TCAQ-DM: Timestep-Channel Adaptive Quantization for Diffusion Models

Haocheng Huang[1,2,*], **Jiaxin Chen**[1,2,*], **Jinyang Guo**[3], **Ruiyi Zhan**[1,2], **Yunhong Wang**[1,2,†]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]School of Artificial Intelligence, Beihang University, Beijing, China

Abstract

Diffusion models have achieved remarkable success in the image and video generation tasks. Nevertheless, they often require a large amount of memory and time overhead during inference, due to the complex network architecture and considerable number of timesteps for iterative diffusion. Recently, the post-training quantization (PTQ) technique has proved a promising way to reduce the inference cost by quantizing the float-point operations to low-bit ones. However, most of them fail to tackle with the large variations in the distribution of activations across distinct channels and timesteps, as well as the inconsistent of input between quantization and inference on diffusion models, thus leaving much room for improvement. To address the above issues, we propose a novel method dubbed Timestep-Channel Adaptive Quantization for Diffusion Models (TCAQ-DM). Specifically, we develop a timestep-channel joint reparameterization (TCR) module to balance the activation range along both the timesteps and channels, facilitating the successive reconstruction procedure. Subsequently, we employ a dynamically adaptive quantization (DAQ) module that mitigate the quantization error by selecting an optimal quantizer for each post-Softmax layers according to their specific types of distributions. Moreover, we present a progressively aligned reconstruction (PAR) strategy to mitigate the bias caused by the input mismatch. Extensive experiments on various benchmarks and distinct diffusion models demonstrate that the proposed method substantially outperforms the state-of-the-art approaches in most cases, especially yielding comparable FID metrics to the full precision model on CIFAR-10 in the W6A6 setting, while enabling generating available images in the W4A4 settings.

Extended version — https://dr-jiaxin-chen.github.io/page/

Introduction

Diffusion models (Ho, Jain, and Abbeel 2020) have emerged as one of the most prevailing generative models, with a wide range of applications including image generation (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021), image translation (Su et al. 2023; Tumanyan et al. 2023), super-resolution (Li et al. 2022; Gao et al. 2023; Wang et al. 2024b), and video generation (Ho et al. 2022; Chen et al. 2024). They gradually transform noises into high-quality images or video clips through an iterative diffusion process, based on a noise estimation network and a denoising sampler. Nevertheless, due to the complex network structure and the massive network forward propagation required during dozens or even hundreds of iterative timesteps, existing models are generally computationally expensive, making it inefficient during inference.

Many efforts have been made to accelerate the diffusion model, which can be roughly divided into two categories. The first category of methods (Chung, Sim, and Ye 2022; Lyu et al. 2022; Franzese et al. 2023) focuses on decreasing

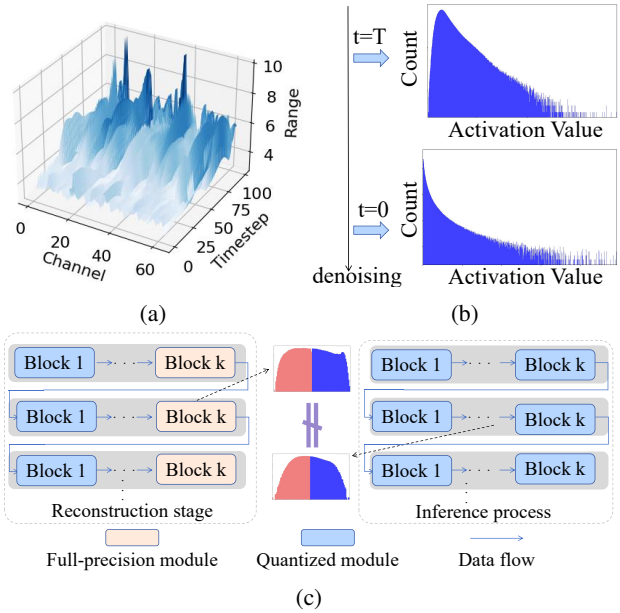


Figure 1: (a) Fluctuated activations per channels and timesteps in the convolutional layers (e.g. *up.0.block.0.conv1* of DDIM). (b) Dynamic changes of activation distributions in the post-Softmax layer (e.g. *down.1.attn.0* of DDIM) in distinct timesteps. (c) Misalignment between the intermediate data of the reconstruction stage in the quantization process and those in the inference process.

*Equal Contribution. †Corresponding Author.

the number of sampling timesteps. It is capable of linearly reducing the inference time cost without modifying the network structure, which however fails to decrease the model size. Alternatively, the second category of methods aims to expedite the inference through compressing the neural network by pruning (Castells et al. 2024; Zhang et al. 2024) and quantization (Shang et al. 2023; Wang et al. 2024a; Huang et al. 2024b).

In this paper, we mainly investigate the post-training quantization (PTQ) technique for diffusion models, considering that it reduces both the storage and inference time cost by mapping the float-point weights and activations of the networks into low-bit integers (Li et al. 2021; Nagel et al. 2020; Kuzmin et al. 2022; Dettmers et al. 2024), and is feasible for fast deployment without expensive re-training. Several works have attempted to explore the PTQ technique for diffusion models by either collecting calibration datasets across all timesteps (Shang et al. 2023; Li et al. 2023b; Huang et al. 2024b) or correcting the accumulation errors on iterative sampling(He et al. 2024; Yao et al. 2024). Nevertheless, most of them suffer from substantial performance degradation especially when quantizing under low bit-widths, as they fails to take the following characteristics of diffusion models into consideration based on our empirical observations: **1)** the range of activation in the convolutions layers often drastically fluctuates in both channels and timesteps as displayed in Fig. 1(a), which is prone to incur large quantization errors; **2)** the distribution of activations in the post-Softmax layers dynamically changes as the timestep decreases, and gradually exhibits a power-law-like shape during the diffusion process as shown in Fig. 1(b), resulting in nonnegligible quantization loss as most exiting works utilize one fix quantizer; **3)** existing reconstruction-based quantization methods often utilizes the output of the quantized model in the precedent block as input for reconstruction stage, which is not consistent with the inference process that adopt iterative sampling strategy as shown in Fig. 1(c), inevitably introducing bias and leaving much room for improvement.

