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

Pioneering 4-Bit FP Quantization for Diffusion Models: Mixup-Sign Quantization and Timestep-Aware Fine-Tuning

Maosen Zhao[1*], Pengtao Chen[1*], Chong Yu[2], Yan Wen[1], Xudong Tan[1], Tao Chen[1†]
[1] School of Information Science and Technology, Fudan University
[2] Academy for Engineering and Technology, Fudan University
20307130202@fudan.edu.cn, eetchen@fudan.edu.cn

Abstract

Model quantization reduces the bit-width of weights and activations, improving memory efficiency and inference speed in diffusion models. However, achieving 4-bit quantization remains challenging. Existing methods, primarily based on integer quantization and post-training quantization fine-tuning, struggle with inconsistent performance. Inspired by the success of floating-point (FP) quantization in large language models, we explore low-bit FP quantization for diffusion models and identify key challenges: the failure of signed FP quantization to handle asymmetric activation distributions, the insufficient consideration of temporal complexity in the denoising process during fine-tuning, and the misalignment between fine-tuning loss and quantization error. To address these challenges, we propose the mixup-sign floating-point quantization (MSFP) framework, first introducing unsigned FP quantization in model quantization, along with timestep-aware LoRA (TALoRA) and denoising-factor loss alignment (DFA), which ensure precise and stable fine-tuning. Extensive experiments show that we are the first to achieve superior performance in 4-bit FP quantization for diffusion models, outperforming existing PTQ fine-tuning methods in 4-bit INT quantization.

1. Introduction

Despite the impressive performance of diffusion models (DMs) in image generation [30, 41], their computational and memory demands, particularly for high-resolution outputs, pose significant challenges for deployment on resource-constrained edge devices, highlighting the need for model compression to address these limitations. Model quantization, a key technique within model compression, reduces the bit-width of model weights and activations, typically stored in 32-bit format, to lower precision. This reduction lowers memory usage and accelerates inference speed. By decreasing the bit-width, quantization improves both temporal and memory efficiency in mainstream models, maintaining robust performance, particularly in resource-constrained environments [13, 18, 21].

The current quantization methods for diffusion models can be broadly categorized into two primary approaches. The first, post-training quantization (PTQ), optimizes the quantization parameters after the model has been trained, typically by minimizing the quantization error [19, 32]. While PTQ is effective for 4-bit quantization of weights, it is limited by its reliance on 8-bit quantization for activations, as further reduction in activation bit-width leads to significant performance degradation. In contrast, quantization-aware training (QAT) integrates quantization into the training process, enabling the model to learn with 4-bit precision from scratch [20]. Although QAT can achieve high-performance 4-bit models, it incurs substantial computational overhead, rendering it less practical for many real-world applications [6, 15].

To achieve fully quantized 4-bit diffusion models with minimal overhead, fine-tuning has emerged as a promising solution. This approach leverages pre-trained models and adjusts a small subset of parameters to narrow the performance gap between the quantized model and its full-precision counterpart. While some studies have explored fine-tuning for 4-bit quantization in diffusion models [8, 39], these methods often fail to achieve consistent performance under standard configurations (e.g., quantizing all layers but not some). Consequently, developing a universally effective and scalable fine-tuning method for 4-bit quantization in diffusion models remains an open challenge.

To be noticed, existing quantization methods for diffusion models primarily rely on integer (INT) quantization, which has long been the dominant approach. However, recent developments have demonstrated the considerable potential of floating-point (FP) quantization. Compared to INT quantization, FP quantization offers greater flexibility in modeling complex weight and activation distributions [27], leading to improved performance in visual

†Corresponding author. *Equal contribution.

tasks at 8-bit precision [17, 36], and remarkable results in large language models under 4-bit precision [23, 40]. Moreover, FP quantization provides significant advantages in inference acceleration, with NVIDIA H100 achieving a 1.45× speedup with FP8 quantization, outperforming INT8 [29, 42]. Despite these advantages, the application of low-bit FP quantization in diffusion models remains largely unexplored, presenting a promising avenue for future research.

In summary, to address the challenges in achieving 4-bit diffusion models, we propose a baseline method using a search-based signed FP quantization framework [2, 23] combined with single-LoRA fine-tuning, aiming to realize 4-bit FP quantized diffusion models. Through this exploration, we uncover several findings that are significant for both diffusion model quantization and FP quantization: (1) There exists many layers which exhibit asymmetric distributions, due to the nonlinear activation function $SiLU$. The application of traditional signed FP quantization with a symmetric distribution leads to significant precision loss in the sub-zero area, causing substantial performance degradation after quantization. (2) Fine-tuning is conducted based on the denoising process, which is regarded as a complex task involving the restoration from outlines to details [37]. However, current methods typically apply a single LoRA to fine-tune severely degraded models across all timesteps, which leads to suboptimal learning at certain timesteps. (3) Predicted noise plays a varying role at different timesteps during denoising, which is the key to diffusion models. Arising from the neglect of this variation across timesteps, we identify a mismatch between the impact of quantization and the loss function in current methods, which undermines the effectiveness of fine-tuning.

