

TerDiT: Ternary Diffusion Models with Transformers

Xudong Lu, Aojun Zhou, Ziyi Lin, Qi Liu, Yuhui Xu, Renrui Zhang, Xue Yang, *Member, IEEE*,
 Junchi Yan, *Senior Member, IEEE*, Peng Gao, Hongsheng Li, *Member, IEEE*



Fig. 1. Sample images (256×256) generated by TerDiT with 4.2B parameters (using 2GB of GPU memory) are shown. For comparison, images generated by full-precision diffusion transformer models—DiT-XL/2 with 675M parameters (using 3GB of GPU memory) and Large-DiT-4.2B with 4.2B parameters (using 17GB of GPU memory)—are provided in Fig. 5.

Abstract—Recent developments in large-scale pre-trained text-to-image diffusion models have significantly improved the generation of high-fidelity images, particularly with the emergence of diffusion transformer models (DiTs). Among diffusion models, diffusion transformers have demonstrated superior image-generation capabilities, boosting lower FID scores and higher scalability. However, deploying large-scale DiT models can be expensive due to their excessive parameter numbers. Although existing research has explored efficient deployment techniques for diffusion models, such as model quantization, there is still little work concerning DiT-based models. To tackle this research gap, we propose TerDiT, the first quantization-aware training (QAT) and efficient deployment scheme for extremely low-bit diffusion transformer models. We focus on the ternarization of DiT networks, with model sizes ranging from 600M to 4.2B, and image resolution from 256×256 to 512×512 . Our work contributes to the exploration of efficient deployment of large-scale DiT models, demonstrating the feasibility of training extremely low-bit DiT models from scratch while maintaining competitive image generation capacities compared to full-precision models. Our code and pre-trained TerDiT checkpoints have been released at <https://github.com/Lucky-Lance/TerDiT>.

Index Terms—Model quantization, quantization-aware training, image generation, diffusion transformer models

I. INTRODUCTION

The advancements in large-scale pre-trained text-to-image diffusion models [1, 2, 3, 4, 5, 6] have led to the successful generation of images characterized by both complexity and

high fidelity to the input conditions. Notably, the emergence of diffusion transformer models (DiTs) [7] represents a significant stride in this research direction. Compared with other diffusion models, diffusion transformers have demonstrated the capability to achieve lower FID scores but with higher computation GFLOPS [7]. Recent research highlights the remarkable image and video generation capabilities of diffusion transformer architectures, as demonstrated in methods like Stable Diffusion 3 [8] and Sora [9]. Given the impressive performance of diffusion transformer models, researchers are now training larger and larger DiTs [10]. For instance, Stable Diffusion 3 trained DiT models range from 800M to 8B parameters. Additionally, there is speculation among researchers that Sora might boast around 3 billion parameters. Given the enormous parameter numbers, deploying these DiT models will be costly, especially on certain end devices (e.g., mobile phones).

To deal with the deployment dilemma, there are recent works on the efficient deployment of diffusion models [11, 12, 13, 14], most of which focus on model quantization. However, as far as we are concerned, there are still two main shortcomings in current research. Firstly, the exploration of quantization methods for transformer-based diffusion models remains quite limited [15, 16] compared to quantizing U-Net-based diffusion models. Secondly, most prevailing quantization approaches heavily rely on post-training quantization (PTQ) [11, 13, 17, 18, 19], which leads to unacceptable performance degradation,

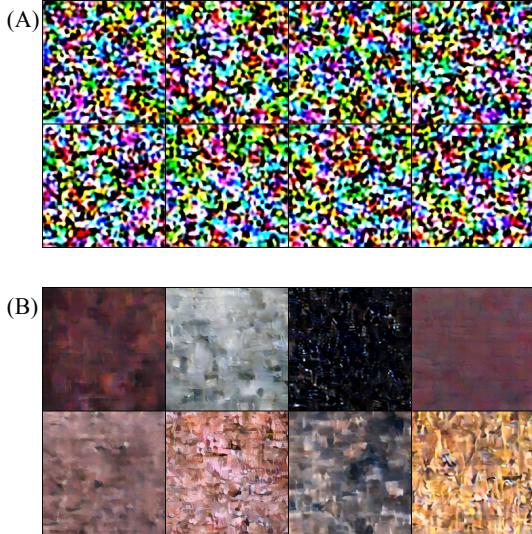


Fig. 2. 2-bit Q-DiT (A) and 2-bit Q-Diffusion (B) quantization results.

particularly with extremely low bit width (e.g., 2-bit and 1-bit). For example, the 2-bit weight quantization results of Q-DiT [15] and Q-Diffusion [11] are shown in Fig. 2. The extremely low-bit quantization of neural networks is essential as it can significantly reduce the computation resources for deployment [20], especially for huge models. During our research, we find that there is still no work considering the extremely low-bit quantization of DiT models.

To tackle these shortcomings, we propose leveraging the quantization-aware training (QAT) technique for the extremely low-bit quantization of large-scale DiT models. Low-bit QAT methods for large-scale models have been discussed in the LLM domain. Recent works have observed that training large language models with extremely low-bit parameters (e.g., binary and ternary) from scratch can also lead to competitive performance [21, 22] compared to their full-precision counterparts. Furthermore, there have been recent advancements in deploying ternary LLMs on CPUs [23], resulting in accelerated inference and reduced energy consumption. These findings highlight the promising potential of ternary quantization and suggest that significant precision redundancy still exists in large-scale models, making QAT a feasible approach for large-scale DiTs.