To address the above issues, we propose a novel PTQ approach dubbed Timestep-Channel Adaptive Quantization for Diffusion Models (TCAQ-DM). As displayed in Fig. 2, we first develop a timestep-channel joint reparameterization (TCR) module tailored for quantizing the convolutional layer with severely fluctuated activations. This module uniformly splits the overall timesteps into groups, and in each group balances the originally unconstrained activations by employing a channel-wise reparameterization transformation with timestep-aware average weighting. Subsequently, we present a dynamically adaptive quantizer (DAQ) specifically designed for quantizing the post-Softmax activations with timestep-varying distributions (Clauset, Shalizi, and Newman 2009). It establishes an estimator to assess the likelihood of the activations from a particular timestep obeying the power-law distribution on each layer. The timesteps with high likelihood are assigned a log2 quantizer (Li et al. 2023c; Lin et al. 2022), which proved effective in quantizing the activations with power-law distributions, and those with low likelihood are dynamically handled by a uniform quantizer that is simple and efficient. Finally, to address the misalignment issue, we employ a progressively aligned reconstruction (PAR) strategy by incorporating the quantized inputs in the reconstruction stage of quantization process, in order to stay consistent with the inference process, thus further boosting the performance.

The main contributions of our work lie in three-fold:

- We propose a novel PTQ approach dubbed Timestep-Channel Adaptive Quantization for Diffusion Models (TCAQ-DM), by flexibly adapting to varying activation ranges and distributions in distinct channels and timesteps, and aligning the intermediate data in the quantization process with those in the inference process.
- We design a timestep-channel joint reparameterization (TCR) module to mitigate the influence of fluctuated activation ranges on quantization, and a dynamically adaptive quantizer (DAQ) to strengthen its flexibility in dealing with timestep-varying activation distributions in the post-Softmax layer, which reduces the quantization error especially under low bit-widths. We also develop a progressively aligned reconstruction (PAR) strategy to avoid the data inconsistency between quantization and inference, further boosting the performance.
- We conduct extensive experiments and ablation studies on various datasets and representative diffusion models, and demonstrate that our method remarkably outperforms the state-of-the-art PTQ approaches for diffusion models in most cases, especially under low bit-widths. Particularly, for the challenging W4A4 setting, our method generate available results, while most compared PTQ approaches yield nearly collapsed performance.

Related Work

Existing approaches for accelerating diffusion models roughly fall into two categories: building efficient diffusion models by reducing the sampling steps and compressing the network structures of diffusion models. For the later, we focus on the quantization based methods, and summarize the related works as below.

Efficient Diffusion Model

Diffusion models gradually apply Gaussian noise to real data in an iteratively process, as the preliminaries are provided in (Song, Meng, and Ermon 2021; He et al. 2024). For this process is time-consuming, many approaches have been proposed to obtain an efficient diffusion model by diminishing the sampling steps or skip some operation in inference time, which can be further divided into the training-based methods and the training-free ones. The former reduces the steps by model distillation (Luhman and Luhman 2021; Salimans and Ho 2022; Huang et al. 2024a) or sample trajectory learning (Lam et al. 2022; Watson et al. 2022; Zhao et al. 2024). And the later usually directly designs efficient samplers on pre-trained diffusion models, by developing implicit samplers (Song, Meng, and Ermon 2021), customized SDE, ODE solvers (Kim and Ye 2023; Zhang and Chen 2023; Zhou et al. 2024), or automatic search (Li et al. 2023a).

Some methods also develop the cache-based strategies (Ma, Fang, and Wang 2024). All of these method diminish the inference time in the timestep dimension without considering the delay in single model forward step. Despite decreasing the time cost, these methods fail to reduce the diffusion model size, thus still suffering from the high computational complexity and extensive storage consumption.

Model Quantization

In contrast to the aforementioned approaches that focus on reducing the number of sampling steps, model quantization takes a different route by aiming to compress diffusion neural networks. This is achieved by mapping the floating-point weights and activations of the network into low-bit representations. The primary objective of this quantization process is to significantly decrease both inference latency and memory overhead associated with the model's operation. By converting high-precision values into lower-bit formats, we can enhance the efficiency of the model while maintaining its performance. We provide a comprehensive review of existing model quantization methods and their respective contributions to this field as below.

General Quantization Methods Current quantization methods for general purpose mainly consists of the quantization-aware training (QAT) (Gong et al. 2019; Zhang et al. 2023; Chu, Li, and Zhang 2024) and the post-training quantization (PTQ) (Nagel et al. 2020; Li et al. 2021; Wei et al. 2022; Wu et al. 2024). The QAT methods simulate the quantization process during the training phase, with the goal of minimizing quantization error. While these methods often achieve high accuracy even at low bit-widths, they come with significant training costs, as they necessitate the retraining of all weights using extensive large-scale training datasets. In contrast, PTQ methods take a more direct approach by quantizing weights and activations based on a smaller-scale calibration set. This approach does not involve fine-tuning the weights during the quantization process, making PTQ considerably more efficient in terms of both data and computational requirements. By leveraging a limited amount of calibration data, PTQ methods can significantly reduce the associated computational costs while still delivering competitive performance.

