MPQ-DM: Mixed Precision Quantization for Extremely Low Bit Diffusion Models

Weilun Feng[1,2*], Haotong Qin[3*], Chuanguang Yang[1†], Zhulin An[1†], Libo Huang[1], Boyu Diao[1], Fei Wang[1], Renshuai Tao[4], Yongjun Xu[1], Michele Magno[3]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]ETH Zurich
[4]Beijing Jiaotong University
{fengweilun24s, yangchuanguang, anzhulin, diaoboyu2012, wangfei, xyj}@ict.ac.cn,
{haotong.qin,michele.magno}@pbl.ee.ethz.ch, www.huanglibo@gmail.com, rstao@bjtu.edu.cn

Abstract

Diffusion models have received wide attention in generation tasks. However, the expensive computation cost prevents the application of diffusion models in resource-constrained scenarios. Quantization emerges as a practical solution that significantly saves storage and computation by reducing the bit-width of parameters. However, the existing quantization methods for diffusion models still cause severe degradation in performance, especially under extremely low bit-widths (2-4 bit). The primary decrease in performance comes from the significant discretization of activation values at low bit quantization. Too few activation candidates are unfriendly for outlier significant weight channel quantization, and the discretized features prevent stable learning over different time steps of the diffusion model. This paper presents **MPQ-DM**, a Mixed-Precision Quantization method for Diffusion Models. The proposed MPQ-DM mainly relies on two techniques: (1) To mitigate the quantization error caused by outlier severe weight channels, we propose an *Outlier-Driven Mixed Quantization* (OMQ) technique that uses $Kurtosis$ to quantify outlier salient channels and apply optimized intra-layer mixed-precision bit-width allocation to recover accuracy performance within target efficiency. (2) To robustly learn representations crossing time steps, we construct a *Time-Smoothed Relation Distillation* (TRD) scheme between the quantized diffusion model and its full-precision counterpart, transferring discrete and continuous latent to a unified relation space to reduce the representation inconsistency. Comprehensive experiments demonstrate that MPQ-DM achieves significant accuracy gains under extremely low bit-widths compared with SOTA quantization methods. MPQ-DM achieves a 58% FID decrease under W2A4 setting compared with baseline, while all other methods even collapse.

Code — https://github.com/cantbebetter2/MPQ-DM

1 Introduction

Diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021) have demonstrated remarkable capabilities in generation tasks (Rombach et al. 2022; Song et al. 2020; Mei and Patel 2023; Yang et al. 2024b; Liu et al. 2023b). However, the iterative denoising process and the massive of parameters in order to cope with the high-resolution demand seriously hinder the wide deployment of the diffusion model in edge devices with limited computing resources (Liu et al. 2024a; Dai et al. 2024).

(a) Baseline (b) MPQ-DM

Figure 1: Visualization of samples generated by baseline and MPQ-DM under W2A6 bit-width.

As an effective model compression approach that reduces the floating-point parameters to low-bit integers, model quantization can simultaneously reduce the model size and improve the inference speed (Gholami et al. 2022). This technique has been widely used in CNNs (Pilipović, Bulić, and Risojević 2018; Ding et al. 2024; Chen et al. 2024) and Transformers (Chitty-Venkata et al. 2023). Model quantization is mainly divided into post-training quantization (PTQ) (Hubara et al. 2020; Wang et al. 2020; Wei et al. 2022; Liu et al. 2023a) and quantization-aware training (QAT) (Esser et al. 2019; Jacob et al. 2018; Krishnamoorthi 1806). QAT often requires a full amount of raw data to fine-tune model weights and quantization parameters. Therefore, it requires a fine-tuning time equivalent to original training but performs well on extremely low bit-width (2-4 bit) or even binarization (Qin et al. 2020; Zheng et al. 2024). PTQ only requires a small amount of calibration data to fine-tune the quantization parameters. Thus, a shorter calibration time is required, but model performance cannot be guaranteed. To solve this problem, quantization-aware low-rank fine-tuning (QLORA-FT) scheme (He et al. 2023) proposes to add a LoRA (Hu et al. 2021) module to perform a low-rank update on model quantized weights to enhance the performance of

*These authors contributed equally.
†Corresponding authors.