Facing the challenges in achieving 4-bit FP quantization, we propose constructive strategies: (1) To handle different distribution effectively, we propose a mixup-sign floating-point quantization (MSFP) framework, where unsigned FP quantization with an added zero point leads to a more compatible distribution of discrete points with the anomalous distributions in activations and signed FP quantization continues to exhibit strong representation capacity on other distributions. (2) Realizing that the current single-LoRA fine-tuning approach is in lack of flexibility across timesteps, we introduce timestep-aware LoRA (TALoRA), incorporating multiple LoRAs and a timestep-aware router to dynamically select the appropriate LoRA for each timestep in the denoising process. (3) To further improve the effectiveness of fine-tuning, we introduce a denoising-factor loss alignment (DFA), ensuring the loss function, and the guidance of the fine-tuning, are consistent with the actual quantization deterioration across timesteps.

In summary, our contributions are as follows:

(i) We are the first to identify that signed FP quantization struggles with asymmetric activations, which arise from the nonlinear behavior of activation functions. To address this, we introduce the MSFP framework, which is also the first effective application of unsigned FP quantization in quantization, offering a novel approach for achieving low-bit quantization.

(ii) For the fine-tuning of low-bit diffusion models, which is based on the denoising process, we precisely define it as a multi-task process across timesteps and introduce an efficient TALoRA module. Furthermore, we improve the alignment of loss function with quantization error via DFA strategy, enabling stable and reliable fine-tuning to achieve low-bit diffusion models.

(iii) We focus on identifying and eliminating three major barriers to effective low-bit FP quantization in diffusion models. Extensive experiments on DDIM and LDM demonstrate that our work achieves SOTA results for 4-bit quantization. The proposed MSFP, TALoRA and DFA have greatly advanced the progress of low-bit quantization in diffusion models.

2. Related Work

2.1. Diffusion Model Quantization

There are two main approaches in the quantization of diffusion models: QAT [6, 14] and PTQ [28]. QAT is particularly effective for low-bit quantization while it retrains the model from scratch, consuming extensive computational resources and time [20]. In contrast, PTQ offers greater time efficiency, making it more practical for large models in real-world applications. Recent advancements in PTQ have concentrated on optimizing calibration datasets to enhance reconstruction accuracy [9, 19, 25, 32] and mitigating quantization errors stemming from the temporal and structural properties of diffusion models [3, 19, 33, 35]. To further achieve fully 4-bit quantization, fine-tuning techniques has been introduced in PTQ-based quantization. EfficientDM [8] develops a LoRA-based fine-tuning framework, while QuEST [39] focuses on optimizing quantization-unfriendly activations. Both works stagnate in addressing data distribution during fine-tuning, failing to account for the specific challenges involved in fine-tuning within the context of diffusion model quantization.

Meanwhile, recent advancements in FP quantization have made significant strides in model quantization [17, 23], highlighting its potential for diffusion models. While existing low-bit quantization methods perform well for linear models, applying FP quantization to convolutional diffusion models remains more challenging. To date, only one study has explored 8-bit activation quantization in diffusion models using a basic search-based approach [2], with no research addressing lower-bit quantization. This underscores both the challenges and the untapped potential of achieving 4-bit FP quantization for diffusion models.

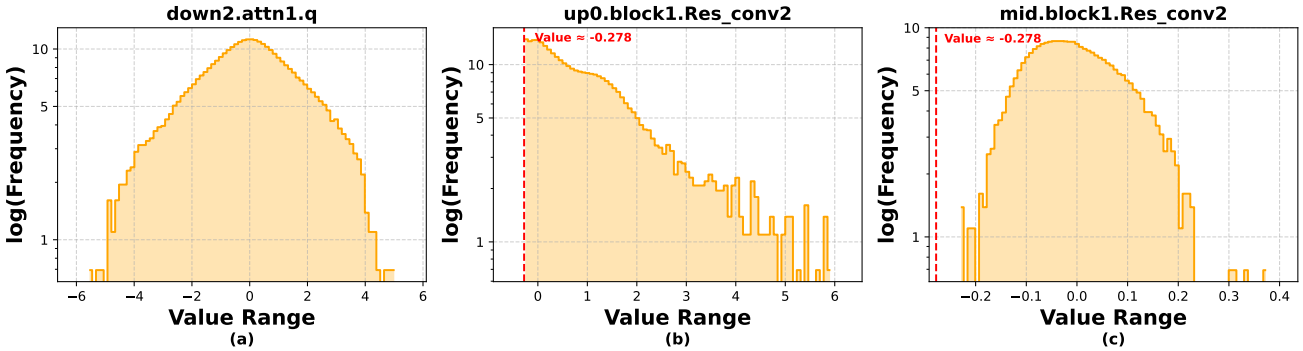

Figure 1. The activation distributions in NALs and AALs, results on the CelebA dataset. (a) The paradigm of NALs with symmetric activations. (b) The typical paradigm of AALs with asymmetric activations, where unsigned FP quantization is more suitable. (c) The infrequent paradigm of AALs with relatively symmetric activations, where either signed or unsigned FP quantization could be applicable.

2.2. Parameter-Efficient Fine-Tuning

Parameter-efficient fine-tuning (PEFT) has emerged as an effective alternative to full model fine-tuning, focusing on adjusting only a subset of parameters while keeping the majority frozen, thereby reducing storage overhead. Low-rank adapters (LoRA) [12], originally developed for large language models, have become one of the most widely used PEFT methods. Leveraging LoRA's strong transferability, QLoRA [5] can be effectively applied to fine-tune low-bit diffusion models. However, previous LoRA-based fine-tuning is suboptimal in practice. In this paper, we make a further exploration on this and incorporate timestep-level adaptation inspired by MoELoRA [7], enhancing performance of low-bit quantized DMs.