Inspired by the ongoing in-depth research on ternary LLMs [22, 23], in this paper, we introduce the first attempt at ternary quantization [24] for the DiT model and provide **TerDiT**, the first quantization scheme for extremely low-bit DiTs to our best knowledge. Our method achieves quantization-aware training (weight only) and efficient deployment for ternary diffusion transformer models. Different from the naive quantization of linear layers in LLMs and CNNs [22, 24], we find that the direct weight ternarization of the adaLN module [25] in DiT blocks [7, 26] leads to large dimension-wise scale and shift values in the normalization layer compared with full-precision models (due to weight quantization, gradient approximation, etc), which results in slower convergence speed and poor model performance. Consequently, we propose a variant of adaLN by applying an RMS Norm [27] after the ternary linear layers of the adaLN module to mitigate this training issue effectively.

With this modification, we scale the ternary DiT model from 600M (size of DiT-XL/2 [7]) to 4.2B (size of Large-DiT-4.2B [28]), with image resolution from 256×256 to 512×512 . We further deploy the trained ternary DiT models with 2-bit CUDA kernels, resulting in over $10 \times$ reduction in checkpoint size and about $8 \times$ reduction in inference memory consumption while achieving competitive (or even better) generation quality compared with full-precision models. The contributions of our work are summarized as follows:

1) Inspired by the quantization-aware training scheme for low-bit LLMs, we study the QAT method for ternary DiT models and introduce DiT-specific techniques for achieving extremely low-bit quantization for DiT models.

2) We scale ternary DiT models from 600M to 4.2B parameters, with resolution from 256×256 to 512×512 , and further deploy the trained ternary DiT on GPUs based on 2-bit CUDA kernels, enabling the inference of a 4.2B DiT with less than 2GB GPU memory (256×256 resolution).

3) Competitive evaluation results compared with full-precision models and baseline quantization methods on the ImageNet [29] benchmark (image generation) showcase the effectiveness of TerDiT.

TerDiT is the first attempt to explore the extremely low-bit quantization of DiT models. We focus on quantization-aware training and efficient deployment for large ternary DiT models, offering valuable insights for future research on deploying extremely low-bit DiT models.

II. RELATED WORKS

A. Diffusion Models

Diffusion models have gained significant attention in recent years due to their ability to generate high-quality images and their potential for various applications. The diffusion model was first introduced by [30], proposing a generative model that learns to reverse a diffusion process. This work laid the foundation for subsequent research in the field. [1] further extended the idea by introducing denoising diffusion probabilistic models (DDPMs), which have become a popular choice for image generation tasks. DDPMs have been applied to a wide range of domains, including unconditional image generation [1], image inpainting [31], and image super-resolution [32]. Additionally, diffusion models have been used for text-to-image synthesis, as demonstrated by the DALL-E model [33] and the Imagen model [5]. These models showcase the ability of diffusion models to generate highly realistic and diverse images from textual descriptions. Furthermore, diffusion models have been extended to other modalities, such as audio synthesis [34], video generation [35], and 3D generation [36], demonstrating their versatility and potential for multimodal applications.

B. Quantization of Diffusion Models

The quantization of diffusion models has been studied in recent years to improve the efficiency of diffusion models. Post-training quantization (PTQ) methods, such as those presented in [11, 13, 15, 16, 17, 18, 19], offer advantages in terms of quantization time and data usage. However, these methods

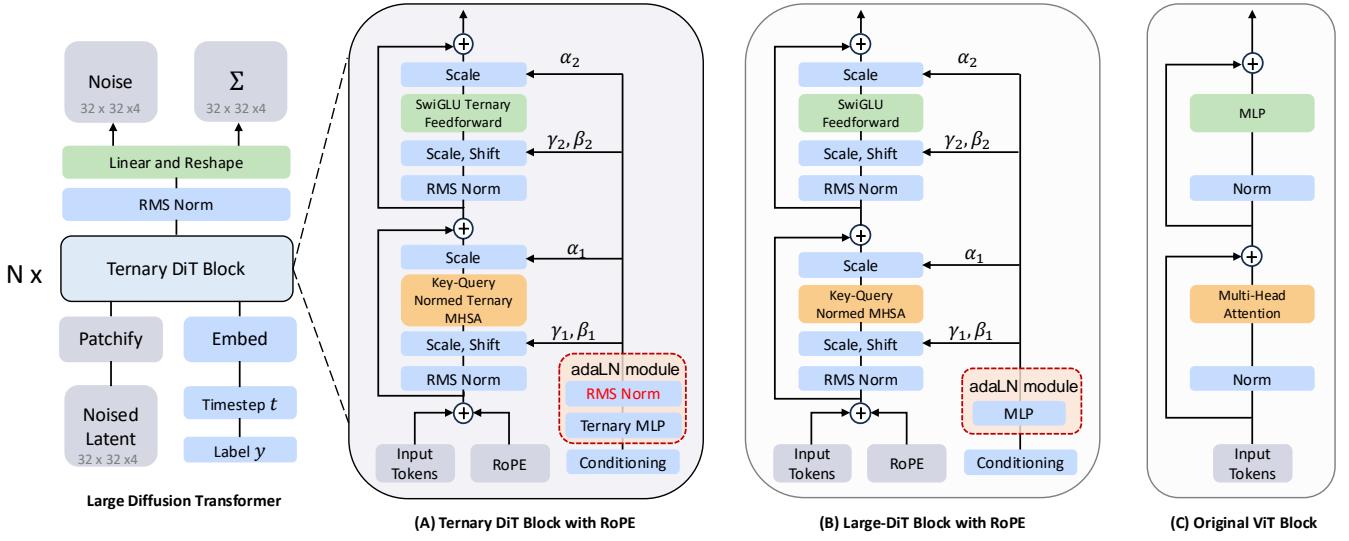


Fig. 3. Model structure comparison between (A) Ternary DiT block, (B) Large-DiT block, and the (C) original ViT block. The Large-DiT (DiT) block adds an adaLN module to the original ViT block for condition injection. Ternary DiT block further adds an RMS Norm in the adaLN module for better ternarization-aware training.