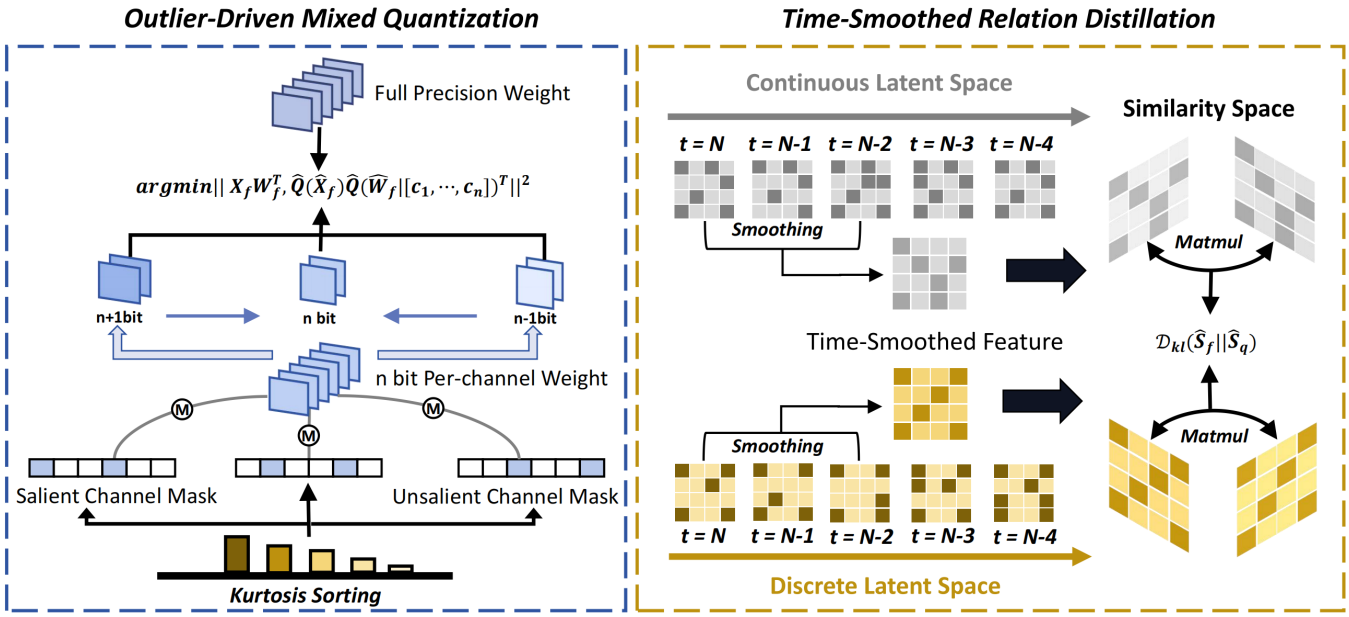


Figure 2: Overview of proposed MPQ-DM, consisting of Outlier-Driven Mixed Quantization to apply intra-layer mixed quantization and Time-Smoothed Relation Distillation to improve optimization robustness. (M) stands for mask operation.

quantized diffusion models, to achieve a calibration time under PTQ-level but accurate quantization performance.

Despite exploring the QLORA-FT scheme in diffusion models, it still experienced significant performance degradation in extremely low bit quantization. The performance degradation mainly comes from the high discretization of activation values. We analyzed the challenges from two different aspects. From the representation perspective, we found significant outliers in some weight channels of the diffusion model, and the presence of outliers leads to a large number of outliers being occupied by stages or a small number of outliers in the bit width. Highly discretized activation values are extremely quantization-unfriendly for channels with severe outliers. This results in severe loss of weight information expression after quantization. Existing unified bit width quantization (He et al. 2023; Wang et al. 2024b) or layer-wise mixed precision quantization (Dong et al. 2019; He et al. 2024; Dong et al. 2020) cannot solve outliers in target weight channel within layers. From the optimization perspective, highly discretized intermediate representation of quantization model resulting in not robust expression of features (Martinez et al. 2020). The multi-step continuous denoising of the diffusion model leads to the accumulation of such errors. Also, the difference in latent space between the discretized features and the fully precision features may lead to negative optimization if we directly align two representations.

To address the aforementioned issues, we propose Mixed Precision Quantization for extremely low bit Diffusion Models (MPQ-DM) consisting of Outlier Driven Mixed Quantization (OMD) and Time Smoothed Relation Distillation (TRD). The overview of MPQ-DM is in Fig. 2. For weight quantization, we use outlier-aware scale to numerically mitigate the outlier degree. Then we use $Kurtosis$ to quantify the presence of outliers and assign higher quantization bits to channels with salient outliers, further ensuring its quantization performance. For model optimization, we select multiple consecutive intermediate representations as smoothed distillation targets to alleviate the problem of abnormal activation values in model optimization. To address the issue of numerical alignment mismatch between discrete and continuous latent spaces, we transfer the numerical alignment of the two latent spaces to a unified similarity space for knowledge distillation, ensuring consistency in feature expression. The contributions of our work are summarized as:

- We identified salient outlier phenomena in different model weight channels as major bottlenecks for extremely low bit quantization. We propose Outlier-Driven Mixed Quantization, which utilizes smooth factor to alleviate outlier phenomenon numerically and dynamically allocates quantization bits of different channels within target bit-width.
- We identified extremely discretized features under extremely low bit quantization suffering from numerical unrobustness. We propose Time-Smoothed Relation Distillation to utilize the features of N time steps as a smoothed distillation objective and transfer two latent spaces with large numerical differences to a unified similarity space for relation distillation.
- We push the limit of efficient diffusion quantization to extremely low bit-widths (2-4 bit). Extensive experiments on generation benchmarks demonstrate that our MPQ-DM surpasses both the baseline and current SOTA PTQ-based diffusion quantization methods by a significant margin.

2 Related Work

2.1 Diffusion Model

Diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) perform a forward sampling process by gradually adding noise to the data distribution $\mathbf{x}_0 \sim q(x)$. In DDPM, the forward noise addition process of the diffusion model is a Markov chain, taking the form:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}),$$
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $\alpha_t = 1 - \beta_t$, β_t is time-related schedule. Diffusion models generate high-quality images by applying a denoising process to randomly sampled Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, taking the form:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \hat{\mu}_{\theta,t}(\mathbf{x_t}), \hat{\beta}_t\mathbf{I}), \tag{2}$$

where $\hat{\mu}_{\theta,t}$ and $\hat{\beta}_t$ are outputed by the diffusion model.

2.2 Diffusion Quantization

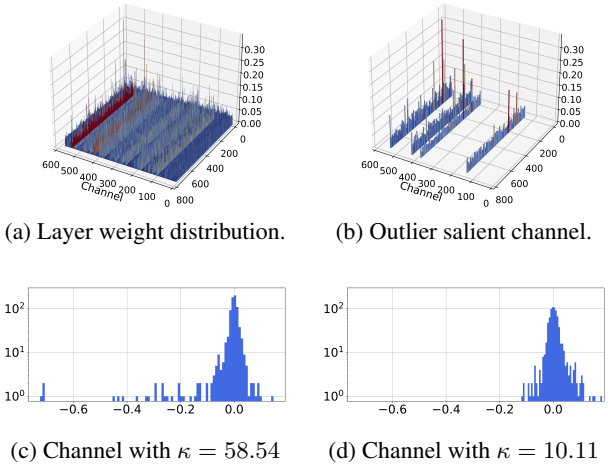
Post-training quantization (PTQ) and Quantization-aware training (QAT) are two main approaches for model quantization. The commonly used QAT methods like LSQ (Esser et al. 2019) and methods for diffusion models Q-dm (Li et al. 2024) and Binarydm (Zheng et al. 2024) ensure the model performance at extremely low bit-width or even binary quantization, but they require a lot of extra training time compared with PTQ methods, resulting larger training burden. PTQ methods for diffusion model PTQ4DM (Shang et al. 2023) and Q-Diffusion (Li et al. 2023) have made initial exploration. The following works PTQ-D (He et al. 2024), TFMQ-DM (Huang et al. 2024), APQ-DM (Wang et al. 2024a) and QuEST (Wang et al. 2024b) have made improvements in the direction of quantization error, temporal feature, calibration data, and calibration module. The performance of diffusion model after quantization is further improved. However, the performance of PTQ-based methods suffers from severe degradation at extremely low bit-width. To combine the advantages of QAT and reduce the required training time, EfficientDM (He et al. 2023) uses LoRA (Hu et al. 2021) method to fine-tune the quantized diffusion model. However, neither of these efficient quantization methods can guarantee the performance of the diffusion model under low bit. Therefore, this paper focuses on maximizing the extremely low bit quantization diffusion models performance.