PTQ for Diffusion Models Directly applying the general quantization methods to diffusion models usually results in poor performance. The primary reason for this discrepancy is that diffusion models employ a distinct inference method compared to traditional models. The variation in activation distributions across timesteps complicates the accurate estimation of quantization parameters. To deal with this problem, PTQ4DM (Shang et al. 2023) collects calibration data from various timesteps, and makes the first attempt on quantizing diffusion models in 8 bit-width with slight performance degradation. Q-Diffusion (Li et al. 2023b) further enhances the performance by dividing the skip connection layer and propose a novel sampling strategy about calibration datasets. PTQD (He et al. 2024) eliminates the accumulation errors by correcting samplers and collecting the output at each timestep for calibration, proposing a novel vision of diminishing the quantiztaion errors. APQ-DM (Wang et al. 2024a) designs a dynamic grouping strategy and chooses a calibration set according to the structural risk minimization principle. The grouping strategy is usually adopted in future works. TFMQ-DM (Huang et al. 2024b) mitigates the information bias at different timesteps caused by quantization and proposes a TIB block to protect the timestep information. PCR (Tang et al. 2024) proposes a progressive quantization method to avoid the accumulation of errors during quantization for diffusion models, along with an activation relaxing strategy to reduce errors in sensitive modules. TMPQ-DM (Sun et al. 2024) propose a new vision of quantization for diffusion models, jointly optimizing timesteps reduction and model quantization to achieve a superior performance-efficiency trade-off.. However, these methods fail to jointly handle the fluctuated activation ranges and distributions in distinct timesteps and channels, and neglect the inconsistency between the inputs of the reconstruction stage in the quantization process and those in the inference process, thus inclining to incur large quantization errors.

Methodology

Framework Overview

Our method is based on PTQ that aims to compute the scaling factor s and the zero point z, and map the float-point data to integers via the following formula:

$$\hat{\mathbf{x}} = \Phi\left(\lfloor\frac{\mathbf{x}}{s}\rceil + z, 0, 2^{bit} - 1\right), \tag{1}$$

where $\hat{\mathbf{x}}$ denotes the quantized value of the float-point weights or activations $\mathbf{x}$, $\mathbf{s}$ and $\mathbf{z}$ denote the scaling factor and zero point, respectively, Φ indicates the function that clips the value range to $[0, 2^{bit} - 1]$, $\lfloor\cdot\rceil$ denotes the rounding operation, and bit is the bit-width. The primary objective of model quantization is to identify the optimal parameters $\mathbf{s}$ and $\mathbf{z}$ that minimize errors. Generally, by following existing works (Li et al. 2023b; Huang et al. 2024b), the overall quantization process includes the initialization stage that roughly search the quantization parameters, and the reconstruction stage that further refine the quantization parameters.

As shown in Fig. 2, different from current PTQ approaches, we propose three novel components, including the timestep-channel joint reparameterization (TCR) and the dynamically adaptive quantizer (DAQ) for the initialization stage, as well as the progressively aligned reconstruction (PAR) strategy for the reconstruction stage. TCR simultaneously mitigates the fluctuations of activation ranges in both distinct timesteps and channels in the specific convolutional layers, and DAQ adapts to varying activation distribution in the post-Softmax layers, both of which facilitate reducing the quantization error. PAR further boosts the performance by iteratively generating a calibration set that is aligned with the data flow in the inference process, which further diminish the loss in reconstruction stage. The technical details are described in the rest part of this section.

Timestep-Channel Joint Reparameterization

The activations of diffusion models' convolution layers exhibit significant fluctuation along the both timesteps and