3. Challenges & Exploration

3.1. Preliminary

Diffusion Models. The diffusion model is a new generation framework that completes learning by adding noise and completes generation in a denoising manner. As for the forward process, the noise is injected into ground-truth images $\boldsymbol{x}_0$ at random timesteps. This enables the DMs to learn the distributions of noise through noisy images $\boldsymbol{x}_t$, which can be obtained as follows [11]:

$$\boldsymbol{x}_t = \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\varepsilon}. \tag{1}$$

Here $\boldsymbol{\varepsilon}$ represents a standard Gaussian noise and $\overline{\alpha}_t$ is the accumulated noise intensity, calculated by:

$$\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i, \tag{2}$$

where α_s governs the noise intensity under each timestep. Once the DMs are well-trained, the Gaussian noise image will be inputted to the DMs, and undergo the iterative denoising process. Specifically, the noise can be predicted by DMs and used to obtain the image $\boldsymbol{x}_t$ under timestep t, which could range from T to 0, with the objective of obtaining the image $\boldsymbol{x}_{t-1}$ in a superior quality:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}} \cdot \boldsymbol{\varepsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right) + \sigma_t \boldsymbol{\delta}, \tag{3}$$

where $\boldsymbol{\varepsilon}_\theta\left(\boldsymbol{x}_t, t\right)$ is the predicted noise at timestep t and $\boldsymbol{\delta}$ is a newly added noise with the factor σ_t to ensure diverse results. Here we define a denoising factor γ_t, which is formulated as:

$$\gamma_t = \frac{1}{\sqrt{\alpha_t}} \cdot \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}. \tag{4}$$

According to Equation 3, γ_t indicates the impact of the prediction noise under timestep t. The greater the factor γ_t, the stronger the predicted noise effect in the denoising.

Model Quantization. The fundamental paradigm of model quantization is the process of transforming a continuous data distribution into a finite set of discrete points. Consequently, the quality of the quantized model is inextricably related to the distribution of discrete points. According to the discrete point type, quantization is defined as two categories, widely-used INT quantization and emerging FP quantization. Analogous to the process of INT quantization, a floating-point vector $\boldsymbol{x}$ can be quantized as follows:

$$\widehat{\boldsymbol{x}} = Clip\left(\lfloor\frac{\boldsymbol{x}}{s}\rceil + z, l, u\right) \cdot s. \tag{5}$$

Here $\lfloor\rceil$ is the rounding operation. l and u are the minimum and maximum quantization thresholds while scaling factor s and zero-point z together constitute the quantization parameters. As shown in Equation 5, INT quantization results in an evenly spaced distribution of discrete points. While INT quantization is straightforward to implement, it may be too simplistic for continuous distributions that have significant variations in density, where evenly spaced intervals in INT quantization are not effective. On the contrary,

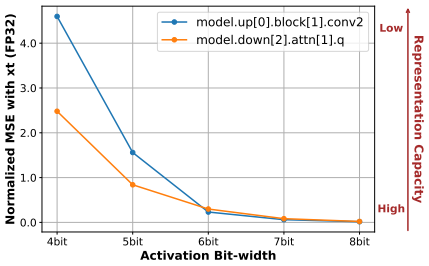

Figure 2. Effect of bit-width reduction on activation representation capacity in AALs (blue) and NALs (orange) under signed FP quantization, evaluated on CelebA dataset.

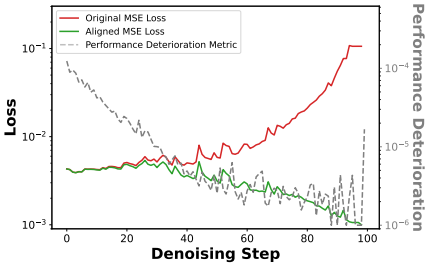

Figure 3. Two loss, and performance degradation between the quantized and full-precision models across steps. Compared with metric, the original loss shows an inverse trend, while the aligned loss remains consistent.

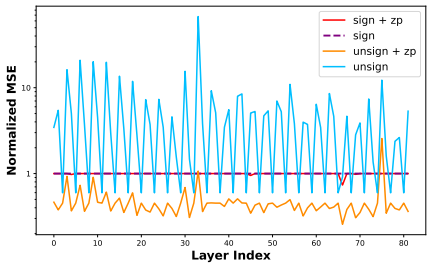

Figure 4. The MSE of activations before and after quantization across all AALs under four different strategies, normalized against the baseline of signed FP quantization without zero point (purple).

the discrete points under FP quantization are not uniformly spaced in distribution, as shown below [22]:

$$f = (-1)^s \, 2^{p-b} \left(1 + \frac{d_1}{2} + \frac{d_2}{2^2} + \ldots \frac{d_m}{2^m}\right), \qquad (6)$$

where s is the sign bit, d_i is the m-bit mantissa, p is the e-bit exponent and b is the bias that serves as both scaling factor and threshold in INT quantization. Previous studies in FP quantization rely on signed FP quantization, where s is set to 1. For an n-bit FP quantization, the bit-width is distributed across the mantissa, exponent, and sign bit, with the condition $m + e + s = n$. The different combinations of m and e allow FP quantization to represent data in multiple formats with a fixed number of bits, denoted as $E_i M_j$, where i-bit exponent and j-bit mantissa are specified. A larger j results in higher precision within each interval, while a larger i expands the range of covered intervals.

There appears to be an inherent suitability of FP quantization for DMs, as the majority of weights and activations follow a normal distribution symmetric around $value = 0$. This aligns well with the unevenly spaced distribution, where discrete points are dense in the small-value region and sparse in the large-value region, in FP quantization [42]. Additionally, the flexible formatting mechanism allows FP quantization to better accommodate complex distribution scenarios.