3 Method

3.1 Model Quantization

Model quantization maps model weights and activations to low bit integer values to reduce memory footprint and accelerate the inference. For a floating vector $\mathbf{x}_f$, the quantization process can be formulated as

$$\hat{\mathbf{x}}_q = Q(\mathbf{x}_f, s, z) = clip(\lfloor \frac{\mathbf{x}_f}{s} \rceil + z, 0, 2^N - 1),$$
$$s = \frac{u - l}{2^N - 1}, z = -\lfloor \frac{l}{s} \rceil, \tag{3}$$



(a) Layer weight distribution. (b) Outlier salient channel.

(c) Channel with $\kappa = 58.54$ (d) Channel with $\kappa = 10.11$

Figure 3: Weight distribution of LDM-4 ImageNet 256x256.

where $\hat{\mathbf{x}}_q$ indicates quantized vector in integer, $\lfloor \cdot \rceil$ is round fuction and $clip(\cdot)$ is function that clamps values into the range of $[0, 2^N - 1]$, s is a scale factor and z is a quantization zero point. l and u are the lower and upper bounds of quantization thresholds respectively. They are determined by $\mathbf{x}_f$ and the target bit-width. Reversely, in order to restore the low bit integer quantization vector $\hat{\mathbf{x}}_q$ to the full precision representation, the dequantization process is formulated as

$$\hat{\mathbf{x}}_f = \hat{Q}(\mathbf{x}_f) = (\hat{\mathbf{x}}_q - z)s, \tag{4}$$

where $\hat{\mathbf{x}}_f$ is the dequantized vector used for forward process.

3.2 Outlier-Driven Mixed Quantization

Studies on diffusion (Wu et al. 2024; Zhao et al. 2024) model find that there are some channels with significant weight values, and these channels are also crucial to the quantization error and model performance. We visualized one layer weight distribution of diffusion model in Fig. 3a and its corresponding outlier salient channel distribution in Fig. 3b. It can be found that not only the weights of different channels have a large gap in the numerical range, but some weight channels also have severe outliers.

In Fig. 3c and Fig. 3d, we visualized two channel weight distribution. The overall weight distribution can be seen as a normal distribution plus an outlier distribution. For model quantization, direct quantization of a normal distribution is well studied (Liu et al. 2020; Qin et al. 2023), but the presence of a small number of outliers can cause some outliers to be clamped or occupy a portion of bit width, making it difficult to quantize the main normal distribution, resulting in quantization errors. This error is greatly amplified at extremely low bit-width (2-4 bit), so we naturally hope to allocate more bit widths to these channels to ensure their performance. However, this outlier phenomenon in specific channels may occur at various layers of the model. Therefore, using traditional methods to allocate different bits to different model layers for mixed precision quantization is clearly

not a good choice. But because the quantization of weights is channel-wise, we can redistribute the quantization bit width between different channels within layers to solve this problem, that is, intra-layer mixed precision quantization.

In order to determine the specific bit width allocation method, we need to quantify the significance of the outliers. $Kurtosis\ \kappa$ can be used to quantify the "tailedness" of a real-valued variable's probability distribution (DeCarlo 1997; Liu et al. 2024b). This aligns naturally with our goal, as outliers add long tails of anomalies to a normal distribution as we mentioned above. As we show in Fig. 3c and Fig. 3d, channels with more salient outlier phenomenon have higher κ values compared to a normal distribution without outliers. Thus, we use $Kurtosis\ \kappa$ as an indicator of the difficulty of quantizing different weight channels to find an optimal bit allocation method for each layer weight $\mathbf{W}_f$.

We hope that the model output to be as consistent as possible with the full precision model after mixed bit width assignment quantization. Since our goal is model output $\mathbf{Y} = \mathbf{X}\mathbf{W}^\top$, and the outliers are the maximum or minimum values in the weights, we can use the property of matrix multiplication to reduce the salient degree of outliers without loss before quantization as follows:

$$\begin{aligned} \delta_i &= \sqrt{\frac{\max(|\mathbf{W}_i|)}{\max(|\mathbf{X}_i|)}}, \\ \mathbf{Y} &= (\mathbf{X}\mathrm{diag}(\delta)) \cdot (\mathrm{diag}(\delta)^{-1}\mathbf{W}^\top) = \hat{\mathbf{X}} \cdot \hat{\mathbf{W}}^\top, \end{aligned} \tag{5}$$

where $\mathbf{Y}$, $\mathbf{X}$, and $\mathbf{W}$ represent the output activation, input activation, and model weights. δ is a channel-wise smooth factor that balances the quantization difficulty of weight and activation by scaling outliers closer to the normal distribution. After pre-scaled, the outlier salient channels are more smooth to be quantized. Using scaled weight $\hat{\mathbf{W}}$, we compute κ for each weight channel and rank them accordingly. Then, we use the following optimization formula to determine the outlier salient and unsalient channels as follows:

$$\begin{aligned} &\arg\min_{c_1,\cdots,c_g} ||\mathbf{X}_f\mathbf{W}_f^\top, \hat{Q}(\hat{\mathbf{X}}_f)\hat{Q}(\hat{\mathbf{W}}_f|[c_1,\cdots,c_n])^\top||^2, \\ &C_{N-1} = \{c_i|c_i = N-1\},\ C_{N+1} = \{c_i|c_i = N+1\}, \\ &|C_{N-1}| = |C_{N+1}|,\ |C_{N-1}| + |C_N| + |C_{N+1}| = n, \end{aligned} \tag{6}$$

where N is the target average bit-width, n is the channel number in weight, $\hat{Q}$ represents the quantization and de-quantization process. We apply channel-wise mixed bit-width quantization for weight as $\hat{Q}(\hat{\mathbf{W}}_f|[c_1,\cdots,c_n])$, where c_i denotes the bit-width for the i_{th} channel. C represents the set of channel shares the same bit-width and $|C|$ stands for the number of channels in set C. For example, for 3-bit quantization, we assign some of the outlier salient channels to 4-bit and reassign the same number of the outlier unsalient channels to 2-bit. In this way, the average total bit-width for the layer weight is still 3-bit without adding any extra parameters.

To accelerate the optimization process, we set k channels as a search group and the search region for outlier salient channel is constrained in $[0, \frac{c_{out}//k}{2}]$. We empirically set k as $\frac{c_{out}}{10}$, thus the search time is 5 times which is efficient enough for layer bit allocation.

3.3 Time-Smoothed Relation Distillation

For diffusion optimization, existing methods like EfficientDM (He et al. 2023) optimize the training parameters of the quantization model by aligning the output of the full-precision (FP) model and the quantization model

$$\mathcal{L}_{task} = ||\theta_f(\mathbf{x}_t, t) - \theta_q(\mathbf{x}_t, t)||^2, \tag{7}$$

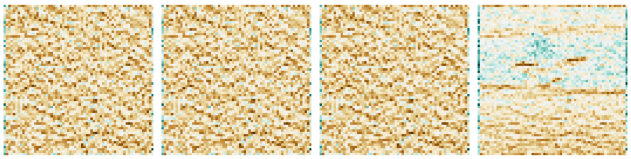
where θ_f and θ_q denotes the FP model and the quantization model respectively, $\mathbf{x}_t$ is obtained by denoising Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with FP model iteratively for $T-t$ steps.

In extremely low bit (2-4 bit) quantization, it is not sufficient to align only the final output of the model because the expressive ability of the quantization model is severely insufficient. We usually set the final projection layer of the quantization model to 8-bit and the layers before the project layer to target low bit (Wang et al. 2024b; He et al. 2023; Huang et al. 2024). In this way, we do not directly perceive the part where the quantization information is severely missing. Therefore, we can give more fine-grained guidance to the extremely low bit quantization model by distilling the model feature layer before the last project layer

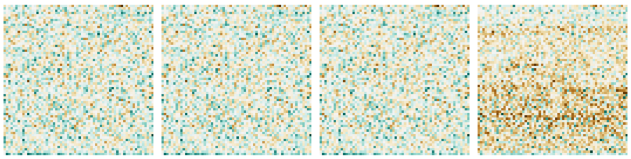
$$\mathcal{L}_{dis} = \mathcal{D}(\mathbf{F}_f, \mathbf{F}_q), \tag{8}$$

where $\mathbf{F}_f$ and $\mathbf{F}_q$ denote the feature map before the last project layer of FP model and quantization model respectively. $\mathcal{D}$ is a metric to measure the distance of the two feature maps.

However, features at extremely low bit quantization present high discretization (e.g., only 16 values for 4-bit activation quantization) (Martinez et al. 2020). This discretized feature shows a high degree of unrobustness in numerical value, and directly distilling it has poor effect or even impairs the normal training of the quantization model (Zheng et al. 2024; Qin et al. 2022). Moreover, due to the unique iterative denoising process of the diffusion model (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020), the outliers of the features will continue to add up with denoising, further amplifying this unrobustness. Typically we can use additional projection heads to map them to a uniform space or to regularize it numerically (Yang et al. 2022, 2023; Feng et al. 2024; Yang et al. 2024a). However, projection heads will bring additional training parameters, and performing regularization cannot solve the problem of error accumulation in iterative calculation. We find that the features of the diffusion model are highly correlated with time steps and show similarity in some time steps. In Fig. 4, we find that the features on consecutive time steps are highly similar, while the time steps far apart are quite different. The similarity of features indicates that the difference of consecutive time steps in denoising trajectory is quite small, so we can alleviate the feature unrobustness at different time steps by fusing the intermediate features of multiple consecutive steps. Instead of forcing the highly discretized quantization model to strictly learn the denoising trajectory of each step

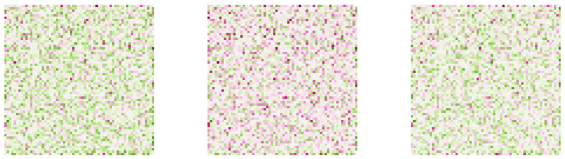


(a) $T = 20$

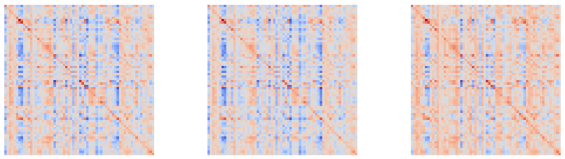


(b) $T = 15$

Figure 4: Visualization of feature map from timesetp (Left) T, (Left-Mid) $T-1$, (Right-Mid) $T-2$, (Right) $T-10$.



(a) Original feature map



(b) Cosine similarity map

Figure 5: Visualization of different activation maps of (Left) FP model, (Mid) well-trained quantization model, (Right) un-trained quantization model.

of FP model, but to learn the denoising trajectory of successive multiple steps.