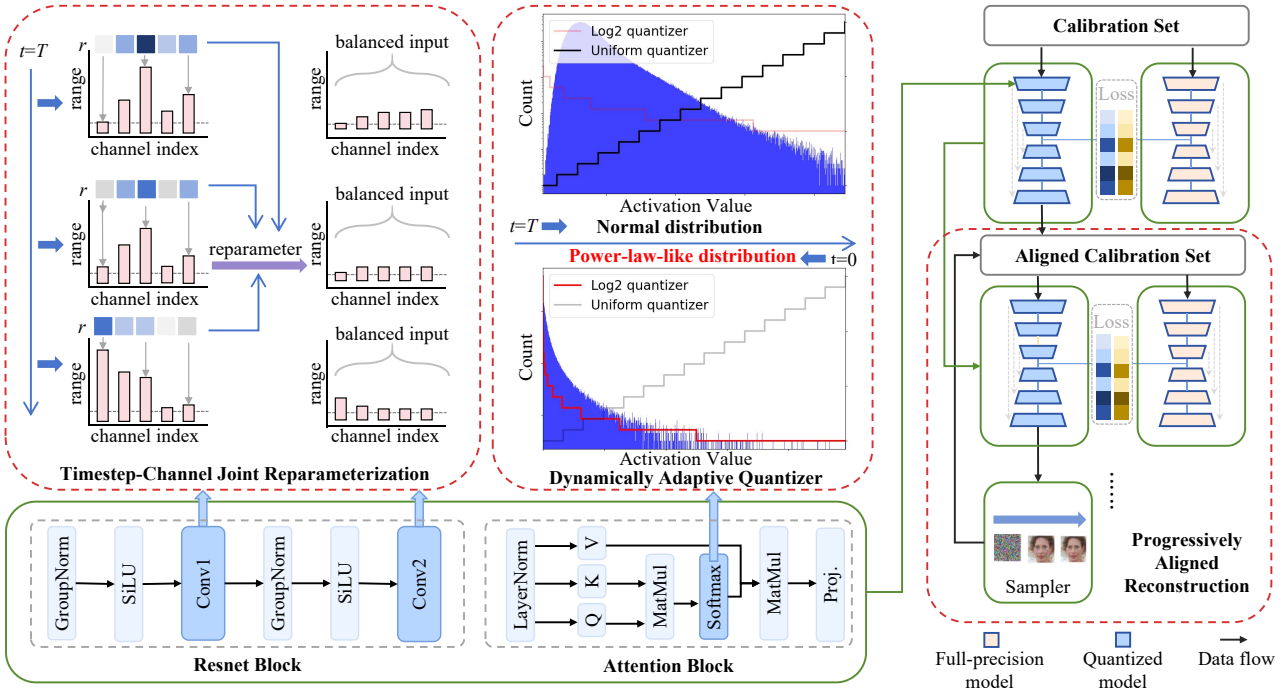


Figure 2: Overview of our proposed method. In the initialization stage, we develop the timestep-channel joint reparameterization (TCR) and the dynamically adaptive quantizer (DAQ) to mitigate the fluctuated activation ranges in the convolutional layers, and the timestep-varying activation distributions in the post-Softmax layers, respectively. In the reconstruction stage, we design the progressively aligned reconstruction (PAR) strategy to further improve the generation performance by aligning the data flow in the quantization process with that in the inference process.

channels. The interplay between these two dimensions renders activation quantization considerably more arduous. To address this issue, we firstly group the activation quantization parameters uniformly under the inference timesteps:

$$S = \{s_0, s_1, ...s_{T-1}\}, \quad Z = \{z_0, z_1, ...z_{T-1}\}, \tag{2}$$

where T denotes the denoising step, S and Z represent the scaling factor and zero point of the activation quantizer, respectively. We assign each inference timestep to a specific group with different quantization parameters. It is noteworthy that the time cost of searching S and Z can be substantially diminished when leveraging a less-step sampling technique.

In addition to the timestep dimension, post-training quantization (PTQ) methods for diffusion models also suffer from activation variability across different channels. Inspired by recent quantization works in Vision Transformers (ViT) (Li et al. 2023c), we propose timestep-channel joint reparameterization module to solve this problem. This module rescales the input range by aggregating the values across all timesteps. Specifically, for a particular convolution layer with weights W and input $\mathbf{X}^t \in \mathbb{R}^{N\times C\times W\times H}$ of timestep t, we aim to find a scaling vector $\mathbf{r}^t \in \mathbb{R}^C$, then reparameterize the activation and corresponding weight as:

$$\mathbf{X}^{t'}_{:,j} = \mathbf{X}^t_{:,j} \oslash \mathbf{r}^t_j, \quad \mathbf{W}'_{:,j} = \mathbf{W}_{:,j} \odot \mathbf{r}^t_j, \tag{3}$$

where $\oslash$ and $\odot$ denote the broadcast division and multiplication, respectively. For a convolution layer, which is equivalent to a linear affine operation, this reparameterization will retain the output while shifting the value range from activations to weights. A tailored $\mathbf{r}^t$ for activation $\mathbf{X}^t$ aligns activations between channels:

$$\mathbf{r}^t_j = \max(\mathbf{X}_{:,j})/s^t_{tar}, \tag{4}$$

where s^t_{tar} is a pre-specific target range of timestep t. However, diffusion models have different activations between timesteps while sharing the same weights. Rescaling for each timestep will produce multiple weights, causing a high storage cost. Therefore, it is necessary to combine $\mathbf{r}^t$ of all timesteps to general scaling vector $\mathbf{r}$.

When performing reparameterization, we find some of the channels have a small value range in most of the denoising steps, but suddenly increase in a few steps and become outliers. To limit these channels' value range, we first set the minimum of the maximum value of all channels as the rescale target s^t, to ensure that the value range of all channels will not be further expanded. Then, we use the maximum value of each channel as the weight to sum the activation across all timesteps, ensuring the scaling vector on the timestep with larger activation receives more attention. The final formula is shown as:

$$\begin{aligned} s^t_{tar} &= \min(\max(\mathbf{X}^t_{:,d})_{1\le d\le D}), \\ \mathbf{r}^t_d = \frac{\max(\mathbf{X}^t_{:,d})}{s^t_{tar}}, \mathbf{r}^s_d &= \frac{\sum_t r^t_d * \max(\mathbf{X}^t_{:,d})}{\sum_t \max(\mathbf{X}^t_{:,d})}, \\ \widetilde{\mathbf{X}}^t_{:,d} = \mathbf{X}^t_{:,d} \oslash \mathbf{r}^s_d, &\quad \widetilde{\mathbf{W}}_{:,d} = \mathbf{W}_{:,d} \odot \mathbf{r}^s_d. \end{aligned} \tag{5}$$

Since this method will enlarge the weight values, which may lead to insignificant performance improvement when applying weight relative low-bit quantization like W4A8, we use a hyper-parameter R_{tru} to truncate the scaling vector to a limit range in these settings.

Dynamically Adaptive Quantizer

The activation of post-Softmax often shows a power-law distribution. The uniform quantizer cannot balance the quantization between the long-tail and the small value peak of this type of distribution, which often leads to performance degradation. Previous methods (Lin et al. 2022) attempt to use a log2 quantizer to fit the feature of the post-Softmax. It maps the float numbers to a logarithmic function with a base of 2:

$$\hat{\mathbf{x}} = \Phi(\lfloor -\log_2 \frac{\mathbf{x}}{s} \rceil, 0, 2^{bit-1}), \tilde{\mathbf{x}} = s * 2^{-\hat{\mathbf{x}}}, \quad (6)$$

where $\hat{\mathbf{x}}$ and $\tilde{\mathbf{x}}$ indicates quantized value and dequantized value of $\mathbf{x}$, respectively.