3.2. Barriers to Effective Low-Bit FP Quantization in Diffusion Models

In order to implement high-performing 4-bit diffusion models, we implemented FP quantization that is more compatible with the data distribution of diffusion models, with the search-based strategy in previous work [23]. Building on this, we deployed a LoRA-based fine-tuning strategy to address the performance degradation caused by 4-bit quantization. Despite this, we find that the performance of the 4-bit FP diffusion model remains satisfactory and we identify two main issues that we are facing: (1) For FP quantization, although FP quantization works well at 8 bits, it experiences a sharp performance degradation at 4 bits. *How to improve the ability of FP quantization to represent data under low-bit quantization*? (2) For standard post-training LoRA-based fine-tuning, exhibits instability and suboptimal results when applied to low-bit quantized diffusion models. *How can we make LoRA more efficient and accurate in learning the loss information at different denoising timesteps*? In the following parts, we will explore the underlying causes of these two issues and the feasible strategies to address them.

Observation 1: Previous signed FP quantization fails to achieve effective low-bit quantization in Activation-Anomalous Layers.

In diffusion models, we observed that the nonlinear activation layer $SiLU$, defined as $SiLU(\boldsymbol{x}) = \frac{\boldsymbol{x}}{1 + e^{-\boldsymbol{x}}}$, is commonly situated between layers.

$SiLU$ causes the abnormal activations for the subsequent layer. As depicted in Panel (b) of Figure 1, all values below 0 are compressed into the range of $[-0.278, 0)$. In this paper, we refer to layers with such asymmetric activations as Anomalous-Activation-Distribution Layers (AALs) and the other layers as Normal-Activation-Distribution Layers (NALs). Mainstream signed FP quantization typically sets the maximum threshold of the quantizer high to accurately represent normal positive activations. However, this approach results in a significant precision loss when dealing with values below 0. Figure 2 illustrates the representation capacity of FP signed quantization in both NALs and AALs under different bit widths. When the bit width drops below 6 bits, AALs suffer more severe performance degradation compared to NALs, which ultimately leads to the failure in low-bit FP quantization. This phenomenon suggests that mitigating the performance decline in AALs is a critical step towards improving low-bit FP quantization in diffusion models.

Observation 2: The single-LoRA-based strategy is overly simplistic for fine-tuning quantized diffusion models across different timesteps.

Previous work has focused on adapting LoRA for the quantization of diffusion models but has not fully explored LoRA's performance in the context of denoising. Considering that the denoising process of diffusion models starts with recovering outlines and progresses to restoring details, we question whether the single-LoRA fine-tuning strategy can handle this complexity. In Table 1, we compare the baseline model, which uses a single-LoRA strategy for fine-tuning, with two alternative strategies. The second strategy assigns a separate LoRA for the first and last 50 timesteps, resulting in a significant improvement over the baseline. In contrast, the third strategy also introduces a dual-LoRA strategy, it randomly selects one for each timestep, resulting in much worse performance. These results suggest that applying multiple LoRAs, allocated in a structured manner across timesteps, enhances model performance, while disordered selection of LoRAs could lead to suboptimal results. This motivates us to approach the fine-tuning of quantized diffusion models as a multi-task process and assign multiple LoRAs across different timesteps with a rational approach in allocation.

Observation 3: The MSE of the predicted noise of full-precision and quantized models does not reflect the actual impact of quantization at different timesteps.

In fine-tuning, a commonly used loss function calculates the MSE between the noise predictions of the full-precision and quantized diffusion models, using denoised images from the full-precision model at the previous timestep as inputs:

$$L_{\boldsymbol{\varepsilon}_\theta}^t = \left\| \boldsymbol{\varepsilon}_\theta \left(\boldsymbol{x}_t, t \right) - \widehat{\boldsymbol{\varepsilon}}_\theta \left(\boldsymbol{x}_t, t \right) \right\|^2 . \tag{7}$$

By observing the variation in this loss during single-LoRA fine-tuning (see Figure 3), we identify an unexpected trend: the loss increases progressively faster as denoising advances. This contradicts the principle of denoising, where image quality should improve with each step, and the impact of predicted noise should diminish over time. By the final step, the impact of quantization on model performance should be negligible, as the predicted noise no longer affects the input. However, the loss is at its maximum, indicating that quantization error is most significant at this stage, contradicting the expectation that its influence should diminish over time.

To highlight the discrepancy between the loss and actual quantization errors, we define the performance gap at each step as the difference in denoising quality between the quantized and full-precision models, as the ultimate goal of denoising is to yield high-quality images. We input the previous output image $\boldsymbol{x}_t$ from the full-precision model and calculate the performance gap between the denoised image $\boldsymbol{x}_{t-1}$ from the full-precision model and the denoised image $\widehat{\boldsymbol{x}}_{t-1}$ from the quantized model, measured by $MSE(\boldsymbol{x}_{t-1}, \widehat{\boldsymbol{x}}_{t-1})$. As shown in Figure 3, this misalignment between the loss and the actual performance gap leads to deviations in LoRA's learning. This necessitates aligning our loss function with the denoising process during fine-tuning.

Method	Bits (W/A)	FID ↓
FP	32/32	6.49
Single-LoRA	4/4	19.41
Dual-LoRA (Split Steps in Half)	4/4	17.07
Dual-LoRA (Random Allocation)	4/4	41.96

Table 1. The impact of the number of LoRAs and their allocation across timesteps on the performance of the fine-tuning. The results is evaluated by 4-bit quantization on CelebA dataset.