Therefore, we use the intermediate features of N consecutive steps as the smoothed feature representation for distillation. For T timestep optimization, we rewrite the distillation formula as

$$\hat{\mathbf{F}}_f = \sum_{t=0}^{N} \mathbf{F}_{T-t,f},\ \hat{\mathbf{F}}_q = \sum_{t=0}^{N} \mathbf{F}_{T-t,q}, \\ \mathcal{L}_{dis} = \mathcal{D}(\hat{\mathbf{F}}_f || \hat{\mathbf{F}}_q), \tag{9}$$

where $\mathbf{F}_{T-t,f}$ and $\mathbf{F}_{T-t,q}$ denotes the last feature map in time step $T-t$ from FP model and quantization model respectively.

In Fig. 5a, we visualized feature maps between FP model, well-trained quantization model, and un-trained quantization model. Although time-smoothed feature improves the robustness of quantized feature, there is still a mismatch in numerical expression between well-trained model and FP model. This is blamed on the difference between the discrete latent space of quantized features and continuous latent space of FP features. Thus, any metrics such as L2 loss that numerically align $\hat{\mathbf{F}}_f$ and $\hat{\mathbf{F}}_q$ cannot avoid this difference between spaces. Therefore, we propose to use relation distillation to replace the strict numerical alignment by learning the feature similarity relation between $\hat{\mathbf{F}}_f$ and $\hat{\mathbf{F}}_q$. In Fig 5b, we transfer the numerical relationship between discrete latent space and continuous latent space to the feature similarity relationship inside each space which unifies the distillation goal into the feature similarity space. This successfully solved the numerical mismatch. Formally speaking, for $\hat{\mathbf{F}} \in \mathbb{R}^{h \times w \times c}$ and to simplify writing, we reshape it as $\hat{\mathbf{F}} \in \mathbb{R}^{s \times c}$, where $s = h \times w$. For i_{th} feature representation $\hat{\mathbf{F}}^i$, we can calculate its cosine similarity distribution with each position feature representation $\hat{\mathbf{S}}^i = \hat{\mathbf{F}}^i \hat{\mathbf{F}}^\top \in \mathbb{R}^s$. Thus, the relation distillation metric is

$$\mathcal{L}_{dis} = \sum_{i=1}^{s} \mathcal{D}_{kl}(\hat{\mathbf{S}}_f^i || \hat{\mathbf{S}}_q^i), \tag{10}$$

where $\hat{\mathbf{S}}_f^i$ and $\hat{\mathbf{S}}_q^i$ denotes time-smoothed feature similarity distribution from FP model and quantization model respectively. $\mathcal{D}_{kl}(\cdot||\cdot)$ stands for the Kullback-Leibler (KL) divergence between two distributions. We utilize KL divergence instead of L2 Loss here. Because L2 Loss can only perceive a single representation but KL divergence can perceive the information of the whole feature map. The overall optimization target is

$$\mathcal{L}_{total} = \mathcal{L}_{target} + \lambda \mathcal{L}_{dis}. \tag{11}$$

4 Experiment

4.1 Experiment Settings

We conduct experiments on commonly used datasets LSUN-Bedrooms 256×256, LSUN-Churches 256×256 (Yu et al. 2015), and ImageNet 256×256 (Deng et al. 2009) for both unconditional and conditional image generation tasks on LDM models. We also conduct text-to-image generation task on Stable Diffusion (Rombach et al. 2022). We use IS (Salimans et al. 2016), FID (Heusel et al. 2017), sFID (Nash et al. 2021) and Precision to evaluate LDM performance. For Stable Diffusion, we use CLIP Score (Hessel et al. 2021) for evaluation. To cope with the extreme expressivity degradation under 2bit quantization, we allocate an additional 10% number of channels for 2bit during the search process of OMQ, named MPQ-DM$^+$. This results in only a 0.6% increase in model size compared with FP model. We compare our MPQ-DM with baseline EfficientDM (He et al. 2023) and layer-wise mixed precision HAWQ-v3 (Yao et al. 2021) and other PTQ-based methods (He et al. 2024; Huang et al. 2024; Wang et al. 2024b) which possess similar time consumption. Details can be found in Appendix.

4.2 Experiment Results

Class-conditional Generation. We conduct conditional generation experiment on ImageNet 256×256 dataset, focusing on LDM-4. Results in Table 1 show that MPQ-DM greatly outperforms existing methods on all bit set-

Method	Bit (W/A)	Size (MB)	IS ↑	FID ↓	sFID ↓	Precision ↑ (%)
FP	32/32	1529.7	364.73	11.28	7.70	93.66
PTQ-D	3/6	144.5	162.90	17.98	57.31	63.13
TFMQ	3/6	144.5	174.31	15.90	40.63	67.42
QuEST	3/6	144.6	194.32	14.32	31.87	72.80
EfficientDM	3/6	144.6	299.63	7.23	8.18	86.27
HAWQ-V3	3/6	144.6	303.79	6.94	8.01	87.76
MPQ-DM	3/6	144.6	**306.33**	**6.67**	**7.93**	**88.65**
PTQ-D	3/4	144.5	10.86	286.57	273.16	0.02
TFMQ	3/4	144.5	13.08	223.51	256.32	0.04
QuEST	3/4	175.5	15.22	202.44	253.64	0.04
EfficientDM	3/4	144.6	134.30	11.02	9.52	70.52
HAWQ-V3	3/4	144.6	152.61	8.49	9.26	75.02
MPQ-DM	3/4	144.6	**197.43**	**6.72**	**9.02**	**81.26**
PTQ-D	2/6	96.7	70.43	40.29	35.70	43.79
TFMQ	2/6	96.7	77.26	36.22	33.05	45.88
QuEST	2/6	96.8	86.83	32.37	31.58	47.74
EfficientDM	2/6	96.8	69.64	29.15	12.94	54.70
HAWQ-V3	2/6	96.8	88.25	22.73	11.68	57.04
MPQ-DM	2/6	96.8	102.51	15.89	10.54	67.74
MPQ-DM$^+$	2/6	101.6	**136.35**	**11.00**	**9.41**	**72.84**
PTQ-D	2/4	96.7	9.25	336.57	288.42	0.01
TFMQ	2/4	96.7	12.76	300.03	272.64	0.03
QuEST	2/4	127.7	14.09	285.42	270.12	0.03
EfficientDM	2/4	96.8	25.20	64.45	14.99	36.63
HAWQ-V3	2/4	96.8	33.21	52.63	14.00	42.95
MPQ-DM	2/4	96.8	43.95	36.59	12.20	52.14
MPQ-DM$^+$	2/4	101.6	**60.55**	**27.11**	**11.47**	**57.84**