often result in suboptimal performance when applied to low-bit settings. To address this issue, [37] proposes combining quantization-aware low-rank adapters (QALoRA) with PTQ methods, leading to improved evaluation results. As an alternative to PTQ, quantization-aware training (QAT) methods have been introduced specifically for low-bit diffusion model quantization [14, 38, 39, 40]. Despite their effectiveness, these QAT methods are currently only limited to small-sized U-Net-based diffusion models, revealing a research gap in applying QAT to large-scale DiT models. Further exploration of QAT techniques for large DiT models with extremely low bit width could potentially unlock even greater efficiency gains and enable the effective deployment of diffusion models in resource-constrained environments.

C. Ternary Weight Networks

Ternary weight networks [24] have emerged as a memory-efficient and computation-efficient network structure, offering the potential for significant reductions in inference memory usage. When supported by specialized hardware, ternary weight networks can also deliver substantial computational acceleration. Among quantization methods, ternary weight networks have garnered notable attention, with two primary approaches being explored: weight-only quantization and weight-activation quantization. In weight-only quantization, as discussed in [41], solely the weights are quantized to ternary values. On the other hand, weight-activation quantization, as presented in [42, 43], involves quantizing both the weights and activations to ternary values. Recent research has demonstrated the applicability of ternary weight networks to the training of large language models [22], achieving results comparable to their full-precision counterparts. Moreover, recent developments have enabled the deployment of ternary LLMs on CPUs [23], leading to fast inference and low energy consumption. Building upon these advancements, our work introduces, for the first time, quantization-aware training and efficient deployment schemes specifically designed for ternary DiT models. By leveraging

the benefits of ternary quantization in the context of DiT models, we aim to push the boundaries of efficiency and enable the deployment of powerful diffusion models in resource-constrained environments, opening up new possibilities for practical applications.

III. TERDiT

In this section, we introduce TerDiT, a framework designed to conduct weight-only quantization-aware training and efficient deployment of large-scale ternary DiT models. We first give a brief review of diffusion transformer (DiT) models in Sec. III-A. Then, building upon the previous open-sourced Large-DiT [28], we illustrate the quantization function and quantization-aware training scheme in Sec. III-B, conduct QAT-specific model structure improvement for better network training in Sec. III-C, and introduce the ternary deployment scheme in Sec. III-D.

A. Preliminary: Diffusion Transformer Models

Diffusion Transformer. Diffusion transformer [7] (DiT) is an architecture that replaces the commonly used U-Net backbone in the diffusion models with a transformer that operates on latent patches. Similar to the Vision Transformer (ViT) architecture shown in Fig. 3 (C), DiT first patches the spatial inputs into a sequence of tokens, and then the denoising process is carried out through a series of transformer blocks (Fig. 3 (B)). To deal with additional conditional information (e.g., noise timesteps t , class labels l , natural language inputs), DiT leverages adaptive normalization modules [44] (adaLN-Zero) to insert these extra conditional inputs to the transformer blocks. After the final transformer block, a standard linear decoder is applied to predict the final noise and covariance. The DiT models can be trained in the same way as U-Net-based diffusion models.

AdaLN Module in DiT. The main difference between DiT and traditional ViT is the need to inject conditional information for image generation. DiT employs a zero-initialized adaptive layer normalization (adaLN-Zero) module in each transformer

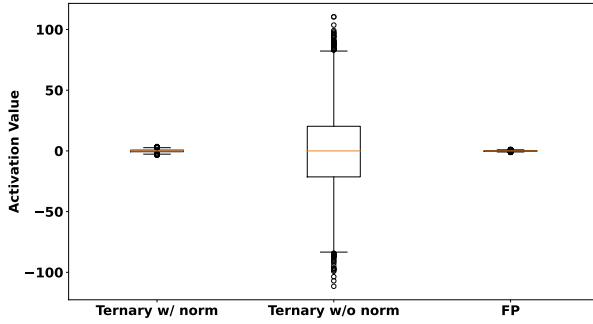


Fig. 4. Activation value analysis. We compare activation values passing through a ternary weight linear layer with and without RMS Norm, using a full-precision linear layer as a reference. The ternary linear layer without RMS Norm results in extremely large activation values, introducing instability in neural network training. However, when the normalization layer is applied, the activation values are scaled to a reasonable range, similar to those observed in the full-precision layer.

block, as shown in the red part of Fig. 3 (B), which calculates the dimension-wise scale and shift values from the input condition c :

$$\text{adaLN}(c) = \text{MLP}(\text{SiLU}(c)). \quad (1)$$

AdaLN is an important component in the DiT model [7]. Within the DiT architecture, the adaLN module integrates an MLP layer with a substantial number of parameters, constituting approximately 10% to 20% of the model’s total parameters. Throughout the training of TerDiT, we observe that the direct weight-ternarization of this module yields undesirable training results (analyzed in Sec. III-C).

B. Model Quantization

As illustrated in Sec. I, there is an increasing popularity in understanding the scaling law of DiT models, which has been proven crucial for developing and optimizing LLMs. In recent explorations, Large-DiT [28] successfully scales up the model parameters from 600M to 7B by incorporating the methodologies of LLaMA [45, 46] and DiT. The results demonstrate that parameter scaling can potentially enhance model performance and improve convergence speed for the label-conditioned ImageNet generation task. Motivated by this, we propose to further investigate the ternarization of DiT models, which can alleviate the challenges associated with deploying large-scale DiT models. In this subsection, we introduce the quantization function and quantization-aware training scheme.