4. Methodology

In this section, we explore the issues identified in Section 3.2 and propose corresponding solutions. As illustrated in Figure 5, we introduce a mixup-sign FP quantization framework to address the diverse activation distributions in the first stage. During fine-tuning, the timestep-aware routing mechanism and denoising-factor loss alignment work in tandem to enable high-quality learning, ultimately enabling the realization of optimized 4-bit FP diffusion models.

4.1. Mixup-Sign Floating Point Quantization

To address the challenges of low-bit FP quantization failure in AALs, we leverage FP quantization by allocating more discrete points to areas with high data concentration. Motivated by the half-normal distribution of activations in AALs, we introduce unsigned floating-point quantization. However, as shown in Equation 6, when using unsigned FP quantization with s set to 0, we round all data in the sub-zero range to zero, losing important negative information. To address this, we introduce a zero point in the range of [-0.278, 0) to recover most sub-zero activations. The updated quantization formula becomes:

$$f_{unsign} = (-1)^s \, 2^{p-b} \left(1 + \frac{d_1}{2} + \frac{d_2}{2^2} + \ldots \frac{d_m}{2^m} \right) + z, \tag{8}$$

where s is set to 0 and z is the newly added zero point. By freeing the 1-bit sign bit, which is ineffective in signed FP quantization, and using it as additional exponent / mantissa bit width, we fully utilize the representation capacity. As illustrated in Figure 4, unsigned FP quantization with a zero point significantly improves representation in over 95% of AALs, compared to traditional signed FP quantization.

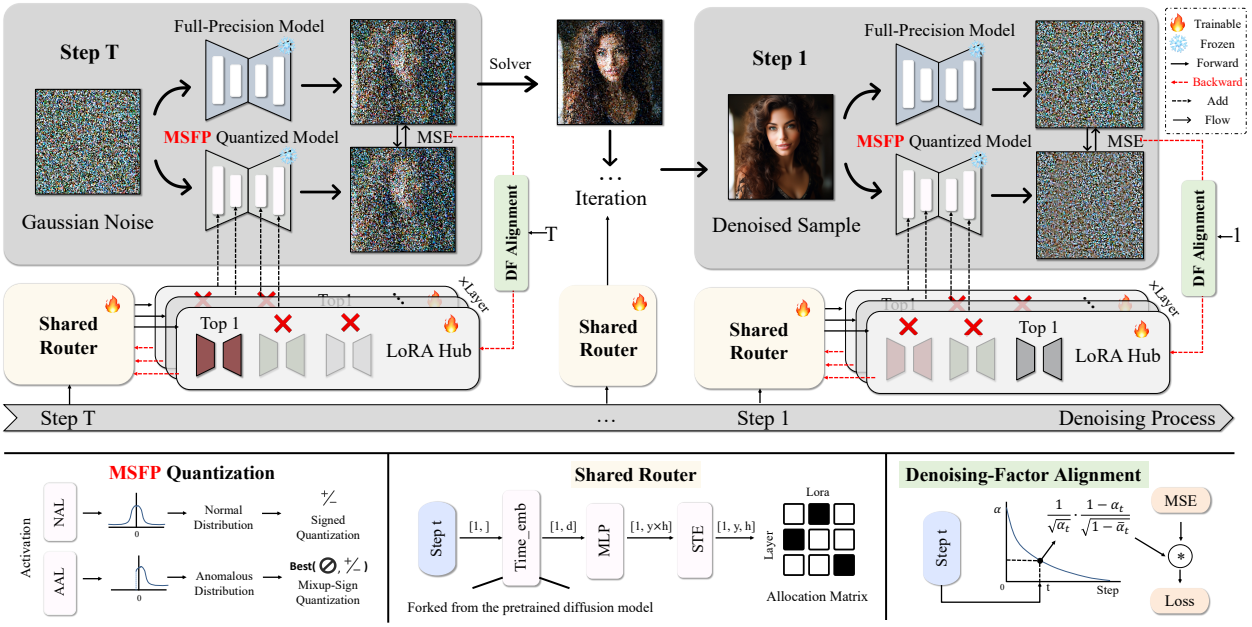

Figure 5. The pipeline of our proposed method. UNets are applied to the Mixup-Sign Floating-Point Quantization (MSFP), where distinct floating-point quantization schemes are employed for Anomalous-Activation-Distribution Layers (AALs) and Normal-Activation-Distribution Layers (NALs). During the fine-tuning stage, multiple LoRA modules are introduced, and a timestep-aware routing mechanism is used for dynamic LoRA allocation across different timesteps. Additionally, a denoising-factor alignment technique is employed to align the loss function with quantization-induced performance degradation.

However, there are rare cases where performance slightly goes worse due to the diversity of anomalous distributions. Panel (c) of Figure 1 shows that in such cases, the activation distribution may resemble a normal distribution, where signed FP quantization might perform better. Figure 4 further indicates that introducing a zero point into signed FP quantization is unnecessary, offering minimal improvement in a few cases.

Given the strong performance of unsigned FP quantization with a zero point and the diversity of AALs, we propose a mixup-sign FP quantization framework. During the search-based initialization phase, we use signed FP quantization for NALs and introduce both unsigned FP quantization with a zero point and signed FP quantization for AALs. This approach addresses AAL challenges in low-bit quantization while minimizing computational overhead by adding the zero point only to unsigned FP quantization.

