage of the differentiable search, the differences between importance weights are not obvious because of insufficient calibration images. When the network gradually converges, the principal impacts on the performance result from the quantization functions with different rounding and clipping errors and differentiate group assignments between different data distributions.

Performance w.r.t. hyperparameters λ and η: The hyperparameter λ controls the importance of the discretization loss in distribution-aware quantization function in the objective of differentiable search, and η balances the empirical risk and the maximal mean discrepancy in the timestep selection. Figure 3b depicts the FID for different hyperparameter settings, where the medium value for both parameters achieves the highest generation quality. The model performance is more sensitive to the hyperparameter λ because the importance weights of quantization functions in different groups usually have similar distribution as one-hot vector, and slight change to λ leads to large perturbation to the overall objective in differentiable search due to the logarithm.

4.3. Comparison with the State-of-the-art Methods

In this section, we compare our proposed method with the state-of-the-art data-free post-training quantization frameworks including LSQ [9] and those specifically designed for diffusion models including PTQ4DM [31] and Q-diffusion [20]. The IS, FID, and sFID scores of the baseline methods are acquired by implementing the officially released code or our re-implementation. For fair comparison of all listed methods, we leverage the rounding function in LSQ for quantization and de-quantization, and generate latent images with 100 iterative timesteps.

Results on unconditional generation: Unconditional generation samples a random variable for diffusion models to yield images with similar distribution of the training datasets. We evaluate our data-free post-training quantization methods on 32×32 Cifar-10 [15], 64×64 CelebA [23], 256×256 LSUN-Church Outdoor and LSUN-Bedroom datasets [39] for DDIM frameworks, and evaluate for LDMs diffusion frameworks on 256×256 CelebA-HQ [16], LSUN-Church Outdoor and LSUN-Bedroom datasets, where the generalization quality and efficiency are reported in Table 3 and Table 4 respectively. LSQ learns the optimal quantization step sizes with minimized discretization errors, while the shared quantization policy across timesteps and randomly constructed calibration set in diffusion models leads to significant quantization loss. PTQ4DM and Q-diffusion employ the step-wise quantization functions to minimize the quantization errors of the diversely distributed activations across timesteps, and presents heuristic timesteps selection criteria for calibration image generation. However, the optimization of large-scale learnable parameters faces the challenges of overfitting due to the very limited quantity of calibration samples, and the data-independent calibration set construction cannot guarantee the optimality of the calibration images. As a result, our method outperforms PTQ4DM by 0.32 (2.55 vs. 2.23) and 1.02 (6.46 vs. 7.48) for IS and FID in LSUN-Bedroom respectively. The computational cost remains the same for baseline methods and our APQ-DM due to the

Method	Bitwidth	CelebA-HQ (U)			Bedroom (U)			Church (U)			ImageNet (C)		
		IS↑	FID↓	sFID↓	IS↑	FID↓	sFID↓	IS↑	FID↓	sFID↓	IS↑	FID↓	sFID↓
Baseline	FP	3.27	6.08	9.36	2.29	3.43	7.68	2.70	4.08	10.99	180.84	11.89	6.86
LSQ	W8A8	3.01	9.75	11.04	2.13	8.11	11.40	2.50	7.10	11.21	154.06	13.26	22.87
PTQ4DM		3.11	8.57	10.36	2.21	4.75	10.76	2.52	5.29	12.49	161.75	12.59	13.53
Q-Diffusion		3.08	8.61	10.43	2.19	4.67	10.51	2.53	4.87	12.95	166.05	12.78	12.21
APQ-DM		**3.22**	**6.30**	**9.25**	**2.35**	**3.88**	**8.55**	**2.69**	**4.02**	**10.70**	**179.13**	**11.58**	**6.31**
LSQ	W6A6	2.09	129.84	135.85	1.34	122.45	148.19	1.82	135.61	77.77	115.71	40.77	48.73
PTQ4DM		2.80	19.53	21.00	2.08	11.10	14.83	2.46	11.05	20.92	140.86	13.68	23.40
Q-Diffusion		2.87	18.39	20.56	2.11	10.10	14.50	2.47	10.90	21.54	146.41	13.94	22.73
APQ-DM		**3.09**	**16.73**	**18.75**	**2.27**	**9.88**	**13.29**	**2.67**	**6.90**	**13.53**	**178.64**	**11.58**	**7.40**

Table 4. The generation quality on unconditional (U) and class-conditional (C) image synthesis for LDMs diffusion models across different datasets and bitwidths.

(a) Full Precision (b) PTQ4DM(6-bit) (c) APQ-DM(6-bit)

Figure 4. The images generated by quantized Stable Diffusion models and the corresponding text prompts, where different post-training quantization methods are employed.

stored rounding parameters. The advantage of our method becomes more obvious for 6-bit diffusion models because quantization errors and calibration sample informativeness are more important for networks with low capacity.