Quantization Function. To construct a ternary weight DiT network, we replace all the linear layers in self-attention, feedforward, and MLP of the original Large-DiT blocks with ternary linear layers, obtaining a set of ternary DiT blocks (Fig. 3 (A)). For ternary linear layers, we adopt an *absmean* quantization function similar to BitNet b1.58 [22]. First, the weight matrix is normalized by dividing each element by the average absolute value of all the elements in the matrix. After normalization, each value in the weight matrix is rounded to the nearest integer and clamped into the set $\{-1, 0, +1\}$.

Referring to current popular quantization methods for LLMs [47, 48], we also multiply a learnable scaling parameter

α (randomly initialized) to each ternary linear matrix after quantization, leading to the final value set as $\{-\alpha, 0, +\alpha\}$. The quantization function is formulated as:

$$\widetilde{W} = \alpha \cdot \text{RoundClip}\left(\frac{W}{\gamma + \epsilon}, -1, 1\right), \quad (2)$$

where ϵ is set to a small value (e.g., 10^{-6}) to avoid division by 0, and

$$\text{RoundClip}(x, a, b) = \text{Clamp}(\text{round}(x), a, b), \quad (3)$$

$$\gamma = \frac{1}{mn} \sum_{ij} |W_{ij}|. \quad (4)$$

TerDiT is a weight-only quantization scheme, and we do not quantize the activations.

Quantization-aware Training Scheme. Based on the above-designed quantization function, we train a DiT model from scratch¹ utilizing the straight-through estimator (STE) [49], allowing gradient propagation through the undifferentiable network components. We preserve the full-precision parameters of the network throughout the training process. For each training step, ternary weights are calculated from full-precision parameters by the ternary quantization function in the forward pass, and the gradients of ternary weights are directly applied to the full-precision parameters for parameter updates in the backward pass.

However, we find the convergence speed is very slow. Even after many training iterations, the loss cannot be decreased to a reasonable range. We find that this issue may arise from the trait that ternary linear layers usually cause large activation values, and propose to tackle the problem with QAT-specific model structure improvement in the following subsection.

C. QAT-specific Model Structure Improvement

Activation Analysis for Ternary Linear Layer. In a ternary linear layer, all the parameters take one value from $\{-\alpha, 0, +\alpha\}$. The values passing through this layer would become large activation values, hampering the stable training of neural networks. We conduct a pilot study to qualitatively demonstrate the impact of ternary linear weights on activation values.

We randomly initialize a ternary linear layer with the input feature dimension set to 1024 and the output feature dimension to 9216 (corresponding to the linear layer of the adaLN module in Large-DiT). The weight parameters pass through the quantization function and receive a 512×1024 sized matrix input (filled with 1). The box plot of activation distribution is shown in the center part of Fig. 4. We also calculate the activation distribution after passing the matrix through a full-precision linear layer generated with the same random seed, shown on the right part of Fig. 4. As can be seen, the ternary linear layer leads to very large activation values compared with full-precision linear layers.

The large-activation problem brought about by the ternary linear weights can be alleviated by applying a layer norm to the output of the ternary linear layer. We add an RMS Norm

¹It is observed in [22] that for ternary LLMs, the conversion or post-training quantization from trained LLMs does not help. So we also train ternary DiT models from scratch.

TABLE I
COMPARISON BETWEEN TERDiT AND FULL-PRECISION DIFFUSION MODELS ON THE IMAGENET DATASET.

| ImageNet 256x256 Benchmark | | | | | | |
|-----------------------------------|------------|-------|--------|-------------------|-------------|----------|
| Models | Images (M) | FID ↓ | sFID ↓ | Inception Score ↑ | Precision ↑ | Recall ↑ |
| BigGAN-deep [50] | - | 6.95 | 7.36 | 171.40 | 0.87 | 0.28 |
| StyleGAN-XL [51] | - | 2.30 | 4.02 | 265.12 | 0.78 | 0.53 |
| ADM [52] | 507 | 10.94 | 6.02 | 100.98 | 0.69 | 0.63 |
| ADM-U [52] | 507 | 7.49 | 5.13 | 127.49 | 0.72 | 0.63 |
| LDM-8 [4] | 307 | 15.51 | - | 79.03 | 0.65 | 0.63 |
| LDM-4 [4] | 213 | 10.56 | - | 103.49 | 0.71 | 0.62 |
| DiT-XL/2 (675M) [7] | 1792 | 9.62 | 6.85 | 121.50 | 0.67 | 0.67 |
| TerDiT-4.2B | 604 | 9.66 | 6.75 | 117.54 | 0.66 | 0.68 |
| Classifier-free Guidance | | | | | | |
| ADM-G [52] | 507 | 4.59 | 5.25 | 186.70 | 0.82 | 0.52 |
| ADM-G, ADM-U [52] | 507 | 3.94 | 6.14 | 215.84 | 0.83 | 0.53 |
| LDM-8-G [4] | 307 | 7.76 | - | 209.52 | 0.84 | 0.35 |
| LDM-4-G [4] | 213 | 3.60 | - | 247.67 | 0.87 | 0.48 |
| DiT-XL/2-G (675M) [7] | 1792 | 2.27 | 4.60 | 278.24 | 0.83 | 0.57 |
| Large-DiT-4.2B-G [28] | 435 | 2.10 | 4.52 | 304.36 | 0.82 | 0.60 |
| TerDiT-600M-G | 448 | 4.34 | 4.99 | 183.49 | 0.81 | 0.54 |
| TerDiT-4.2B-G | 604 | 2.42 | 4.62 | 263.91 | 0.82 | 0.59 |
| ImageNet 512x512 Benchmark | | | | | | |
| ADM [52] | 1385 | 23.24 | 10.19 | 58.06 | 0.73 | 0.60 |
| ADM-U [52] | 1385 | 9.96 | 5.62 | 121.78 | 0.75 | 0.64 |
| ADM-G [52] | 1385 | 7.72 | 6.57 | 172.71 | 0.87 | 0.42 |
| ADM-G, ADM-U [52] | 1385 | 3.85 | 5.86 | 221.72 | 0.84 | 0.53 |
| DiT-XL/2-G [7] | 768 | 3.04 | 5.02 | 240.82 | 0.84 | 0.54 |
| Large-DiT-4.2B-G [28] | 472 | 2.52 | 5.01 | 303.70 | 0.82 | 0.57 |
| TerDiT-4.2B-G | 696 | 2.81 | 4.96 | 267.86 | 0.84 | 0.55 |