However, the post-Softmax activation in diffusion models also suffers from timestep variance. In the early denoising steps of certain blocks, activations is only distributed within a limited range, where the log quantizer will lead to a larger quantization error. Directly applying the log2 quantizer in diffusion models may even perform poorly in high-bit setting, as shown in Table 6. Therefore, we propose a dynamically adaptive quantizer that could select whether to use the log2 quantizer for a specific timestep of a post-Softmax layer, based on its mathematical properties. Specifically, we use the Maximum Likelihood Estimation method to fit the each layers activation on every timestep to a power-law distribution (Clauset, Shalizi, and Newman 2009):

$$P(X \geq x) = cx^{-\alpha}. \quad (7)$$

Since the activations of model could be collected in advance, this operation could be conducted offline. Then we calculate the ratio of the likelihood estimation results for the power-law distribution and other distributions (e.g., log-normal or exponential) as R_g, and perform log2 quantizer to the specific timestep where the ratio is greater than zero:

$$\hat{\mathbf{x}} = \begin{cases} \lfloor \frac{\mathbf{x}}{s_g} + z_g \rceil, & if\ R_g \leq 0; \\ \lfloor -\log_2 \frac{\mathbf{x}}{s_g} \rceil, & if\ R_g > 0. \end{cases} \quad (8)$$

It is worth noting that DAQ performs offline, and only introduces a small amount of extra computational overhead, roughly 3% of the overall cost, which is affordable in our implementation.

Progressively Aligned Reconstruction

Existing post-training quantization (PTQ) methods frequently incorporate a block reconstruction stage aimed at enhancing overall performance. However, the iterative inference process inherent to diffusion models introduces a significant inconsistency issue between the reconstruction stage and the inference process, as illustrated in Figure 1c. When existing PTQ methods are applied to diffusion models, they tend to generate biased input distributions. This bias arises because the model is quantized in a single forward pass during the reconstruction phase, while it is subsequently invoked iteratively during the inference phase.

For diffusion models sharing weights across all timesteps, quantizing blocks in the same order as the denoising process is challenging. As an alternative method, we propose progressively aligned reconstruction to iteratively align the inputs. In particular, after the basic reconstruction with BRECQ (Li et al. 2021), we continuously sample a new calibration set using the quantized model and then utilize this aligned set to reconstruct the model. This phase will repeated in multiple rounds with fewer iterations than the first one. We refer to the *Supplementary Material* for the detailed algorithm of the proposed PAR method.

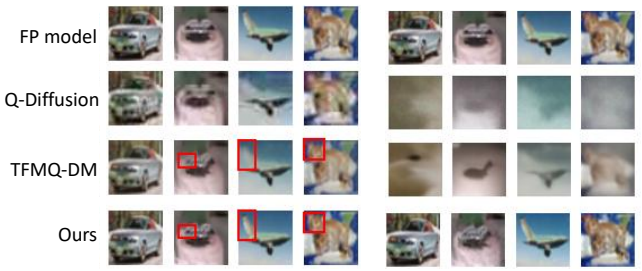
Method	Bit-width	FID (↓)	IS (↑)
FP model	W32A32	4.14	9.12
PTQ4DM	W4A32	5.65	9.02
Q-Diffusion	W4A32	5.09	8.78
TFMQ-DM	W4A32	4.73	**9.14**
Ours	W4A32	**4.28**	9.09
PTQ4DM	W8A8	5.69	9.31
Q-Diffusion*	W8A8	4.78	8.89
APQ-DM	W8A8	4.24	9.07
TFMQ-DM	W8A8	4.24	9.07
TAC-Diffusion	W8A8	**3.68**	**9.49**
Ours	W8A8	4.09	9.08
PTQ4DM	W4A8	10.12	**9.31**
Q-Diffusion	W4A8	4.93	9.12
TFMQ-DM	W4A8	4.78	9.13
TAC-Diffusion	W4A8	4.89	9.15
Ours	W4A8	**4.59**	9.17
PTQ4DM*	W6A6	61.83	7.10
Q-Diffusion*	W6A6	26.06	9.02
TFMQ-DM*	W6A6	9.59	8.84
Ours	W6A6	**4.40**	**9.04**
PTQ4DM*	W4A4	375.12	0.45
Q-Diffusion*	W4A4	384.21	0.71
TFMQ-DM*	W4A4	236.63	3.19
Ours	W4A4	**6.38**	**8.70**

Table 1: Comparison results on CIFAR-10 based on DDIM model with 100 timesteps. * means directly rerunning the open-resource code.

Experimental Results and Analysis

Experimental Settings

By following existing works (Li et al. 2023b; Huang et al. 2024b), we evaluate our proposed method on the *ImageNet* dataset (Deng et al. 2009) by using LDM-4 for the conditional generation task. For the unconditional generation task, we conduct experiments on the *CIFAR-10* dataset (Krizhevsky, Hinton et al. 2009) by using DDIM (Song, Meng, and Ermon 2021), *LSUN-Bedrooms* and *LSUN-Churches* dataset (Yu et al. 2015) based on LDM-4. Similar to (Li et al. 2023b; Huang et al. 2024b), we adopt the evaluation metrics including Fréchet Inception Distance (FID)



(a) W6A6 (b) W4A4

Figure 3: Visualization of images generated by quantized models via various PTQ methods, indicating that our method generates images with better visual details in W6A6, and outputs available images in the challenging W4A4 setting.

(Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016) on CIFAR-10 and ImageNet, and additionally report sliced FID (sFID) (Salimans et al. 2016) when using LDM. FID and sFID are based on measuring the similarity between generated and real images by calculating the mean and covariance of features extracted by the Inception network, while IS assesses the quality and diversity of the generated images based on the predicted distribution from the classification model like Inception.

Implementation Details

By following (Li et al. 2023b; Huang et al. 2024b), we perform the channel-wise quantization on weights and layer-wise quantization on activations. Similar to (Huang et al. 2024b), we maintain the input and output layers of the model in full precision and generate calibration sets by the full-precision model. For the weight quantization, we conduct BRECQ with 20,000 iterations for initialization, and 10,000 iterations for each progressive round with a batch size of 16. For the activation quantization, we use the commonly used hyper-parameter search method as depicted in RepQ-ViT (Li et al. 2023c) with a batch size of 64. All experiments are conducted with an 8-bit post-Softmax layer unless being specifically claimed. We set R_{tru} to 3 when performing TCR on W4A8 bit-width, while not implementing truncation operation on remaining experiments. All experiments are conducted on a single RTX4090 GPU.

Comparison to the State-of-the-Art Methods

We compare our method to the state-of-the-art PTQ approaches, including PTQ4DM (Shang et al. 2023), Q-Diffusion (Li et al. 2023b), PTQD (He et al. 2024), APQ-DM (Wang et al. 2024a), TFMQ-DM (Huang et al. 2024b) and TAC-Diffusion (Yao et al. 2024).

Unconditional Image Generation. On CIFAR-10 with DDIM, we follow the same setting as Q-Diffusion (Li et al. 2023b). As shown in Table 1 and Fig. 3, our method reaches competitive FIDs compared to the full-precision model, and outperforms the state-of-the-art approaches in most cases. As for W4A4, the performance of existing approaches is collapsed with large FIDs. In contrast, our method achieves promising results with less than 2.3 increase in FID, compared to the full-precision model. We provide more visualization results in the *Supplementary Material*.

Method	Bit-width	FID (↓)	sFID (↓)
FP model	W32A32 S32	2.98	7.09
Q-Diffusion	W4A32 S8	4.20	7.66
PTQD	W4A32 S8	4.42	7.88
TFMQ-DM	W4A32 S32	3.60	7.61
Ours	W4A32 S8	**3.55**	**7.54**
Q-Diffusion	W8A8 S8	4.51	8.17
PTQD	W8A8 S8	3.75	9.89
TFMQ-DM	W8A8 S32	3.14	**7.26**
Ours	W8A8 S8	3.21	7.59
Ours	W8A8 S32	**3.11**	7.34
Q-Diffusion	W4A8 S8	6.40	17.93
PTQD	W4A8 S8	5.94	15.16
TFMQ-DM	W4A8 S32	3.68	**7.65**
TAC-Diffusion	W4A8 S8	4.94	-
Ours	W4A8 S8	3.70	7.69
Ours	W4A8 S32	**3.65**	7.67
Q-Diffusion*	W4A4 S8	334.83	190.89
PTQD*	W4A4 S8	321.47	181.61
TFMQ-DM*	W4A4 S32	118.70	80.85
Ours	W4A4 S8	**16.43**	**23.85**

Table 2: Comparison results on LSUN-bedrooms with LDM-4 using the DDIM sampler with 200 timesteps.

On LSUN-bedrooms with Latent Diffusion Model (LDM), we adopt the same settings as TFMQ-DM (Huang et al. 2024b), except for the post-Softmax quantization bit-width. As shown in Table 2, our method with 8-bit post-Softmax significantly promotes the FID, compared to the other methods using the same setting. And when using 32 bits, our method is comparable to TFMQ-DM in most cases, and remarkably outperforms it under the W4A4 setting.

Conditional Image Generation. On ImageNet, we employ a denoising process with 20 iterations, following the same setting as TFMQ-DM. As shown in Table 4, our method improves FIDs of TFMQ-DM by 0.21 and 1.32 under W8A8 and W4A32, respectively. As for the challenging W4A4 settings, despite that there is a gap between the full-precision model and the quantized model, our method still reaches a comparable performance, while the compared methods performs poorly with extremely large FIDs and sFIDs.

Ablation Study

To evaluate the effectiveness of each proposed component, we perform ablation study on CIFAR-10 based on DDIM, by employing BRECQ as the baseline method. The effectiveness of different modules show as below.

Effectiveness of TCR As summarized in Table 5, the proposed TCR module reduces the FID by 0.66 and 22.01 in the W8A8 setting and the W6A6 setting respectively, compared to the baseline. Moreover, this module plays a crucial role

Method	Bit-width	FID (↓)	sFID (↓)
FP model	W32A32	4.08	10.89
Q-Diffusion	W4A32	4.55	11.90
PTQD	W4A32	4.67	13.68
TFMQ-DM	W4A32	4.07	**11.41**
Ours	W4A32	**4.00**	11.72
Q-Diffusion	W8A8	4.87	12.23
PTQD	W8A8	4.89	12.23
TFMQ-DM	W8A8	**4.01**	10.98
Ours	W8A8	4.05	**10.82**
Q-Diffusion	W4A8	4.66	13.97
PTQD	W4A8	5.10	13.23
TFMQ-DM	W4A8	4.14	**11.46**
Ours	W4A8	**4.13**	11.57
Q-Diffusion*	W4A4	360.32	191.75
PTQD*	W4A4	358.34	180.26
TFMQ-DM*	W4A4	236.52	186.44
Ours	W4A4	**29.17**	**35.89**

Table 3: Comparison results on LSUN-Churches based on LDM-4 by the DDIM sampler with 500 timesteps.

