Memory-Efficient Fine-Tuning for Quantized Diffusion Model

Hyogon Ryu, Seohyun Lim, and Hyunjung Shim

Korea Advanced Institute of Science and Technology (KAIST)
{hyogon.ryu, seohyunlim, kateshim}@kaist.ac.kr

Abstract. The emergence of billion-parameter diffusion models such as Stable Diffusion XL, Imagen, and DALL-E 3 has significantly propelled the domain of generative AI. However, their large-scale architecture presents challenges in fine-tuning and deployment due to high resource demands and slow inference speed. This paper explores the relatively unexplored yet promising realm of fine-tuning quantized diffusion models. Our analysis revealed that the baseline neglects the distinct patterns in model weights and the different roles throughout time steps when finetuning the diffusion model. To address these limitations, we introduce a novel memory-efficient fine-tuning method specifically designed for quantized diffusion models, dubbed TuneQDM. Our approach introduces quantization scales as separable functions to consider inter-channel weight patterns. Then, it optimizes these scales in a timestep-specific manner for effective reflection of the role of each time step. TuneQDM achieves performance on par with its full-precision counterpart while simultaneously offering significant memory efficiency. Experimental results demonstrate that our method consistently outperforms the baseline in both single-/multi-subject generations, exhibiting high subject fidelity and prompt fidelity comparable to the full precision model.

Keywords: Quantization · Diffusion Model · Transfer Learning

1 Introduction

Diffusion models have been a de facto standard in generative models, especially in image synthesis [6,16,34,40,44]. They are widely used in various applications, such as image super-resolution [28,45], inpainting [31,50], and text-to-image generation [1,7,9,40,42]. However, their slow generation process and high memory and computational requirements pose significant challenges for practical use.

With the emergence of billion-parameter diffusion models such as Stable Diffusion XL [36], Imagen [44], and DALL-E 3 [3], the issues of slow inference and computational load are becoming more pronounced. Recent studies have focused on model quantization to address these concerns. Quantization [11,20,21,27,29,46] is a key model compression technique that uses lower-bit representations (e.g., 4-bit, 8-bit) for model parameters, thus drastically improving computational and memory efficiency. Notably, PTQ4DM [46] has achieved 8-bit quantization in

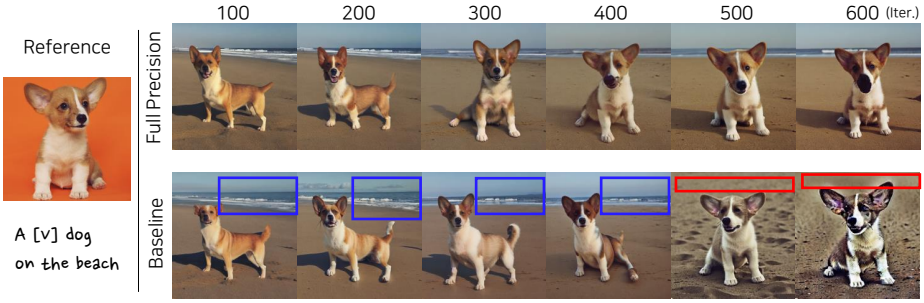


Fig. 1: Comparison between fine-tuning a full precision model and fine-tuning a quantized diffusion model with the baseline. Unlike the `fp` model, the `baseline` cannot achieve both prompt fidelity and subject fidelity simultaneously. Up to 400 iterations, it retains high image quality but fails to accurately reflect reference features. After 500 iterations, the ocean disappears, and further training leads to noticeable artifacts. Blue boxes indicate where the ocean is present, while red boxes highlight areas where the ocean should be but is missing. A unique token, `[V]`, is used as an identifier describing images provided by users.

diffusion models by constructing a timestep-aware calibration dataset and Q-Diffusion [29] has accomplished both 8-bit and 4-bit quantization by separating the shortcut layer through activation analysis.

Given the growing role of diffusion models as vision foundation models, the direct fine-tuning of quantized diffusion models for specific applications is an unexplored yet highly impactful research direction. This approach is inspired by recent developments in the large language model (LLM), where techniques like Alpha Tuning, PreQuant, and PEQA [10, 22, 24] have been investigated for fine-tuning quantized LLMs.

Building on the success of the LLM community, we developed a baseline for fine-tuning quantized diffusion models. Leveraging publicly available quantized checkpoints from Q-Diffusion, we trained the model using the PEQA methodology, commonly used for fine-tuning quantized LLMs. For fine-tuning diffusion models, we utilized the common diffusion personalization technique, DreamBooth. Combining these three cutting-edge methods, we established a baseline and observed its performance trend. As seen in Fig. 1, the `baseline` model produced unsatisfactory results in terms of either fidelity or quality. The generated results either fail to reflect the concept of text prompt (e.g., ocean) or personalized categories denoted by an identifier token `[V]`. Moreover, image quality degraded with more training iterations.