(similar to LLaMA) after the ternary linear layer and obtain the activation distribution in the left part of Fig. 4. In this case, the activation values are scaled to a reasonable range after passing the normalization layer and lead to more stable training behavior. The observation also aligns with [21], where a layer normalization function is applied before the activation quantization for each quantized linear layer.

RMS Normalized AdaLN Module. We analyze the DiT model for QAT-specific model structure improvement based on the above insights. In a standard ViT transformer block, the layer norm is applied to every self-attention layer and feedforward layer. This is also the case with the self-attention layers and feedforward layers in the DiT block, which can help to properly scale the range of activations. However, the DiT block differs from traditional transformer blocks due to the presence of the adaLN module, as introduced in Sec. III-A. Notably, there is no layer normalization applied to this module. In the context of full-precision training, the absence of layer normalization does not have a significant impact. However, for ternary DiT networks, its absence can result in large dimension-wise scale and shift values in the adaLN (normalization) module, posing bad influences on model training. To mitigate this issue, we introduce an RMS Norm after the MLP layer of the adaLN module in each ternary DiT block:

$$\text{adaLN}(c) = \text{RMS}(\text{MLP}(\text{SiLU}(c))), \quad (5)$$

and the final model structure of TerDiT is illustrated in Fig. 3 (A). This minor modification can result in faster convergence speed and lower training loss, leading to better quantitative and qualitative evaluation results. To better showcase the effect,

the **actual** activation distribution with/without the RMS Norm after model training is analyzed in Sec. IV-D, which shows similar distribution features.

Remark: What distinguishes TerDiT from classic low-bit quantization methods like BitNet b1.58 [22]?

In BitNet b1.58, a BitLinear layer is designed, which applies a layer norm to the input activation of the Linear layer. BitNet then replaces the Linear layers in LLaMA with BitLinear layers and removes the RMS Norm before the attention and SwiGLU layers. Similar approaches can be found in earlier works like Xnor-net [53], and Bi-real net [54] to construct a quantized layer. Different from applying layer norm to all the BitLinear layers of the model, TerDiT focuses on **simple yet effective** modifications to the model structure by simply adding an RMS Norm within the adaLN module of each DiT block. With fewer norms added, our method leads to faster training speed and better evaluation scores. Comparative experiments are introduced in Sec. IV-B.

D. Deployment Scheme

Although there have been recent advancements in deploying ternary LLMs on CPUs [23], there are currently no effective open-source deployment solutions for ternary DiT models. This requires joint efforts from both the academic and engineering communities. To demonstrate the performance and deployment efficiency of ternary DiT and provide a reference for further advancing deployment on hardware, we deploy the trained networks with a 2-bit implementation. To be specific, we pack the ternary linear weights to INT8 values (4 ternary numbers into one INT8 number) with the `pack_2bit_u8()`