Table 1: Performance comparisons of fully-quantized LDM-4 models on ImageNet 256×256. Best results are in bold and second best are in underlined.

tings. MPQ-DM generally performs better than the layer-wise approach HAWQ-v3, demonstrating the necessity of mixed precision quantization within layers. MPQ-DM W3A4 model even surpasses FP model on FID. In W2A4 setting, PTQ-based methods fail to generate images, while EfficientDM performs poorly. MPQ-DM greatly improves baseline with a notable 27.86 decrease in FID. MPQ-DM$^+$ even further leads to 9.48 decrease in FID using only 4.8 MB additional model size.

Unconditional Generation. We conduct unconditional generation experiment on LSUN-Bedrooms dataset over LDM-4 and LSUN-Churches dataset over LDM-8 with 256×256 resolution. In Table 2 and Table 3, MPQ-DM still outperforms all other existing methods under all bit settings. For LSUN-Bedrooms dataset, we achieved FID decrease of 5.59 on W3A4, 8.53 on W2A6, and even 12.81 on W2A4 setting compared with baseline. Under W2A4 setting, we are the first method pushing sFID under 20 which leads to 6.12 decrease compared with baseline.

Method	Bit (W/A)	Size (MB)	FID ↓	sFID ↓	Precision ↑ (%)
FP	32/32	1045.4	7.39	12.18	52.04
PTQ-D	3/6	98.3	113.42	43.85	10.06
TFMQ	3/6	98.3	26.42	30.87	38.29
QuEST	3/6	98.4	21.03	28.75	40.32
EfficientDM	3/6	98.4	13.37	16.14	44.55
MPQ-DM	3/6	98.4	**11.58**	**15.44**	**47.13**
PTQ-D	3/4	98.3	100.07	50.29	11.64
TFMQ	3/4	98.3	25.74	35.18	32.20
QuEST	3/4	110.4	19.08	32.75	40.64
EfficientDM	3/4	98.4	20.39	20.65	38.70
MPQ-DM	3/4	98.4	**14.80**	**16.72**	**43.61**
PTQ-D	2/6	65.7	86.65	53.52	10.27
TFMQ	2/6	65.7	28.72	29.02	34.57
QuEST	2/6	65.7	29.64	29.73	34.55
EfficientDM	2/6	65.7	25.07	22.17	34.59
MPQ-DM	2/6	65.7	17.12	19.06	40.90
MPQ-DM$^+$	2/6	68.9	**16.54**	**18.36**	**41.80**
PTQ-D	2/4	65.7	147.25	49.97	9.26
TFMQ	2/4	65.7	25.77	36.74	32.86
QuEST	2/4	77.7	24.92	36.33	32.82
EfficientDM	2/4	65.7	33.09	25.54	28.42
MPQ-DM	2/4	65.7	21.69	21.58	38.69
MPQ-DM$^+$	2/4	68.9	**20.28**	**19.42**	**38.92**

Table 2: Unconditional image generation results of LDM-4 models on LSUN-Bedrooms 256×256.

Text-to-image Generation. We conduct text-to-image generation experiment on randomly selected 10k COCO2014 validation set prompts over Stable Diffusion v1.4 model with 512×512 resolution. In Table 4, our method achieves better performance over baseline and SOTA PTQ methods. In W3A4 and W2A6 settings, we achieve over 0.3 CLIP Score improvement. MPQ-DM$^+$ even further achieves 1.79 improvement in CLIP Score with only 10.2 MB additional model size.

4.3 Ablation Study

Component Study. In Table 5, we perform comprehensive ablation studies on LDM-4 ImageNet 256×256 model to evaluate the effectiveness of each proposed component. Our proposed OMQ solves the existing layer-wise bit allocation methods from the perspective of quantization, allocating more bit width to the channels with salient outlier phenomenon within layer while the total average bit width is unchanged. This intra-layer mixed-precision quantization method greatly improves the performance of baseline, gaining IS increases of 58.88. In addition, TSD improves the robustness in the distillation process from the perspective of model optimization, and also achieves a certain improvement. Through the parallel improvement of the two perspectives of quantization and optimization, MPQ-DM achieves state-of-the-art quantization performance.

Outlier Selection Method Study. In Table 6, we study different outlier salient channel selection methods in mixed

Method	Bit (W/A)	Size (MB)	FID ↓	sFID ↓	Precision ↑
FP	32/32	1125.2	5.55	10.75	67.43
PTQ-D	3/6	106.0	59.43	40.26	13.37
TFMQ	3/6	106.0	13.53	22.10	62.74
QuEST	3/6	106.1	22.19	32.79	60.73
EfficientDM	3/6	106.1	9.53	13.70	62.92
MPQ-DM	3/6	106.1	**9.28**	**13.37**	**63.73**
PTQ-D	3/4	106.0	77.08	49.63	10.25
TFMQ	3/4	106.0	35.51	48.59	55.32
QuEST	3/4	122.4	40.74	53.63	52.78
EfficientDM	3/4	106.1	15.59	18.16	57.92
MPQ-DM	3/4	106.1	**14.08**	**16.91**	**59.68**
PTQ-D	2/6	70.9	63.38	46.63	12.14
TFMQ	2/6	70.9	25.51	35.83	54.75
QuEST	2/6	70.9	23.03	35.13	56.90
EfficientDM	2/6	70.9	16.98	18.18	57.39
MPQ-DM	2/6	70.9	15.61	17.44	59.03
MPQ-DM$^+$	2/6	74.4	**13.38**	**15.59**	**61.00**
PTQ-D	2/4	70.9	81.95	50.66	9.47
TFMQ	2/4	70.9	51.44	64.07	42.25
QuEST	2/4	86.9	50.53	63.33	45.86
EfficientDM	2/4	70.9	22.74	22.55	53.00
MPQ-DM	2/4	70.9	21.83	21.38	53.99
MPQ-DM$^+$	2/4	74.4	**16.91**	**18.57**	**58.04**

Table 3: Unconditional image generation results of LDM-8 models on LSUN-Churches 256×256.