To understand the performance limitations, we first fine-tune the model and compare the weight update patterns before and after quantization. Interestingly, the distinct inter-channel patterns appear in the change ratio map between the full-precision (`fp`) weight matrix and its fine-tuned version. Fig. 2 illustrates the ratio map, highlighting the changes made by fine-tuning. Ideally, the ratio map

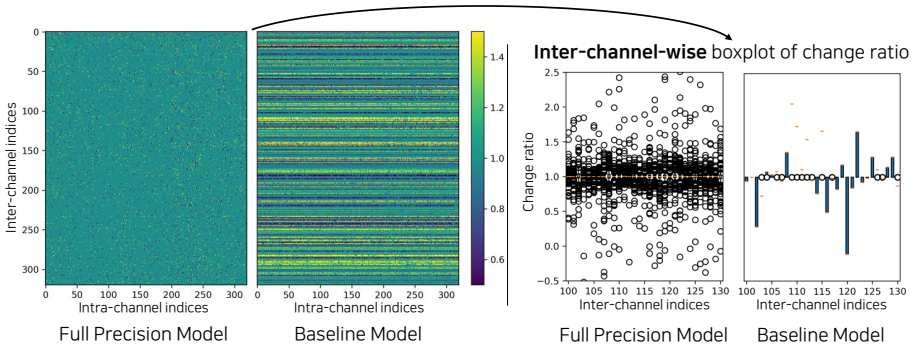


Fig. 2: Weight change ratio after the fine-tuning. The left side describes the weight change ratio of `fp` model and `baseline` in 2D image plots, and the right side describes it in an inter-channel-wise boxplot. There is a clear difference between the `baseline` and the `fp` model.

from the full-precision model should be the target. However, after quantization, these maps are clearly different from those of the full-precision model. Inter-channel patterns disappear in the `baseline` because the scale parameters, the only parameter trainable in the `baseline`, vary only within the channel. As observed in Fig. 2, the `fp` model shows distinct inter-channel patterns in weight change ratio while the `baseline` fails to capture these weight updates effectively.

Secondly, we focus on timestep-aware training for fine-tuning quantized diffusion models. Previous studies [1, 4, 26] discuss the significant role of timesteps in diffusion training. It is known that during the denoising step of diffusion models, intervals affecting content features differ from those affecting coarse features or noise-cleaning. Therefore, we designed a method to separate the timesteps and assign different roles to each timestep during fine-tuning.

To address these challenges, we propose a novel method for fine-tuning quantized diffusion models called TuneQDM. Our method tackles the aforementioned issues by (1) decomposing the quantization scales as separable functions to consider inter-channel weight patterns and (2) fine-tuning these scales timestep-wisely to reflect the role of the timestep. TuneQDM improves personalization performance while preserving the memory and computational efficiency of quantized diffusion models.

We evaluated our method on personalization (*i.e.* single-/multi-subject generation) and unconditional generation. The results demonstrate that our fine-tuned model can generate images at a level comparable to full precision fine-tuning on both personalization and unconditional generation. Compared to the baseline, TuneQDM addresses issues such as degradation of text prompt and subject fidelity. Particularly, even when fine-tuning the 4-bit (8× compressed) model, we achieved performance levels similar to those of the full precision model.

In summary, our contributions are as follows: (1) We established a strong baseline by combining existing state-of-the-art methods, identifying limitations,

and observing performance trends. (2) We introduced a novel memory-efficient fine-tuning method for quantized diffusion models named TuneQDM. By introducing a multi-channel-wise scale update, we addressed the issue of inter-channel patterns during weight updates. Additionally, by fine-tuning independent scale parameters for each timestep interval, we enabled the quantized diffusion model to effectively reflect the role of each timestep interval. (3) TuneQDM achieves both parameter efficiency and a substantial reduction in memory footprint during fine-tuning. Experimental results demonstrate that our method consistently outperforms the baseline in the single-/multi-subject generation, achieving high subject fidelity and prompt fidelity comparable to the full precision model.

2 Related work

2.1 Quantizing diffusion model

Quantization reduces model complexity and enhances speed by representing model weights with fewer bits. The two main approaches in quantization are Quantization Aware Training (QAT, integrated during training) [8, 20, 21], and Post-Training Quantization (PTQ, applied after model training) [2, 14, 19, 29, 33, 46, 48, 49].

Quantizing diffusion models tends to align well with PTQ [14, 29, 46] because diffusion models often serve as foundation models pre-trained on extensive datasets, and retraining the entire model involves high computational overheads. PTQ4DM [46] constructed a calibration dataset by considering the multi-step process of the diffusion model, and Q-Diffusion [29] proposed a quantization method that takes into account the shortcut connections in the UNet. PTQD [14] decomposed the quantization noise and corrected it. However, research on fine-tuning quantized diffusion models for downstream tasks has not yet been conducted, and we are the first to fine-tune the quantized diffusion model for the downstream task.

2.2 Personalizing diffusion model

Text-to-image (T2I) [3, 7, 36, 38, 44] generation has garnered significant attention for its ability to produce diverse and realistic images in response to textual prompts. While large models trained on extensive text and image-paired datasets excel in general tasks, they often face difficulties in generating highly personalized or novel images aligned with specific user concepts. Personalization emerges as a prominent downstream task for general diffusion models, aiming to tailor the models to individual preferences or user-defined concepts for image generation.