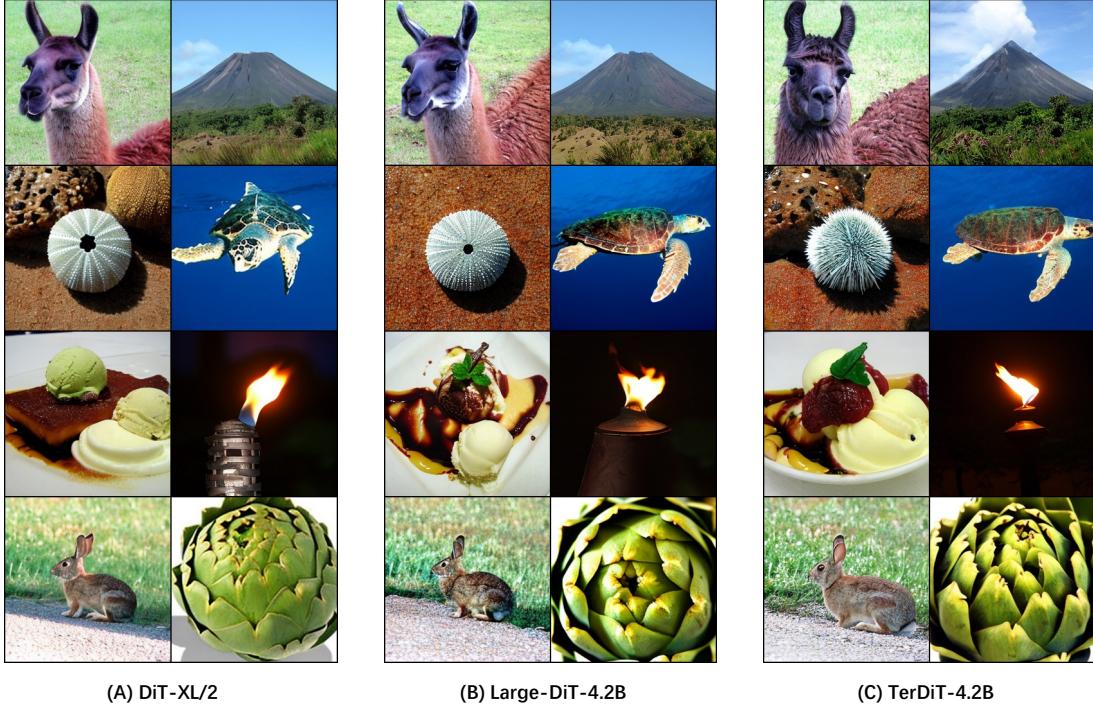


Fig. 5. Qualitative results analysis (256×256). We compare images generated by DiT-XL/2 (A), Large-DiT-4.2B (B), and TerDiT-4.2B (C) with class labels [355, 980, 328, 33, 928, 862, 330, 944] and $\text{cfg}=4$. TerDiT-4.2B generates images of the same visual quality as two full-precision DiT models.

function provided by [55]. During the inference process of the DiT model, we call the complementary `unpack_2bit_u8()` function on the fly to recover packed 2-bit numbers to FP values, then perform subsequent calculations. The addition of the unpacking operation will slow down the inference process, but we believe that with further research into model ternarization, more hardware support for inference speedup will be available.

IV. EXPERIMENTS

In this section, a set of experiments are designed to evaluate TerDiT. We compare TerDiT with full-precision diffusion models in Sec. IV-A, with other quantization methods (PTQ and QAT) in Sec. IV-B, carry out deployment efficiency comparison in Sec. IV-C, illustrate the effectiveness of the RMS Normalized adaLN module in Sec. IV-D, and carry out more ablation studies in Sec. IV-E. To further enrich the content of the paper, we show the generation results with high cfg values in Sec. IV-F and provide additional qualitative results in Sec. IV-G to visually demonstrate the capabilities of TerDiT. Our DiT implementation is based on the open-sourced code of Large-DiT-ImageNet [28]². We conduct experiments on ternary DiT models with 600M (size of DiT-XL/2) and 4.2B³ (size of Large-DiT-4.2B) parameters, respectively.

A. Comparison with Full-precision Models

We provide quantitative and qualitative comparison results between TerDiT and representative full-precision models in this subsection.

²<https://github.com/Alpha-VLLM/LLaMA2-Accessory/tree/main/Large-DiT-ImageNet>

³The provided 3B model in the Large-DiT-ImageNet repository actually has 4.2B parameters.

Experiment Setup. Following the evaluation setting of the original DiT paper [7], we train 600M and 4.2B ternary DiT models on the ImageNet dataset. We start from training the 256×256 image resolution model (600M and 4.2B) and continue to train the 512×512 resolution model (4.2B) based on the 256×256 checkpoint. We compare TerDiT with a series of full-precision diffusion models, report FID [56], sFID [57], Inception Score, Precision, and Recall (50k generated images) following [52]. We also provide the total number of images (million) during the training stage as in [28] to offer further insights into the convergence speed of different models.

Training Details. For the 256×256 resolution model, we train the 600M TerDiT model on 8 A100-80G GPUs for 1750k steps with batchsize set to 256, and the 4.2B model on 16 A100-80G GPUs for 1180k steps with batchsize set to 512. We set the initial learning rate as $5e-4$ ⁴. After 1550k steps of training for the 600M and 550k steps for the 4.2B model, we reduce the learning rate back to $1e-4$ for more fine-grained parameter updates.

For 512×512 resolution, we start from the 4.2B trained 256×256 ternary model (trained with 604M images) and continue to train it on ImageNet with 512×512 resolution and only 92M images. The learning rate is set to $1e-4$.

Quantitative Results Analysis. The evaluation results are listed in Tab. I. TerDiT is a QAT scheme for DiT models, so among all the full-precision models, we pay special attention to DiT-XL/2 (675M) and Large-DiT-4.2B. (Large-DiT-4.2B and TerDiT-4.2B-G share the exact same architecture.)

In the 256×256 case, without classifier-free guidance, TerDiT-4.2B achieves very similar testing results with DiT-

⁴The default learning rate of DiT is $1e-4$. For TerDiT, we increase the initial learning rate to $5e-4$ following [21] that a larger learning rate is needed for the faster convergence of low-bit QAT.

TABLE II
DEPLOYMENT EFFICIENCY COMPARISON WITH CFG=1.5.

| | Resolution | Checkpoint Size | Max Memory Allocated | Inference Time | FID |
|------------------|------------|-----------------|----------------------|----------------|------|
| DiT-XL/2-G | 256 | 2.6GB | 3128.42MB | 15s | 2.27 |
| DiT-XL/2-G | 512 | 2.6GB | 3728.61MB | 80s | 3.04 |
| Large-DiT-4.2B-G | 256 | 16GB | 17027.29MB | 83s | 2.10 |
| Large-DiT-4.2B-G | 512 | 16GB | 18614.29MB | 365s | 2.52 |
| TerDiT-600M-G | 256 | 168M | 762.30MB | 20s | 4.34 |
| TerDiT-4.2B-G | 256 | 1.1GB | 1919.67MB | 97s | 2.42 |
| TerDiT-4.2B-G | 512 | 1.1GB | 3506.67MB | 376s | 2.81 |

XL/2 (with much fewer training images). With classifier-free guidance (cfg=1.5), TerDiT-4.2B-G outperforms LDM-G while bringing a very slight performance degradation compared with two full-precision DiT-structured models. Besides, TerDiT-4.2B-G achieves better evaluation results than TerDiT-600M-G, implying that models with more parameters can incur smaller performance degradation after quantization. In the more commonly used 512×512 setting, TerDiT-4.2B-G outperforms DiT-XL/2-G in all aspects. It also surpasses Large-DiT-4.2B-G in terms of sFID and Precision. This result demonstrates the generative capabilities of TerDiT compared with full-precision models with the same architecture and parameter numbers.