4.2. Timestep-Aware Router for LoRA Allocation

In Section 3.2, we observe that a single LoRA cannot fully capture all the information across timesteps due to the diverse generative characteristics at different timesteps. We also find that a reasonable allocation of different LoRAs for timesteps will be beneficial to fine-tuning. In this section, we introduce a timestep-aware LoRA allocation method that dynamically assigns optimal LoRA to each timestep, maximizing fine-tuning effectiveness.

Our method relies on a learnable router (illustrated in Figure 5), a module shared across all timesteps. It takes the timestep as input and outputs selection probabilities for each LoRA across UNet layers. For each timestep, the LoRA with the highest probability corresponding to the router's output is inserted into the quantized model for fine-tuning or inference. In a router network, the main components are a time embedding layer and an MLP layer. The time embedding layer, derived from a pre-trained diffusion model, converts the scalar into a d-dimensional embedding. The MLP layer then maps the embedding to a LoRA allocation distribution, where y is the number of quantized layers and h is the size of the LoRA Hub. Using an STE method [1], this distribution is converted into 0/1 probabilities to allocate suitable LoRAs to the diffusion model's quantized layers.

4.3. Denoising Factor Aligned Loss

To address the mismatch between the actual performance gap and the loss used during quantization, we review the denoising principle depicted in Equation 3 and pinpoint the cause of the mismatch: the time-step-dependent constraint on the predicted noise is not sufficiently accounted for during the denoising process. Therefore, we implement a modification to the loss function based on predicted noise:

$$L^t = \gamma_t \cdot L^t_{\boldsymbol{\epsilon}_\theta}. \tag{9}$$

By introducing γ_t, which accurately reflects the utilization of the predicted noise at each time step, we achieve a preliminary alignment between the loss and the actual quantization error, as shown in Figure 3. This facilitates more accurate fine-tuning, leading to better performance recovery in the low-bit diffusion model.

5. Experiment

5.1. Experimental Setup

Models and Metrics. To verify the effectiveness of the proposed method, we evaluate it with two widely adopted diffusion paradigms: DDIM [34] and LDM [30]. For DDIM experiments, we evaluate on CIFAR-10 [16] and CelebA [26]. For LDM, we test unconditional generation on LSUN-Bedroom [41] and LSUN-Church [41] and conditional generation on ImageNet [4]. The performance of the diffusion models is evaluated with Inception Score (IS) [31], Fréchet Inception Distance (FID) [10] and Sliding FID (sFID) [31]. All metrics are evaluated based on 50k samples generated by the DDIM solver [34].

Quantization Detail. We employ standard layer-wise quantization for both weights and activations. Except for the input and output layers, which are typically set to 8-bit, all other convolution and linear layers are quantized to the target bit-width. Furthermore, we generate 256 samples for the calibration set based on Q-Diffusion [19] for bias initialization and use the method from [2] to obtain the optimal quantization parameters.

Baseline. We compare two main types of quantization methods: PTQ methods (Q-Diffusion [19] and EDA-DM [38]) and fine-tuning methods (EfficientDM [8] and QuEST [39]). Since these fine-tuning methods involve special settings like non-full-layer quantization, we standardize the settings of EfficientDM in a consistent manner and procure other standardized results from DilateQuant [24]. Comparison with special settings is provided in Appendix.

5.2. Quantization Performance

Unconditional Generation. Table 2 presents the results of unconditional image generation across multiple datasets. With 6-bit quantization, our method achieves nearly identical performance to full precision. Our 4-bit quantized models achieve SOTA results in all tasks, significantly outperforming previous baseline methods. Notably, on CIFAR-10, our 4-bit quantized model results in an FID score that is only 1.84 worse than full precision, an almost negligible degradation, while previous methods struggle with 4-bit quantization. Compared to the fine-tuning method [8], our 4-bit quantized models improve FID by 32.38, 24.15, and 9.63 on three datasets, respectively, and still maintain remarkable IS. Additionally, we provide the performance of 4-bit and 6-bit quantized models on CelebA in Appendix.

Task	Method	Prec. (W/A)	FID ↓	IS ↑
CIFAR-10 32x32 DDIM steps = 100	FP	32/32	4.26	9.03
	Q-Diffusion	6/6	9.19	8.76
	EDA-DM	6/6	26.68	**9.35**
	EfficientDM	6/6	25.03	8.08
	Ours (h=2)	6/6	4.26	9.04
	Ours (h=4)	6/6	**4.23**	9.06
	Q-Diffusion	4/4	N/A	N/A
	EDA-DM	4/4	120.24	4.42
	EfficientDM	4/4	38.40	7.32
	Ours (h=2)	4/4	**6.02**	8.79
	Ours (h=4)	4/4	6.10	**8.90**
LSUN (Bedroom) 256x256 LDM-4 steps = 100 eta = 1.0	FP	32/32	3.02	2.29
	Q-Diffusion	6/6	10.10	2.11
	EDA-DM	6/6	10.56	2.12
	QuEST	6/6	10.10	2.20
	EfficientDM	6/6	12.95	**2.57**
	Ours (h=2)	6/6	8.42	2.49
	Ours (h=4)	6/6	**8.40**	2.49
	Q-Diffusion	4/4	N/A	N/A
	EDA-DM	4/4	N/A	N/A
	QuEST	4/4	N/A	N/A
	EfficientDM	4/4	36.36	**2.69**
	Ours (h=2)	4/4	**12.21**	2.47
	Ours (h=4)	4/4	12.34	2.48
LSUN (Church) 256x256 LDM-8 steps = 100 eta = 0.0	FP	32/32	4.06	2.70
	Q-Diffusion	6/6	10.90	2.47
	EDA-DM	6/6	10.76	2.43
	QuEST	6/6	6.83	2.65
	EfficientDM	6/6	7.45	**2.80**
	Ours (h=2)	6/6	**6.24**	2.73
	Ours (h=4)	6/6	6.38	2.73
	Q-Diffusion	4/4	N/A	N/A
	EDA-DM	4/4	N/A	N/A
	QuEST	4/4	13.03	2.63
	EfficientDM	4/4	18.40	**2.97**
	Ours (h=2)	4/4	8.81	2.70
	Ours (h=4)	4/4	**8.77**	2.71