The users provide several image examples representing the personal concept, while additional scene components, such as backgrounds or attributes, are defined through textual prompts. Textual Inversion [9] proposed an optimization approach for word embeddings that effectively represents a given image. Meanwhile, DreamBooth [42] employs the strategy of fine-tuning a pre-trained model

to generate images with a novel perspective of the input target. Custom Diffusion [23] introduces the fine-tuning of only the key and value components of the cross-attention layer, enabling multi-subject image generation.

2.3 Parameter-efficient fine-tuning

Fine-tuning enables pretrained models to be adapted to specific tasks. However, full model fine-tuning requires significant computational resources. As an alternative to full fine-tuning, parameter-efficient fine-tuning methods have been proposed, where most of the model's parameters are frozen, and only a subset is updated. Adapter modules [17, 30, 39, 41] suggest inserting task-specific parameters within pretrained model layers. LoRA [18, 43] represents the gap between fully fine-tuned weights and pretrained weights as low-rank matrices, allowing the addition of trainable weights for task adaptation while preserving the pretrained weights.

However, parameter-efficient fine-tuning methods still struggle to handle a vast number of parameters and are less suitable for scenarios requiring smaller model sizes, such as low-power mobile devices [37,47]. Recently, methods for fine-tuning quantized models have been proposed. In the field of LLMs, QLoRA [5] proposed a low-rank adaptor applicable to quantized LLMs, while PEQA [22] suggested updating scale parameters while freezing quantized weights. Prequant [10] and OWQ [25] performed task-agnostic quantization followed by tuning a small subset of weights. Our method applies this concept to fine-tuning diffusion models and demonstrates its effectiveness in personalization and unconditional generation.

3 Motivation

The recent success of large-scale foundation models has led to their widespread adoption in numerous downstream tasks. In the field of computer vision, the diffusion model has emerged as a representative foundation model and is popularly used in various fields such as personalized generation, 3D generation combined with NeRF [32], and improving discriminative model training, leveraging models like Stable Diffusion [40] or Imagen [44]. As foundation models expand exponentially in scale over time, it has led to a growing interest in fine-tuning reduced (a.k.a., quantized) foundation models for specific downstream tasks. This concept has been actively explored in the natural language processing field using LLMs. In this study, we are the first to apply the same philosophy to the vision foundation model, the diffusion model. We specifically introduce a new problem of directly applying quantized diffusion models to a key downstream task–personalization.

Advantages of fine-tuning quantized diffusion model. Diffusion models have increasingly utilized larger UNets to improve image quality. For example, the Stable Diffusion model has expanded to 2.6 billion parameters in its SDXL variant. Similarly, Imagen's model has grown to 3 billion, while DALL-E 2 has

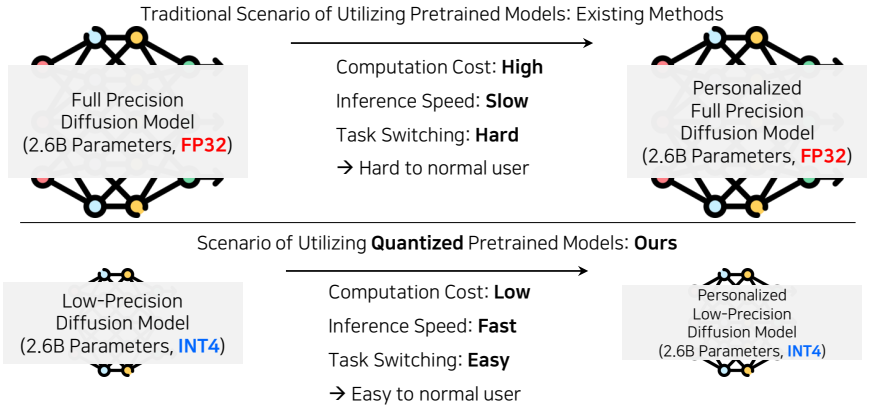


Fig. 3: Scenario of utilizing quantized pretrained models. Above: Full model is loaded and fine-tuned. Below: Quantized model is loaded and used. As the model sizes increase, fine-tuning requires significant computational cost. Therefore, directly fine-tuning the quantized model offers various efficiency advantages for users.

escalated to 5.5 billion parameters. This upward trend in the sizes of foundation models raises concerns about the efficiency of deploying and utilizing pretrained diffusion models.

To align with this trend, we introduce the concept of fine-tuning quantized diffusion models. As described in Fig 3, directly fine-tuning quantized diffusion models for downstream tasks offers several advantages. First, quantized checkpoints require significantly less memory storage, DRAM, and fewer trainable parameters than their full-precision weights. Besides, we store only the scale parameters per dataset ($\sim$ 3MB) and reuse the quantized checkpoint across different datasets. This approach eliminates the need for separate quantization processes (*i.e.*, PTQ) for each dataset and task, which is particularly time-consuming for diffusion models. Finally, the computational costs for deployment are substantially reduced compared to their full-precision counterparts.