Qualitative Results Analysis. To visually demonstrate the effectiveness of TerDiT, we also show some qualitative comparison results in Fig. 5 (256×256), concerning TerDiT-4.2B, DiT-XL/2, and Large-DiT-4.2B. In terms of visual perception, there is no significant difference between the images generated by TerDiT and those by the full-precision models.

B. Extremely Low-bit Quantization Baselines

The topic of our paper is to explore an algorithm for the ternary (1.58-bit) quantization of extremely low-bit DiT models. In this section, we make a comparison between TerDiT and PTQ/QAT baselines.

PTQ Baselines. We choose PTQ algorithms Q-DiT [15] for DiT models and Q-Diffusion [11] for U-Net-based models and perform 2-bit weight quantization. We find they both fail to generate images, detailed examples are shown in Fig. 2.

QAT Baselines.⁵ We adapt BitNet b1.58 [21, 22] for the DiT model (256×256, 600M parameters). We also use EfficientDM [37] for a more comprehensive analysis. For BitNet, we replace the linear layers in DiT (which correspond to TerDiT) with BitLinear and remove the norms before attention and SwiGLU layers. We then train BitNet using the same process as TerDiT and measure the FID score (↓), which yields 6.60 for BitNet and 4.34 for TerDiT. An intuitive explanation for the phenomenon is that adding more norms can stabilize the training, but it may also slow down the convergence speed. Moreover, due to the additional norms, the training speed of BitNet drops to 0.9× that of TerDiT, while inference latency increases to 1.15×. This demonstrates the efficiency of our proposed TerDiT. For EfficientDM, we apply 2-bit quantization and train the quantized model, but it fails to generate normal images, as detailed in Fig. 6.

⁵We do not find code for replicating Q-DM [39] and they do not provide the ImageNet 256×256/512×512 results.



Fig. 6. Generation results of EfficientDM with 2-bit weight-only quantization.

C. Deployment Efficiency Comparison

The improvement in deployment efficiency is the motivation of our proposed TerDiT scheme. In this subsection, we provide a comparison between TerDiT-600M/4.2B, DiT-XL/2, and Large-DiT-4.2B to discuss the actual deployment efficiency TerDiT can bring about. Tab. II shows the checkpoint sizes of the four DiT models. We also record the memory usage and inference time of the total diffusion sample loop (step = 250) on a single A100-80G GPU.

From the table, we can see that TerDiT greatly reduces checkpoint size and memory usage. The checkpoint size and memory usage of the 4.2B ternary DiT model are significantly smaller than those of Large-DiT-4.2B, even smaller than DiT-XL/2. This brings significant advantages to deploying the model on end devices (e.g., mobile phones, FPGA). However, the absence of open-source software deployment frameworks and efficient hardware support for ternary DiTs leads to slower inference speeds compared to their full-precision counterparts. Despite this, recent engineering advancements in deploying ternary LLMs on CPUs [23] suggest that the computational benefits of ternary-weight networks will become more apparent as software and hardware co-design continue to evolve.

D. Ablation Study: RMS Normalized AdaLN Module

The main modification of TerDiT to the structure of the DiT model is the addition of an RMS Norm after the MLP in the adaLN module. In this part, we compare with the baseline ternary model to demonstrate the influence of RMS Norm on both the training process and the training outcomes.

Experiment Setup. We train ternary DiT models with 600M and 4.2B parameters on the ImageNet [29] dataset in 256×256 resolution. For each parameter size, we train two models, one with RMS Norm in the adaLN module and one without. We record the loss curves during training and measure the FID-50k score (cfg=1.5) every 100k training steps.

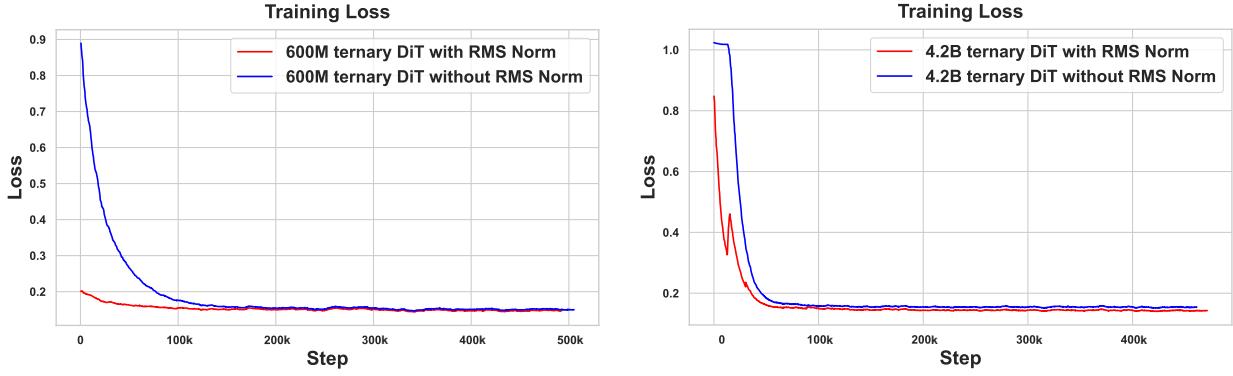


Fig. 7. Training loss comparison of ternary DiT models with/without RMS Norm in the adaLN module. We show the loss curves training both 600M (left) and 4.2B (right) DiT models. Adding the RMS Norm will lead to faster convergence speed and lower training loss.