Method	Bit (W/A)	Size (MB)	CLIP Score ↑
FP	32/32	3279.1	31.25
QuEST	3/4	332.9	26.55
EfficientDM	3/4	309.8	26.63
MPQ-DM	3/4	309.8	**26.96**
QuEST	2/6	207.4	22.88
EfficientDM	2/6	207.4	22.94
MPQ-DM	2/6	207.4	23.23
MPQ-DM$^+$	2/6	217.6	**25.02**

Table 4: Text-to-image generation results (512×512) of Stable Diffusion v1.4 using 10k COCO2014 validation set prompts.

precision quantization on LDM-4 ImageNet 256×256 model. We find that even randomly selecting some channels to higher or lower bits leads to a certain performance gain. This indicates that in extremely low bit quantization, the gain brought by increasing the bit of some channels is far greater than the impact brought by decreasing some channels, which proves the necessity of mixed quantization. While there are some gains in selecting channels randomly or from the head and tail of weights, our outlier selection method based on $Kurtosis$ achieves the most significant performance improvement. This shows that $Kurtosis$ selects the outlier salient channels with the most significant performance improvement, which proves the effectiveness of our $Kurtosis$-based channel selection method.

Method	Bit (W/A)	IS ↑	FID ↓	sFID ↓	Precision ↑
PTQD	3/4	10.86	286.57	237.16	0.05
Baseline	3/4	134.30	11.02	9.52	70.52
+OMQ	3/4	193.18	6.91	9.12	80.77
+TSD	3/4	135.91	10.38	9.38	72.21
MPQ-DM	3/4	**197.43**	**6.72**	**9.02**	**81.26**

Table 5: Ablation study on proposed methods.

Method	Bit (W/A)	FID ↓	sFID ↓	Precision ↑
Baseline	3/4	11.08	22.02	75.86
Random	3/4	9.69	22.23	79.63
Head-tail	3/4	9.44	21.71	79.95
$Kurtosis$ κ	3/4	**9.05**	**21.51**	**80.60**

Table 6: Ablation study on outlier selection function. We sample 10k samples for evaluation.

Distillation Metrics Study. In Table 7, we study different distillation metrics used in Eq. 9 on LDM-4 ImageNet model. We compare with without distillation to valid different metrics. Using L2 Loss to align $\hat{\mathbf{F}}_f$ and $\hat{\mathbf{F}}_q$ only shows little improvement on sFID, but decreases FID and Precision. This indicates that the discrete features of and continuous features cannot be well aligned by numerical values directly, which leads to negative optimization. However, our proposed relation distillation can transfer all features into a unified similarity space. This breaks the difference between the two latent spaces and improves model performance.

Method	Bit (W/A)	FID ↓	sFID ↓	Precision ↑
w/o Distillation	3/4	9.12	21.59	81.15
L2 Loss	3/4	9.13	21.47	80.76
Relation Distillation	3/4	**9.10**	**21.40**	**81.25**

Table 7: Ablation study on distillation metrics. We sample 10k samples for evaluation.

5 Conclusion

In this paper, we have proposed MPQ-DM, a Mixed-Precision Quantization method for extremely low bit diffusion quantization. To address severe model quantization error caused by outlier salient weight channel, we have proposed Outlier-Driven Mixed Quantization to apply optimized intra-layer mixed-precision bit-width allocation that automatically resolves performance degradation introduced by the outlier. To robustly learn representations across time steps, we have constructed a Time-Smoothed Relation Distillation scheme to obtain feature representations that are more suitable for quantitative model learning. Our extensive experiments have demonstrated the superiority of MPQ-DM over baseline and other previous PTQ-based methods.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No.62476264 and No.62406312), China National Postdoctoral Program for Innovative Talents (No.BX20240385) funded by China Postdoctoral Science Foundation, Beijing Natural Science Foundation (No.4244098), Science Foundation of the Chinese Academy of Sciences, Swiss National Science Foundation (SNSF) project 200021E_219943 Neuromorphic Attention Models for Event Data (NAMED), Baidu Scholarship, and Beijing Natural Science Foundation (L242021).

References

Chen, Q.; Diao, B.; Yang, Y.; and Xu, Y. 2024. SCP: A Structure Combination Pruning Method via Structured Sparse for Deep Convolutional Neural Networks. In *International Conference on Pattern Recognition*, 238–253. Springer.

Chitty-Venkata, K. T.; Mittal, S.; Emani, M.; Vishwanath, V.; and Somani, A. K. 2023. A survey of techniques for optimizing transformer inference. *Journal of Systems Architecture*, 102990.

Dai, L.; Gong, L.; An, Z.; Xu, Y.; and Diao, B. 2024. Sketchfusion: A gradient compression method with multi-layer fusion for communication-efficient distributed training. *Journal of Parallel and Distributed Computing*, 185: 104811.

DeCarlo, L. T. 1997. On the meaning and use of kurtosis. *Psychological methods*, 2(3): 292.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Ding, Y.; Feng, W.; Chen, C.; Guo, J.; and Liu, X. 2024. Reg-PTQ: Regression-specialized Post-training Quantization for Fully Quantized Object Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16174–16184.

Dong, Z.; Yao, Z.; Arfeen, D.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2020. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33: 18518–18529.