4 Methodology

In this section, we propose a memory-efficient fine-tuning method using quantized diffusion models. We first provide a brief overview of uniform quantization and introduce the baseline using the PEQA technique. Then, we analyze the challenges of the baseline method and propose our solutions. Finally, we introduce TuneQDM.

4.1 Post-training quantization

Post-training quantization is the prevalent method to quantize the model weights with low precision. In this paper, we utilize the hardware-friendly quantization

method, uniform quantization. For a given full-precision model's weight matrix W_f, the quantization and de-quantization operation can be expressed as:

$$W_q = \texttt{clamp}\big(\texttt{round}\big(\frac{W_f}{s}\big) + z, 0, 2^b - 1\big), \tag{1}$$

$$\hat{W}_f = s \cdot (W_q - z), \tag{2}$$

where W_q and $\hat{W}_f$ represent the quantized weight indices and de-quantized weight matrix, respectively. Here, s, z, and b are per-channel scaling factors, zero points, and number of bits, respectively. The function $\texttt{round}(\cdot)$ and $\texttt{clamp}(\cdot, a, b)$ are used for rounding and clamping within the range $[a, b]$.

4.2 Baseline

We have constructed a baseline by applying PEQA, a method for fine-tuning quantized LLMs, to quantized diffusion models. PEQA involves freezing the weight integers in the quantized pretrained model and updating only the quantization scale for fine-tuning. Through PEQA, the fine-tuned weight W_{tuned} can be expressed as follows:

$$W_{tuned} = (s + \Delta s) \cdot (W_q - z) \tag{3}$$

$$= (s + \Delta s) \cdot \big(\texttt{clamp}\big(\texttt{round}\big(\frac{W_f}{s}\big) + z, 0, 2^b - 1\big) - z\big). \tag{4}$$

Here, only s is trainable, while the other parameters remain frozen. The weight indices, scale, and zero-point parameters are initialized using the quantized diffusion model checkpoint (*i.e.*, Q-Diffusion).

Limitation of baseline. During our experiments with the baseline, we identified two key issues in the fine-tuning process of the quantized model: (**P1**) **Weight update across channels:** As depicted in Fig. 2, weight updates during the fine-tuning should occur independently of the channels. However, in the baseline approach, only the intra-channel-wise scale parameters are updated, restricting changes in the inter-channel components of the weight matrix. This limitation is observed consistently in both linear and conv2d layers. (**P2**) **Limited capacity to learn denoising timestep variability:** The role of the UNet architecture in the diffusion model varies depending on the denoising timestep. However, we found that the baseline's capacity to learn these variations is limited. Even with extended training iterations, the output images that are not exactly the same with target images. There are often some quality degradation or difference in details.

4.3 Multi-channel-wise scale update

To address the aforementioned issue (**P1**), we propose the multi-channel-wise scale update method. Given the quantized weight $W_q \in \mathbb{R}^{n \times m}$ (for conv2d layers,

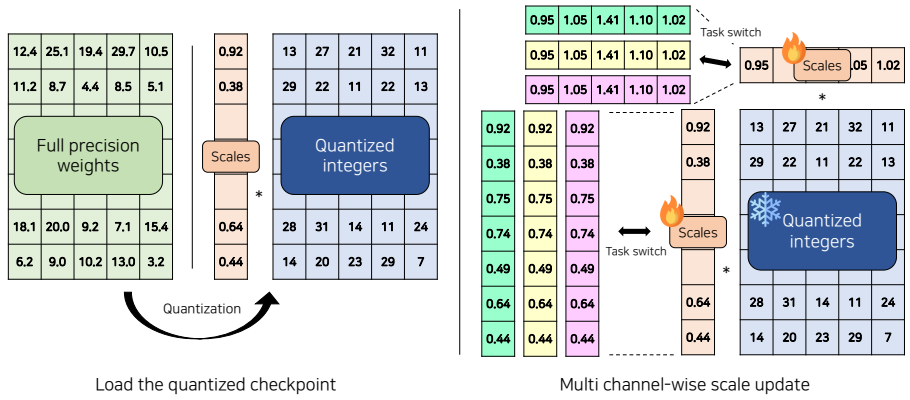


Fig. 4: Multi-channel-wise-scale. Left: Our method requires quantization to be performed only once on a pretrained model, enabling subsequent fine-tuning across various tasks without large computation costs. Right: When switching tasks, the scale pairs should be switched together. This simplifies task switching and allows the quantized model to be easily adapted to different tasks.

$W_q \in \mathbb{R}^{n\times m\times k\times k}$), and the intra per-channel quantization scale $s_{\text{out}} \in \mathbb{R}^m$, we define the inter per-channel quantization scale $s_{\text{in}} \in \mathbb{R}^n$. Subsequently, while freezing the quantized integer values W_q, we only fine-tune s_{in} and s_{out}. The fine-tuned weights W_{tuned} for downstream task are expressed as:

$$W_{\text{tuned}} = (s_{\text{out}} + \Delta s_{\text{out}}) \cdot (W_q^* - z^*) \cdot (s_{\text{in}} + \Delta s_{\text{in}}). \tag{5}$$

Here, $\Delta s_{\text{in}} \in \mathbb{R}^n$ and $\Delta s_{\text{out}} \in \mathbb{R}^m$ represent the gradient updates generated during fine-tuning for the downstream task. * indicates the frozen parameters.