Method	Bit-width	FID (↓)	IS (↑)	sFID (↓)
FP model	W32A32	10.91	235.64	7.67
Q-Diffusion	W4A32	11.87	213.56	8.76
PTQD	W4A32	11.65	210.78	9.06
TFMQ-DM	W4A32	10.50	223.81	7.98
Ours	W4A32	**10.50**	**234.51**	**6.66**
Q-Diffusion	W8A8	12.80	187.65	9.87
PTQD	W8A8	11.94	153.92	8.03
TFMQ-DM	W8A8	10.79	198.86	7.65
Ours	W8A8	**10.58**	**239.41**	**7.54**
Q-Diffusion	W4A8	10.68	212.51	14.85
PTQD	W4A8	10.40	214.73	12.63
TFMQ-DM	W4A8	10.29	221.82	**7.35**
Ours	W4A8	**9.97**	**232.87**	7.67
Q-Diffusion*	W4A4	376.54	1.69	165.39
PTQD*	W4A4	361.29	1.87	190.48
TFMQ-DM*	W4A4	210.06	2.95	192.81
Ours	W4A4	**30.69**	**86.11**	**18.92**

Table 4: Comparison results on ImageNet based on LDM-4 by using the DDIM sampler with 20 timesteps.

in maintaining a comparable performance in low-bit quantization such as W4A4, with a substantial improvement of FID.

Effectiveness of DAQ In terms of the DAQ module, it further promotes the performance across all bit-widths, especially improving the FID by 0.17 and 0.45 in the W4A8 and W4A4 settings, respectively. Moreover, as displayed in Table 6, DAQ achieves stable improvements across all Softmax bit-widths, compared with the uniform and log2 quantizers.

Effectiveness of PAR As shown in Table 5, the proposed PAR module also obtains improvements, reducing the FID significantly by 2.71 in the W4A4 setting and promoting the performance in other bit-widths.

More experimental results about hyper-parameters are provided in the *Supplementary Material*.

Method	Bit-width	FID (↓)	IS (↑)
Baseline	W8A8	4.78	8.87
+TCR	W8A8	4.12	9.04
+TCR+DAQ	W8A8	4.11	9.06
+TCR+DAQ+PAR	W8A8	**4.09**	**9.08**
Baseline	W4A8	4.93	9.12
+TCR	W4A8	4.76	9.02
+TCR+DAQ	W4A8	4.59	8.97
+TCR+DAQ+PAR	W4A8	**4.59**	**9.17**
Baseline	W6A6	26.60	9.02
+TCR	W6A6	4.59	8.99
+TCR+DAQ	W6A6	4.47	**9.09**
+TCR+DAQ+PAR	W6A6	**4.40**	9.04
Baseline	W4A4	371.61	0.41
+TCR	W4A4	9.54	8.57
+TCR+DAQ	W4A4	9.09	8.37
+TCR+DAQ+PAR	W4A4	**6.38**	**8.70**

Table 5: Ablation results of the proposed main components on CIFAR-10 based on DDIM with 100 timesteps.

Method	Bit-width	FID (↓)	IS (↑)
Log2 quantizer	W6A6 S8	4.97	8.95
Uniform quantizer	W6A6 S8	4.66	**9.08**
DAQ (Ours)	W6A6 S8	**4.42**	9.04
Log2 quantizer	W6A6 S6	4.77	**8**.63
Uniform quantizer	W6A6 S6	4.87	8.53
DAQ (Ours)	W6A6 S6	**4.61**	**9.05**
Log2 quantizer	W6A6 S4	4.76	9.01
uniform quantizer	W6A6 S4	14.20	8.06
DAQ (Ours)	W6A6 S4	**4.66**	**9.07**

Table 6: Ablation results on distinct quantizers under different post-Softmax bit-widths on CIFAR-10 based on DDIM with 100 timesteps.

Conclusion

In this work, we propose a novel post-training quantization method, dubbed Timestep-Channel Adaptive Quantization for Diffusion Models (TCAQ-DM). We first develop the timestep-channel joint reparameterization (TCR) to mitigate the fluctuated activation ranges. Subsequently, we employ a dynamically adaptive quantizer (DAQ) to reduce the quantization errors caused by the timestep-varying activation distributions. Moreover, we design a progressively aligned reconstruction (PAR) strategy to align the data in the reconstruction stage of the quantization process with that during inference, further boosting the performance. Extensive experimental results on distinct dataset and diffusion models as well as extensive ablation results clearly demonstrate the superiority of the proposed approach under low bit-widths.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China (Nos. 62202034, 62306025,

92367204), the Beijing Natural Science Foundation (No. 4242044), the Beijing Municipal Science and Technology Project (No. Z231100010323002), the Aeronautical Science Foundation of China (No. 2023Z071051002), the Research Program of State Key Laboratory of Virtual Reality Technology and Systems, and the Fundamental Research Funds for the Central Universities.

References

Castells, T.; Song, H.-K.; Kim, B.-K.; and Choi, S. 2024. LD-Pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 821–830.

Chen, S.; Xu, M.; Ren, J.; Cong, Y.; He, S.; Xie, Y.; Sinha, A.; Luo, P.; Xiang, T.; and Perez-Rua, J.-M. 2024. GenTron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6441–6451.

Chu, X.; Li, L.; and Zhang, B. 2024. Make repvgg greater again: A quantization-qware approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, 11624–11632.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12413–12422.

Clauset, A.; Shalizi, C. R.; and Newman, M. E. 2009. Power-law distributions in empirical data. *SIAM review*, 661–703.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 248–255.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2024. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, 10088–10115.