Dong, Z.; Yao, Z.; Gholami, A.; Mahoney, M. W.; and Keutzer, K. 2019. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, 293–302.

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2019. Learned step size quantization. *arXiv preprint arXiv:1902.08153*.

Feng, W.; Yang, C.; An, Z.; Huang, L.; Diao, B.; Wang, F.; and Xu, Y. 2024. Relational diffusion distillation for efficient image generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 205–213.

Gholami, A.; Kim, S.; Dong, Z.; Yao, Z.; Mahoney, M. W.; and Keutzer, K. 2022. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, 291–326. Chapman and Hall/CRC.

He, Y.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2023. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. *arXiv preprint arXiv:2310.03270*.

He, Y.; Liu, L.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2024. Ptqd: Accurate post-training quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

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. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Hubara, I.; Nahshan, Y.; Hanani, Y.; Banner, R.; and Soudry, D. 2020. Improving post training neural quantization: Layer-wise calibration and integer programming. *arXiv preprint arXiv:2006.10518*.

Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.; Adam, H.; and Kalenichenko, D. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2704–2713.

Krishnamoorthi, R. 1806. Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv 2018. *arXiv preprint arXiv:1806.08342*.

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

Li, Y.; Xu, S.; Cao, X.; Sun, X.; and Zhang, B. 2024. Q-dm: An efficient low-bit quantized diffusion model. *Advances in Neural Information Processing Systems*, 36.

Liu, H.; Diao, B.; Chen, W.; and Xu, Y. 2024a. A resource-aware workload scheduling method for unbalanced GEMMs on GPUs. *The Computer Journal*, bxae110.

Liu, J.; Niu, L.; Yuan, Z.; Yang, D.; Wang, X.; and Liu, W. 2023a. Pd-quant: Post-training quantization based on prediction difference metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24427–24437.

Liu, Z.; Shen, Z.; Savvides, M.; and Cheng, K.-T. 2020. Reactnet: Towards precise binary neural network with generalized activation functions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 143–159. Springer.

Liu, Z.; Zhang, F.; He, J.; Wang, J.; Wang, Z.; and Cheng, L. 2023b. Text-guided mask-free local image retouching. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2783–2788. IEEE.

Liu, Z.; Zhao, C.; Fedorov, I.; Soran, B.; Choudhary, D.; Krishnamoorthi, R.; Chandra, V.; Tian, Y.; and Blankevoort, T. 2024b. SpinQuant–LLM quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

Martinez, B.; Yang, J.; Bulat, A.; and Tzimiropoulos, G. 2020. Training binary neural networks with real-to-binary convolutions. *arXiv preprint arXiv:2003.11535*.

Mei, K.; and Patel, V. 2023. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9117–9125.

Nash, C.; Menick, J.; Dieleman, S.; and Battaglia, P. W. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.

Pilipović, R.; Bulić, P.; and Risojević, V. 2018. Compression of convolutional neural networks: A short survey. In *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–6. IEEE.

Qin, H.; Ding, Y.; Zhang, M.; Yan, Q.; Liu, A.; Dang, Q.; Liu, Z.; and Liu, X. 2022. Bibert: Accurate fully binarized bert. *arXiv preprint arXiv:2203.06390*.

Qin, H.; Gong, R.; Liu, X.; Shen, M.; Wei, Z.; Yu, F.; and Song, J. 2020. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2250–2259.

Qin, H.; Zhang, M.; Ding, Y.; Li, A.; Cai, Z.; Liu, Z.; Yu, F.; and Liu, X. 2023. Bibench: Benchmarking and analyzing network binarization. In *International Conference on Machine Learning*, 28351–28388. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

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. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

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, H.; Shang, Y.; Yuan, Z.; Wu, J.; and Yan, Y. 2024b. Quest: Low-bit diffusion model quantization via efficient selective finetuning. *arXiv preprint arXiv:2402.03666*.

Wang, P.; Chen, Q.; He, X.; and Cheng, J. 2020. Towards accurate post-training network quantization via bit-split and stitching. In *International Conference on Machine Learning*, 9847–9856. PMLR.

Wei, X.; Gong, R.; Li, Y.; Liu, X.; and Yu, F. 2022. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*.

Wu, J.; Wang, H.; Shang, Y.; Shah, M.; and Yan, Y. 2024. PTQ4DiT: Post-training Quantization for Diffusion Transformers. *arXiv preprint arXiv:2405.16005*.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024a. CLIP-KD: An Empirical Study of CLIP Model Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, C.; An, Z.; Zhou, H.; Zhuang, F.; Xu, Y.; and Zhang, Q. 2023. Online knowledge distillation via mutual contrastive learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10212–10227.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12319–12328.

Yang, Y.; Cheng, D.; Fang, C.; Wang, Y.; Jiao, C.; Cheng, L.; and Wang, N. 2024b. Diffusion-based Layer-wise Semantic Reconstruction for Unsupervised Out-of-Distribution Detection.

Yao, Z.; Dong, Z.; Zheng, Z.; Gholami, A.; Yu, J.; Tan, E.; Wang, L.; Huang, Q.; Wang, Y.; Mahoney, M.; et al. 2021. Hawq-v3: Dyadic neural network quantization. In *International Conference on Machine Learning*, 11875–11886. PMLR.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; 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*.

Zhao, T.; Fang, T.; Liu, E.; Rui, W.; Soedarmadji, W.; Li, S.; Lin, Z.; Dai, G.; Yan, S.; Yang, H.; et al. 2024. ViDiT-Q: Efficient and Accurate Quantization of Diffusion Transformers for Image and Video Generation. *arXiv preprint arXiv:2406.02540*.

Zheng, X.; Qin, H.; Ma, X.; Zhang, M.; Hao, H.; Wang, J.; Zhao, Z.; Guo, J.; and Liu, X. 2024. Binarydm: Towards accurate binarization of diffusion model. *arXiv preprint arXiv:2404.05662*.