The multi-channel-wise scale update is a memory-efficient fine-tuning method directly applicable to quantized diffusion models. It requires only $(m+n)$ trainable parameters, as opposed to the full weight matrix W_f, making it parameter-efficient. Moreover, as W_q and z remain fixed, adjusting only $(s_{\text{out}} + \Delta s_{\text{out}})$ and $(s_{\text{in}} + \Delta s_{\text{in}})$ values allows for easy application to other downstream tasks, facilitating task switching. While the baseline represents a specific case with all elements of s_{in} being 1 and frozen, our method can be considered as a generalized approach. The overall process of the multi-channel-wise scale update is illustrated in Fig 4.

4.4 Timestep-aware scale update strategy

To address the issue (**P2**), we propose a timestep-aware scale update strategy. P2weighting [4] elucidates the varied contributions of training timesteps to image generation. Similarly, e-Diffi [1] and MEME [26] enhance text-to-image generation performance by employing multi-expert models that adapt based on the training timesteps.

Inspired by the above-mentioned approaches, we independently fine-tune quantization scales for each timestep to update multi-experts. Assuming we propose n experts, we uniformly divide the timestep interval into n segments. For each segment, we clone and update quantization scales separately. After training, we obtain n optimized quantization scales S_i for each timestep interval I_i:

$$S_n = \{s_1, s_2, \ldots, s_n\}, \quad \mathcal{I}_n = \{I_i | (\frac{i \times t}{n}, \frac{(i+1) \times T}{n})\} \quad \text{for } i = 1 \ldots n, \tag{6}$$

where T is the total denoising steps. Our approach is much more memory-efficient compared to conventional methods that train full models to create multi-experts because we only update scale parameters. During inference, the quantized integer values remain fixed, and only the scale parameters switch according to each timestep.

Algorithm 1 TuneQDM pipeline

Require: Quantized model weight set $\{W_q^{(l)}\}_{l=1}^{L}$, Number of layer L
Require: Quantization parameter set $\{(\mathbf{s}, \mathbf{z})_l\}_{l=1}^{L}$
Require: Number of expert N and number of denoising step T
Require: Training dataset $\{(x, c, t)\}_{m=1}^{M}$, x, c, t are the noisy image, condition timestep, repectively.
for $l = 1 : L$ **do**
  for $n = 1 : N$ **do**
    Initialize intra-channel scale $\mathbf{s}^{intra} \sim \mathcal{N}(1, 0.01)$
  end for
end for
Freeze W_q and set only $\mathbf{s}$ and $\mathbf{s}^{intra}$ are trainable
for *training epoch* **do**
  for (x, c, t) *in train dataset* **do**
    Expert index $i \leftarrow \lfloor \frac{t*N}{T} \rfloor$
    Update *i-th expert* $\mathbf{s}$ and $\mathbf{s}^{intra}$ for each layer l.
  end for
end for
return Fine-tuned inter/intra-channel scale parameters $\{(\mathbf{s}, \mathbf{s}^{intra})_l\}_{l=1}^{L}$

4.5 Memory-efficient fine-tuning for quantized diffusion model

We propose TuneQDM, a novel fine-tuning method for quantized diffusion models. The overall pipeline is outlined in Algorithm 1. This method supports the previously introduced multi-channel-wise scale update and timestep-aware scale update strategies.

Firstly, we initialize the weights and quantization parameters using the quantized diffusion checkpoint and initialize the multi-channel-wise scales. Then, we

Table 1: Quantitative comparison of single-subject generation.

Method	Bits(W)	Size	# Params	DINO-I	CLIP-I	CLIP-T
Full prec.	32	3.20GB	859M	0.431	0.746	0.316
Baseline	4	0.40GB + 1.32MB	0.33M	0.519	0.787	0.313
TuneQDM	4	0.40GB + 2.48MB	0.62M	0.551 (+6.16%)	0.802 (+1.91%)	0.306 (−2.23%)
Baseline	8	0.80GB + 1.32MB	0.33M	0.581	0.824	0.300
TuneQDM	8	0.80GB + 2.48MB	0.62M	0.578 (−0.52%)	0.816 (−0.97%)	0.307 (+2.33%)

fine-tune the quantized model with the downstream task's loss function. If the timestep-aware scale update strategy is employed, additional training is conducted according to the timestep interval.

This systematic approach integrates key strategies such as multi-channel-wise scale update and timestep-aware scale update to optimize model performance for downstream tasks.

5 Experiments

In this section, we evaluate the performance of the TuneQDM across various tasks (*i.e.*, single-, multi-subject and unconditional generation. Unless specified otherwise, we utilize the DDIM sampler with $\eta = 0$ and 50 steps for single-subject generation, and 100 steps for multi-subject and unconditional generation. Implementation details can be found in the supplementary material.