Table 2. Quantization performance of unconditional generation. 'Prec. (W/A)' denotes the quantization bit-width. 'N/A' denotes failed image generation. h denotes the size of LoRA Hub.

Conditional Generation. Table 3 presents the results of our conditional generation experiments on ImageNet. We observe that FID is not always a reliable metric in this context, as an unexpected trend emerges: as the bit-width of the quantized model decreases, the FID score improves, which contradicts the expected trend. Therefore, our discussion of the ImageNet results focuses on other evaluation metrics. As shown by the IS and sFID, our method achieves

Figure 6. A visual comparison of generation results using our method across different quantization bit-widths, with the LSUN-Church dataset as an example.

performance comparable to that of the full-precision model with 6-bit quantization. Even with 4-bit quantization, we achieve SOTA results in terms of sFID, improving by 7.00 over the previous best method, EfficientDM [8]. Furthermore, our method significantly outperforms two other fine-tuning methods in the IS metric. Visual evaluations further confirm that the generated images maintain high quality, exhibiting clear and coherent content, as shown in Figure 6. More visualization results will be presented in Appendix.

Method	Prec. (W/A)	sFID ↓	FID ↓	IS ↑
FP	32/32	7.67	11.69	364.72
EDA-DM	6/6	8.02	11.52	**360.77**
QuEST	6/6	9.36	**8.45**	310.12
EfficientDM	6/6	6.88	9.54	351.79
Ours (h=2)	6/6	7.43	10.10	349.91
Ours (h=4)	6/6	**6.65**	10.10	351.79
EDA-DM	4/4	36.66	20.02	**204.93**
QuEST	4/4	29.27	38.43	69.58
EfficientDM	4/4	14.42	12.73	139.45
Ours (h=2)	4/4	**7.42**	**6.50**	190.74
Ours (h=4)	4/4	8.23	7.43	177.40

Table 3. Quantization performance of conditional generation for fully-quantized LDM-4 models on ImageNet 256×256 with 20 steps. 'Prec. (W/A)' denotes the quantization bit-width. h denotes the size of LoRA Hub.

5.3. Ablation Study

The ablation experiments are conducted on the 4-bit quantization using the CelebA dataset, which is challenging for low-bit quantization, further demonstrating the effectiveness of our approach. The baseline uses signed FP quantization combined with single LoRA fine-tuning. As shown in Table 4, all three proposed modules lead to significant performance improvements, with their combination

MSFP	Method TALoRA	DFA	Prec. (W/A)	FID ↓
✗	✗	✗	4/4	16.02
✓	✗	✗	4/4	9.60
✗	✓	✗	4/4	10.66
✓	✗	✓	4/4	8.39
✓	✓	✗	4/4	8.79
✓	✓	✓	4/4	7.69

Table 4. Ablation study on different modules we proposed. Testing on CelebA dataset with $h = 2$ LoRA Hub size.

yielding a synergistic effect. The baseline FID score is 9.53 higher than that of the full-precision model (6.49). By applying our technique, we reduce the FID by 8.18 compared to the baseline. More interesting is that we visualize the LoRA allocation distribution learned by the router as shown in Figure 7. We find that the distribution of the router-learned allocation over timesteps is consistent with the finding that the diffusion model focuses on contour generation early and on detail generation later [37].

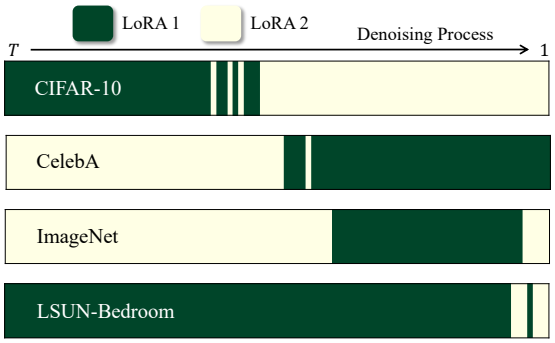

Figure 7. Distribution of LoRA allocations over timesteps obtained after router training on different datasets, when $h = 2$.

6. Conclusion

In this paper, we focus on exploring low-bit FP quantization for diffusion models. For model initialization, we innovatively introduce unsigned FP quantization with zero point to address AALs. For the fine-tuning based on the denoising process, we formulate it as a multi-task procedure. We introduce multiple LoRAs along with a router for their allocation at different timesteps, and further align the loss function, originally based on estimated noise, with the actual quantization error. We introduce unsigned FP quantization and achieve 4-bit FP quantized diffusion models. Our FP PTQ-based fine-tuning method sets a new precedent for 4-bit diffusion models, offering insight into the deployment of low-bit diffusion models in the future.