Results on conditioned image generation: Conditioned image generation synthesizes samples according to text including class names or descriptions. For class-conditional image generation, we discretize the LDMs model that is pre-trained on the 256×256 class-conditional ImageNet [8] dataset, where the guidance strength is set to 3.0 to balance the generation quality and diversity. Table 4 shows the quantitative experimental results for different post-training quantization methods, while our method increases the FID and IS by 1.01 (11.58 vs.12.59) and 17.38 (179.13 vs. 161.75) respectively compared with the state-of-the-art method PTQ4DM. For description-conditional image generation, Figure 4 demonstrates some examples of the text prompts images that are generated by different quantized Stable Diffusion models, where our method can still acquire plausible images with high-quality details with weights and activations in low bitwidths. Since conditioned image generation is widely adopted in many realistic multimedia applications, our method brings potential to deploy large pre-trained diffusion models on mobile devices or robots under limited resource constraints with satisfying generation quality.

5. Conclusion

In this paper, we have presented a novel post-training quantization framework of diffusion model for efficient image generation. We design a differentiable search framework that assigns the optimal partition for each timestep, where network activations are discretized with distribution-aware quantization functions for rounding error minimization. By generalizing the structural risk minimization principle, we select the optimal timesteps for calibration image construction to provide effective supervision in quantization parameter learning. Extensive experiments demonstrate that our methods outperform the state-of-the-art post-training quantization methods across various diffusion architectures.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62125603, Grant 62321005, and Grant 62336004, and Shenzhen Key Laboratory of Ubiquitous Data Enabling (Grant No. ZDSYS20220527171406015).

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002. 5

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 1

[3] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022. 2

[4] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. 5

[5] Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Zeroq: A novel zero shot quantization framework. In *CVPR*, pages 13169–13178, 2020. 2, 3

[6] Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018. 3

[7] Yoni Choukroun, Eli Kravchik, Fan Yang, and Pavel Kisilev. Low-bit quantization of neural networks for efficient inference. In *ICCVW*, pages 3009–3018, 2019. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 8

[9] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019. 7

[10] Jun Fang, Ali Shafiee, Hamzah Abdel-Aziz, David Thorsley, Georgios Georgiadis, and Joseph H Hassoun. Post-training piecewise linear quantization for deep neural networks. In *ECCV*, pages 69–86, 2020. 3

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 2

[12] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016. 2

[13] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *NeurIPS*, 29, 2016. 2

[14] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021. 2

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 7

[17] Junghyup Lee, Dohyung Kim, and Bumsub Ham. Network quantization with element-wise gradient scaling. In *CVPR*, pages 6448–6457, 2021. 2

[18] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 1

[19] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. *arXiv preprint arXiv:2211.02048*, 2022. 2

[20] Xiuyu Li, Long Lian, Yijiang Liu, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. *arXiv preprint arXiv:2302.04304*, 2023. 2, 3, 6, 7

[21] Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021. 3

[22] Zhikai Li, Liping Ma, Mengjuan Chen, Junrui Xiao, and Qingyi Gu. Patch similarity aware data-free quantization for vision transformers. In *ECCV*, pages 154–170, 2022. 3

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 7

[24] Zechun Liu, Wenhan Luo, Baoyuan Wu, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Binarizing deep network towards real-network performance. *IJCV*, 128:202–219, 2020. 3

[25] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *NeurIPS*, 34:28092–28103, 2021. 3

[26] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *ICML*, pages 7197–7206. PMLR, 2020. 3

[27] Yury Nahshan, Brian Chmiel, Chaim Baskin, Evgenii Zheltonozhskii, Ron Banner, Alex M Bronstein, and Avi Mendelson. Loss aware post-training quantization. *Machine Learning*, 110(11-12):3245–3262, 2021. 6

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 6

[30] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *TPAMI*, 2022. 1

[31] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. *arXiv preprint arXiv:2211.15736*, 2022. 2, 3, 4, 7

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 6

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[34] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *ICLR*, 2022. 1

[35] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021. 2

[36] Daniel Watson, William Chan, Jonathan Ho, and Mohammad Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2022. 3

[37] Xiuying Wei, Ruihao Gong, Yuhang Li, Xianglong Liu, and Fengwei Yu. Qdrop: randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022. 6

[38] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. *arXiv preprint arXiv:2303.08622*, 2023. 1

[39] 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. *arXiv preprint arXiv:1506.03365*, 2015. 7

[40] Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, and Qi Tian. Variational convolutional neural network pruning. In *CVPR*, pages 2780–2789, 2019. 2

[41] Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *ICML*, pages 7543–7552, 2019. 3

[42] Yunshan Zhong, Mingbao Lin, Gongrui Nan, Jianzhuang Liu, Baochang Zhang, Yonghong Tian, and Rongrong Ji. Intraq: Learning synthetic images with intra-class heterogeneity for zero-shot network quantization. In *CVPR*, pages 12339–12348, 2022. 2