5.1 Main results

Comparison on single-subject generation. Table 1 and Fig. 5 show the quantitative and qualitative comparison between our TuneQDM, `fp` and the `baseline` models. For quantitative evaluation, we assess both image and prompt fidelity. Image fidelity is measured using the CLIP [15], and DINO [35] image similarity. Prompt fidelity is evaluated using the CLIP text-to-image similarity.

In the 8-bit setting, the difference between TuneQDM and the `baseline` model is minimal. However, in the 4-bit setting, TuneQDM outperforms the `baseline` model in subject fidelity while maintaining a similar level of prompt fidelity. Specifically, we highlight that, as mentioned in previous studies [12, 42], DINO-I better reflects prompt fidelity. This can be more clearly observed through qualitative comparison.

As shown in Fig. 5, TuneQDM consistently outperforms the `baseline` by generating images that more accurately reflect both subject features and prompts. In the 4-bit setting, TuneQDM significantly performs better in accurately representing subject features compared to the `baseline` (row 1, 3, 4). In row 2, TuneQDM effectively reflects the prompt (*i.e.*, the swimming pool) into the image, unlike the `baseline` model. Moreover, TuneQDM often produces higher quality images than `baseline`. Compared to the `fp` model, TuneQDM achieves

Fig. 5: Qualitative comparisons of single-subject generation. We compared the `fp` model, TuneQDM, and baseline that fine-tuned on the target images. Subject fidelity and prompt fidelity were assessed for images generated by both TuneQDM and the baseline.

similar results despite using a $\times 8$ compressed quantization model with much fewer parameters.

Comparison on multi-subject generation Table 2 and Fig. 6 summarize the quantitative and qualitative comparison results. The `baseline` model frequently fails to reflect the features of the subjects and the prompt. In contrast, TuneQDM successfully captures both the subjects' features and the prompt. As shown in Table 2, while DINO-I and CLIP-I scores are nearly identical, TuneQDM shows a higher CLIP-T score. However, both the `baseline` and TuneQDM show a performance drop compared to the `fp` model.

In Fig. 6, the differences between each model are more evident. In row 1, TuneQDM effectively captures the features of both the "[V1] cat" and the "[V2] pot", whereas the `baseline` fails to represent the characteristics of the "[V1] cat". In row 2, the `baseline` does not depict the "[V2] sofa" at all, while TuneQDM successfully represents the features of both the "[V1] cat" and the "[V2] sofa". However, there were some cases where TuneQDM cannot perfectly reflect both prompt and subject fidelity. In multi-subject generation, which requires learning two concepts simultaneously, there was a slight performance gap compared to the `fp` model, unlike in single-subject generation. More qualitative results can be found in the supplementary material.

Table 2: Quantitative comparison of multi-subject generation.

Method	Bits(W)	Size	# Params	DINO-I	CLIP-I	CLIP-T
Full prec.	32	3.20GB	859M	0.345	0.706	0.304
Baseline	4	0.40GB + 1.32MB	0.33M	0.275	0.677	0.314
TuneQDM	4	0.40GB + 2.48MB	0.62M	0.276 (+0.36%)	0.675 (−0.30%)	0.317 (+0.96%)
Baseline	8	0.80GB + 1.32MB	0.33M	0.330	0.704	0.286
TuneQDM	8	0.80GB + 2.48MB	0.62M	0.329 (−0.30%)	0.708 (+0.57%)	0.295 (+3.15%)

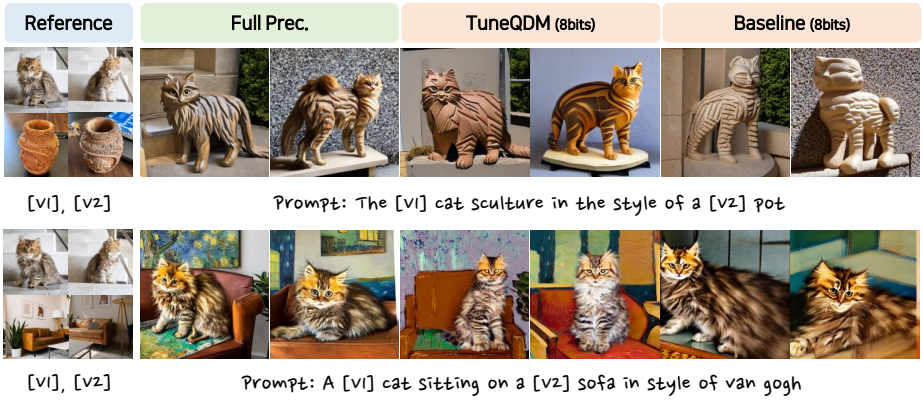


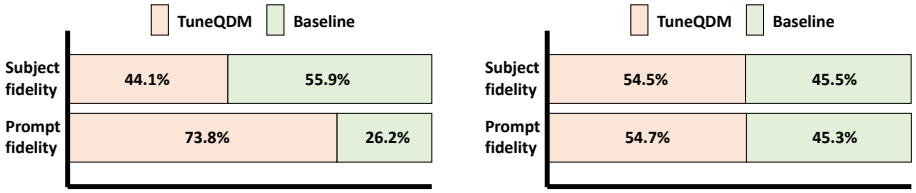
Fig. 6: Qualitative comparisons of multi-subject generation. The images generated by TuneQDM exhibit excellent quality by capturing details effectively and reflecting the prompts and subject features accurately.