Franzese, G.; Rossi, S.; Yang, L.; Finamore, A.; Rossi, D.; Filippone, M.; and Michiardi, P. 2023. How much is enough? A study on diffusion times in score-based generative models. *Entropy*, 633–643.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10021–10030.

Gong, R.; Liu, X.; Jiang, S.; Li, T.; Hu, P.; Lin, J.; Yu, F.; and Yan, J. 2019. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4852–4861.

He, Y.; Liu, L.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2024. Ptqd: Accurate post-training quantization for diffusion models. In *Advances in Neural Information Processing Systems*, 13237–13249.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. In *Advances in Neural Information Processing Systems*, 8633–8646.

Huang, T.; Zhang, Y.; Zheng, M.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024a. Knowledge diffusion for distillation. In *Advances in Neural Information Processing Systems*, 65299–65316.

Huang, Y.; Gong, R.; Liu, J.; Chen, T.; and Liu, X. 2024b. TFMQ-DM: Temporal feature maintenance quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7362–7371.

Kim, B.; and Ye, J. C. 2023. Denoising MCMC for accelerating diffusion-based generative models. In *Proceedings of the International Conference on Machine Learning*, 16955–16977.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, Toronto, ON, Canada.

Kuzmin, A.; Van Baalen, M.; Ren, Y.; Nagel, M.; Peters, J.; and Blankevoort, T. 2022. Fp8 quantization: The power of the exponent. In *Advances in Neural Information Processing Systems*, 14651–14662.

Lam, M. W. Y.; Wang, J.; Su, D.; and Yu, D. 2022. BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis. In *Proceedings of the International Conference on Learning Representations*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 47–59.

Li, L.; Li, H.; Zheng, X.; Wu, J.; Xiao, X.; Wang, R.; Zheng, M.; Pan, X.; Chao, F.; and Ji, R. 2023a. AutoDiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7082–7091.

Li, X.; Liu, Y.; Lian, L.; Yang, H.; Dong, Z.; Kang, D.; Zhang, S.; and Keutzer, K. 2023b. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17535–17545.

Li, Y.; Gong, R.; Tan, X.; Yang, Y.; Hu, P.; Zhang, Q.; Yu, F.; Wang, W.; and Gu, S. 2021. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *Proceedings of the International Conference on Learning Representations*.

Li, Z.; Xiao, J.; Yang, L.; and Gu, Q. 2023c. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17227–17236.

Lin, Y.; Zhang, T.; Sun, P.; Li, Z.; and Zhou, S. 2022. FQ-ViT: Post-training quantization for fully quantized vision transformer. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1173–1179.

Luhman, E.; and Luhman, T. 2021. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*.

Lyu, Z.; Xu, X.; Yang, C.; Lin, D.; and Dai, B. 2022. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*.

Ma, X.; Fang, G.; and Wang, X. 2024. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15762–15772.

Nagel, M.; Amjad, R. A.; Van Baalen, M.; Louizos, C.; and Blankevoort, T. 2020. Up or down? Adaptive rounding for post-training quantization. In *Proceedings of the International Conference on Machine Learning*, 7197–7206.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2226–2234.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. In *Proceedings of the International Conference on Learning Representations*.

Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; and Yan, Y. 2023. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1972–1981.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*.

Su, X.; Song, J.; Meng, C.; and Ermon, S. 2023. Dual diffusion implicit bridges for image-to-image translation. In *Proceedings of the International Conference on Learning Representations*.

Sun, H.; Tang, C.; Wang, Z.; Meng, Y.; Jiang, J.; Ma, X.; and Zhu, W. 2024. TMPQ-DM: Joint timestep reduction and quantization precision selection for efficient diffusion models. *arXiv preprint arXiv:2404.09532*.

Tang, S.; Wang, X.; Chen, H.; Guan, C.; Wu, Z.; Tang, Y.; and Zhu, W. 2024. Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models. In *Proceedings of the European Conference on Computer Vision*, 404–420.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, C.; Wang, Z.; Xu, X.; Tang, Y.; Zhou, J.; and Lu, J. 2024a. Towards accurate post-training quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16026–16035.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2024b. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25796–25805.

Watson, D.; Chan, W.; Ho, J.; and Norouzi, M. 2022. Learning fast samplers for diffusion models by differentiating through sample quality. In *Proceedings of the International Conference on Learning Representations*.

Wei, X.; Gong, R.; Li, Y.; Liu, X.; and Yu, F. 2022. QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. In *Proceedings of the International Conference on Learning Representations*.

Wu, Z.; Chen, J.; Zhong, H.; Huang, D.; and Wang, Y. 2024. AdaLog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *Proceedings of the European Conference on Computer Vision*, 411–427.

Yao, Y.; Tian, F.; Chen, J.; Lin, H.; Dai, G.; Liu, Y.; and Wang, J. 2024. Timestep-aware correction for quantized diffusion models. In *Proceedings of the European Conference on Computer Vision*, 215–232.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*.

Zhang, D.; Li, S.; Chen, C.; Xie, Q.; and Lu, H. 2024. Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models. *arXiv preprint arXiv:2404.11098*.

Zhang, L.; He, Y.; Lou, Z.; Ye, X.; Wang, Y.; and Zhou, H. 2023. Root quantization: A self-adaptive supplement STE. *Appl. Intell.*, 6266–6275.

Zhang, Q.; and Chen, Y. 2023. Fast sampling of diffusion models with exponential integrator. In *Proceedings of the International Conference on Learning Representations*.

Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; and Lu, J. 2024. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 49842–49869.

Zhou, Z.; Chen, D.; Wang, C.; and Chen, C. 2024. Fast ode-based sampling for diffusion models in around 5 steps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7777–7786.