7. Acknowledgments

This work is supported by Shanghai Science and Technology Commission Explorer Program Project (24TS1401300), National Key Research and Development Program of China (No.2022ZD0160101). The computations in this research were performed using the CFFF platform of Fudan University.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 6

[2] Cheng Chen, Christina Giannoula, and Andreas Moshovos. Low-bitwidth floating point quantization for efficient high-quality diffusion models. *arXiv preprint arXiv:2408.06995*, 2024. 2, 7

[3] Huanpeng Chu, Wei Wu, Chengjie Zang, and Kun Yuan. Qncd: Quantization noise correction for diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10995–11003, 2024. 2

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms (2023). *arXiv preprint arXiv:2305.14314*, 52:3982–3992, 2023. 3

[6] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *International Conference on Learning Representations*, 2019. 1, 2

[7] Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. Mixture-of-loras: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11371–11380, 2024. 3

[8] Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. *arXiv preprint arXiv:2310.03270*, 2023. 1, 2, 7, 8

[9] Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[13] Wei Huang, Xingyu Zheng, Xudong Ma, Haotong Qin, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. An empirical study of llama3 quantization: From llms to mllms. *Visual Intelligence*, 2(1): 36, 2024. 1

[14] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018. 2

[15] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018. 1

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

[17] Andrey Kuzmin, Mart Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. Fp8 quantization: The power of the exponent. *Advances in Neural Information Processing Systems*, 35:14651–14662, 2022. 2

[18] Min Li, Zihao Huang, Lin Chen, Junxing Ren, Miao Jiang, Fengfa Li, Jitao Fu, and Chenghua Gao. Contemporary advances in neural network quantization: A survey. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2024. 1

[19] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545, 2023. 1, 2, 7

[20] Yanjing Li, Sheng Xu, Xianbin Cao, Xiao Sun, and Baochang Zhang. Q-dm: An efficient low-bit quantized diffusion model. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[21] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403, 2021. 1

[22] Fangxin Liu, Wenbo Zhao, Zhezhi He, Yanzhi Wang, Zongwu Wang, Changzhi Dai, Xiaoyao Liang, and Li Jiang. Improving neural network efficiency via post-training quantization with adaptive floating-point. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5281–5290, 2021. 4

[23] Shih-yang Liu, Zechun Liu, Xijie Huang, Pingcheng Dong, and Kwang-Ting Cheng. Llm-fp4: 4-bit floating-point quantized transformers. *arXiv preprint arXiv:2310.16836*, 2023. 2, 4

[24] Xuewen Liu, Zhikai Li, and Qingyi Gu. Dilatequant: Accurate and efficient diffusion quantization via weight dilation. *arXiv preprint arXiv:2409.14307*, 2024. 7

[25] Xuewen Liu, Zhikai Li, Junrui Xiao, and Qingyi Gu. Enhanced distribution alignment for post-training quantization

of diffusion models. *arXiv preprint arXiv:2401.04585*, 2024. 2

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

[27] Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, et al. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022. 1

[28] Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pages 7197–7206. PMLR, 2020. 2

[29] NVIDIA. Blackwell platform sets new llm inference records in mlperf inference v4.1, 2024. Available at: https://developer.nvidia.com/blog/nvidia-blackwell-platform-sets-new-llm-inference-records-in-mlperf-inference-v4-1, Accessed: 2024-11-14. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 7

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7

[32] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1972–1981, 2023. 1, 2

[33] Junhyuk So, Jungwon Lee, Daehyun Ahn, Hyungjun Kim, and Eunhyeok Park. Temporal dynamic quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 7

[35] Haojun Sun, Chen Tang, Zhi Wang, Yuan Meng, Xinzhu Ma, Wenwu Zhu, et al. Tmpq-dm: Joint timestep reduction and quantization precision selection for efficient diffusion models. *arXiv preprint arXiv:2404.09532*, 2024. 2

[36] Mart van Baalen, Andrey Kuzmin, Suparna S Nair, Yuwei Ren, Eric Mahurin, Chirag Patel, Sundar Subramanian, Sanghyuk Lee, Markus Nagel, Joseph Soriaga, et al. Fp8 versus int8 for efficient deep learning inference. *arXiv preprint arXiv:2303.17951*, 2023. 2

[37] Binxu Wang and John J Vastola. Diffusion models generate images like painters: an analytical theory of outline first, details later. *arXiv preprint arXiv:2303.02490*, 2023. 2, 8

[38] Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate data-free quantization for diffusion models. *arXiv preprint arXiv:2305.18723*, 2(5), 2023. 7

[39] Haoxuan Wang, Yuzhang Shang, Zhihang Yuan, Junyi Wu, Junchi Yan, and Yan Yan. Quest: Low-bit diffusion model quantization via efficient selective finetuning. *arXiv preprint arXiv:2402.03666*, 2024. 1, 2, 7

[40] Jie Wang, Huanxi Liu, Dawei Feng, Jie Ding, and Bo Ding. Fp4-quantization: Lossless 4bit quantization for large language models. In *2024 IEEE International Conference on Joint Cloud Computing (JCC)*, pages 61–67. IEEE, 2024. 2

[41] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 7

[42] Yijia Zhang, Lingran Zhao, Shijie Cao, Sicheng Zhang, Wenqiang Wang, Ting Cao, Fan Yang, Mao Yang, Shanghang Zhang, and Ningyi Xu. Integer or floating point? new outlooks for low-bit quantization on large language models. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2, 4