Comparison on unconditional generation To evaluate the performance of TuneQDM for tasks other than personalization, we conducted fine-tuning to enhance the original purpose of the quantized diffusion model, similar to the previous study [13]. For this, we tested the performance on unconditional generation using the CIFAR-10 dataset. For this task, we compared our method with the baseline and QLoRA. The performance of diffusion models is evaluated with Inception Score(IS) and Fréchet inception distance(FID). As shown in Table 3, the baseline performed similarly to QLoRA, while TuneQDM outperformed QLoRA in both metrics with fewer parameters.

User study. To evaluate prompt and subject fidelity accurately, we conducted a user study. As shown in Fig. 7, TuneQDM exhibits slightly lower subject fidelity but significantly higher prompt fidelity compared to the baseline in single-subject generation. In a multi-subject generation, TuneQDM shows slightly higher fidelity in both subject and prompt fidelity. As known from previous research [23], prompt and subject fidelity generally have a trade-off relationship. Therefore, achieving better performance in both prompt and subject, even with slight differences, signifies a meaningful performance improvement. Experimental details can be found in the supplementary material.

Table 3: Performance comparison of fine-tuned quantized diffusion models on CIFAR-10 32 × 32.

Model	Bits	Model Size	# Params	IS	FID
Full Prec.	32	143.1MB	35.8M	9.00	4.53
Q-Diffusion	8	35.8MB	-	8.97	4.45
+ QLoRA (r=32)	8	35.8MB	8.64M	9.03	4.30
+ QLoRA (r=2)	8	35.8MB	0.57M	9.03	4.15
+ Baseline	8	35.8MB	0.03M	8.96	4.39
+ TuneQDM	8	35.8MB	0.13M	**9.17**	**3.80**



(a) Single-subject generation **(b)** Multi-subject generation

Fig. 7: User study. All models are 4-bit quantized diffusion models.

5.2 Analysis

Ablation study. To assess the impact of each component, we conducted ablation studies. As shown in Table 4, applying multi-channel-wise scale update (MCSU) and Timestep-aware scale update (TAS) with two experts to the baseline achieved the best performance in terms of FID. There was an improvement in performance when each component was applied.

Limitation. As shown in Fig 8, our method has several failure cases. In a multi-subject generation, there were some cases where only one subject was reflected, or the prompt was not reflected. The first row shows cases where the full precision model also failed to reflect the subjects (case 1). As shown in the second row, there were cases where subjects were well reflected in the full precision model but not in the quantized model (case 2). Case 1 can be attributed to the limitations of the personalization methodology, whereas case 2 specifically occurs when using the quantized model, indicating the need for additional solutions to address this problem.

6 Conclusion

This paper addressed the problem of the fine-tuning of quantized diffusion models for the first time. Inspired by the unique characteristics of diffusion models, we proposed TuneQDM, a memory-efficient fine-tuning method for quantized models. Our approach represents scales as separable functions to account for weight update patterns and customizes quantization scales for distinct intervals,

Table 4: Ablation study. Results showing the impact of MCSU and TAS on IS and FID.

Model	Bits	MCSU	TAS	IS	FID
Full Prec.	32	-	-	9.00	4.53
Baseline	8	✗	1	8.96	4.39
TuneQDM	8	✗	2	**9.19**	4.24
	8	✗	4	9.03	4.25
	8	✓	1	8.97	4.33
	8	✓	2	<u>9.17</u>	**3.80**
	8	✓	4	9.02	<u>4.15</u>

Fig. 8: Failure cases. Above: Both the `fp` and TuneQDM models fail to accurately reflect the prompt. Below: The `fp` model successfully captures both the subject and prompt, while TuneQDM does not.

effectively enhancing model capacity with minimal memory overhead. TuneQDM achieves both parameter efficiency and a substantial reduction in memory footprint during fine-tuning. Our experimental results demonstrate that our method achieves high subject fidelity and prompt fidelity while mitigating overfitting, significantly outperforming the baseline approach.

We believe fine-tuning the low-precision vision foundation models, the quantized diffusion models, holds great potential for diverse computer vision applications, alleviating slow inference and resource demands. This work can facilitate the practical deployment of diffusion models in real-world computer vision scenarios.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (NRF-2022R1A2C3011154, RS-2023-00219019, RS-2023-00240135), KEIT grant funded by the Korea government (MOTIE) (No. 2022-0-00680, 2022-0-01045, 2021-0-02068, Artificial Intelligence Innovation Hub), the IITP grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)).

References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Banner, R., Nahshan, Y., Soudry, D.: Post training 4-bit quantization of convolutional networks for rapid-deployment. Advances in Neural Information Processing Systems **32** (2019)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf (2023)
4. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11472–11481 (2022)
5. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems **36** (2024)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
7. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217 (2022)
8. Esser, S.K., McKinstry, J.L., Bablani, D., Appuswamy, R., Modha, D.S.: Learned step size quantization. arXiv preprint arXiv:1902.08153 (2019)
9. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
10. Gong, Z., Liu, J., Wang, Q., Yang, Y., Wang, J., Wu, W., Xian, Y., Zhao, D., Yan, R.: Prequant: A task-agnostic quantization approach for pre-trained language models. arXiv preprint arXiv:2306.00014 (2023)
11. Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015)
12. Hao, S., Han, K., Zhao, S., Wong, K.Y.K.: Vico: Plug-and-play visual condition for personalized text-to-image generation. arXiv preprint arXiv:2306.00971 (2023)
13. He, Y., Liu, J., Wu, W., Zhou, H., Zhuang, B.: Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. arXiv preprint arXiv:2310.03270 (2023)

14. He, Y., Liu, L., Liu, J., Wu, W., Zhou, H., Zhuang, B.: Ptqd: Accurate post-training quantization for diffusion models. Advances in Neural Information Processing Systems **36** (2024)
15. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7514–7528. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.595, https://aclanthology.org/2021.emnlp-main.595
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
17. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
19. Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., Soudry, D.: Accurate post training quantization with small calibration sets. In: International Conference on Machine Learning. pp. 4466–4475. PMLR (2021)
20. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2704–2713 (2018)
21. Jung, S., Son, C., Lee, S., Son, J., Han, J.J., Kwak, Y., Hwang, S.J., Choi, C.: Learning to quantize deep networks by optimizing quantization intervals with task loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4350–4359 (2019)
22. Kim, J., Lee, J.H., Kim, S., Park, J., Yoo, K.M., Kwon, S.J., Lee, D.: Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. Advances in Neural Information Processing Systems **36** (2024)
23. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1931–1941 (2023)
24. Kwon, S.J., Kim, J., Bae, J., Yoo, K.M., Kim, J.H., Park, B., Kim, B., Ha, J.W., Sung, N., Lee, D.: Alphatuning: Quantization-aware parameter-efficient adaptation of large-scale pre-trained language models. arXiv preprint arXiv:2210.03858 (2022)
25. Lee, C., Jin, J., Kim, T., Kim, H., Park, E.: Owq: Lessons learned from activation outliers for weight quantization in large language models. arXiv preprint arXiv:2306.02272 (2023)
26. Lee, Y., Kim, J.Y., Go, H., Jeong, M., Oh, S., Choi, S.: Multi-architecture multi-expert diffusion models. arXiv preprint arXiv:2306.04990 (2023)
27. Li, F., Liu, B., Wang, X., Zhang, B., Yan, J.: Ternary weight networks. arXiv preprint arXiv:1605.04711 (2016)
28. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022)
29. Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., Keutzer, K.: Q-diffusion: Quantizing diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17535–17545 (2023)

30. Lin, Z., Madotto, A., Fung, P.: Exploring versatile generative language model via parameter-efficient transfer learning. arXiv preprint arXiv:2004.03829 (2020)
31. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
33. Nagel, M., Baalen, M.v., Blankevoort, T., Welling, M.: Data-free quantization through weight equalization and bias correction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1325–1334 (2019)
34. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
36. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
37. Przewlocka-Rus, D., Sarwar, S.S., Sumbul, H.E., Li, Y., De Salvo, B.: Power-of-two quantization for low bitwidth and hardware compliant neural networks. arXiv preprint arXiv:2203.05025 (2022)
38. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
39. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. Advances in neural information processing systems **30** (2017)
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
41. Rücklé, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., Gurevych, I.: Adapterdrop: On the efficiency of adapters in transformers. arXiv preprint arXiv:2010.11918 (2020)
42. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023)
43. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning, `https://github.com/cloneofsimo/lora`
44. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)
45. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)

46. Shang, Y., Yuan, Z., Xie, B., Wu, B., Yan, Y.: Post-training quantization on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1972–1981 (2023)
47. Wu, D., Tang, Q., Zhao, Y., Zhang, M., Fu, Y., Zhang, D.: Easyquant: Post-training quantization via scale optimization. arXiv preprint arXiv:2006.16669 (2020)
48. Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., Han, S.: Smoothquant: Accurate and efficient post-training quantization for large language models. In: International Conference on Machine Learning. pp. 38087–38099. PMLR (2023)
49. Yao, Z., Yazdani Aminabadi, R., Zhang, M., Wu, X., Li, C., He, Y.: Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. Advances in Neural Information Processing Systems **35**, 27168–27183 (2022)
50. Yeh, R.A., Chen, C., Yian Lim, T., Schwing, A.G., Hasegawa-Johnson, M., Do, M.N.: Semantic image inpainting with deep generative models. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5485–5493 (2017)