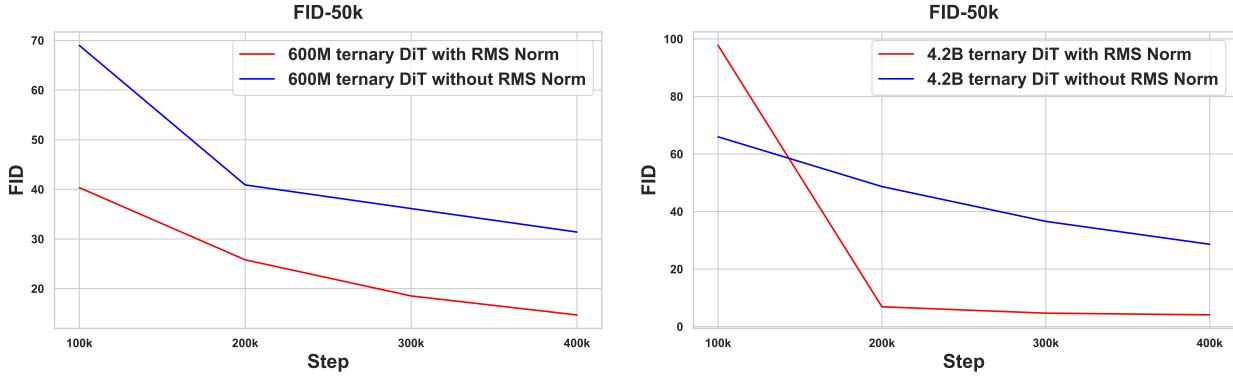


Fig. 8. FID-50k score comparison on class-conditional ImageNet 256×256 generation task (cfg=1.5) with/without RMS Norm for both 600M and 4.2B ternary DiT models (100k steps to 400k steps). Training with RMS Norm will lead to lower FID scores.

Training Details. For a fair comparison, we train all the ternary DiT models on 8 A100-80G GPUs with batchsize set to 256. The learning rate is set to 5e-4 throughout training.

Quantitative Results Analysis. The training loss⁶ and evaluation scores are shown in Fig. 7 and Fig. 8 respectively. As can be seen, training with the RMS Normalized adaLN module will lead to faster convergence speed and lower FID scores. Another observation is that models with more parameters tend to achieve faster and better training compared to models with fewer parameters. This also, to some extent, reflects the scaling law of the ternary DiT model.

Qualitative Results Analysis. We compare the TerDiT-600M and TerDiT-4.2B models trained with/without RMS Norm in the adaLN module. We sample images from the models trained for 400k steps and also show the results of the fully trained models. The comparisons are demonstrated in Fig. 10. We can come to two conclusions:

1) For both the 600M and 4.2B TerDiT model, training with the RMS Normalized adaLN module will lead to better qualitative results.

2) Models with more parameters show more learning ability and can achieve better training results compared to models with fewer parameters.

Activation Distribution after Training. We also conduct an analysis of the activation distribution during inference after model training as a supplementary for the experiment and

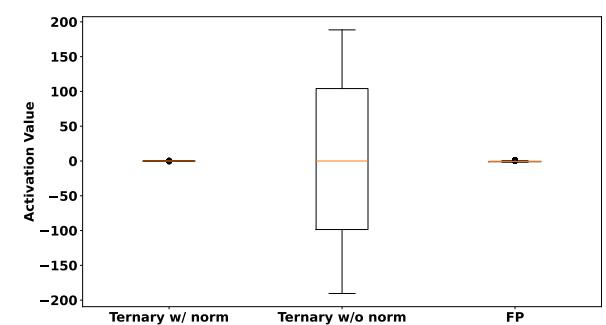


Fig. 9. Activation value analysis. We train the TerDiT-4.2B model with/without RMS Norm in the adaLN module for 50k steps and show the activation distribution of the ‘scale_mlp’ output in the second ternary DiT block during inference (at the first sampling step). The activation distribution of the original full-precision model is also provided.

explanations in Sec. III-C. We train the TerDiT-4.2B model (256×256) with/without RMS Norm in the adaLN module for 50k steps and then analyze the activation distribution of the ‘scale_mlp’ output in the adaLN module during inference, specifically at the first sampling step within the second ternary DiT block. For comparison, we also calculate the activation distribution of the original full-precision Large-DiT-4.2B model at the same layer. As shown in Fig. 9, training with RMS Norm can help limit the range of the activation values.

E. Ablation Study: Learning Rate Reduction

In Sec. IV-A, we adopt a learning rate reduction after training for certain steps for more fine-grained parameter updates. Here,

⁶The training process of diffusion models is not as ‘smooth’ as one might assume. To better illustrate the training dynamics, we employ exponential smoothing (with a smoothing factor of 0.995) during visualization.

TABLE III
EFFECTIVENESS OF LEARNING RATE REDUCTION.

| ImageNet 256×256 Benchmark, Classifier-free Guidance | | | | | | |
|--|------|-------|--------|-------------------|-------------|----------|
| Models | LR | FID ↓ | sFID ↓ | Inception Score ↑ | Precision ↑ | Recall ↑ |
| TerDiT-600M-G (cfg=1.50) | 1e-4 | 4.34 | 4.99 | 183.49 | 0.81 | 0.54 |
| TerDiT-600M-G (cfg=1.50) | 5e-4 | 6.38 | 5.00 | 147.79 | 0.76 | 0.54 |

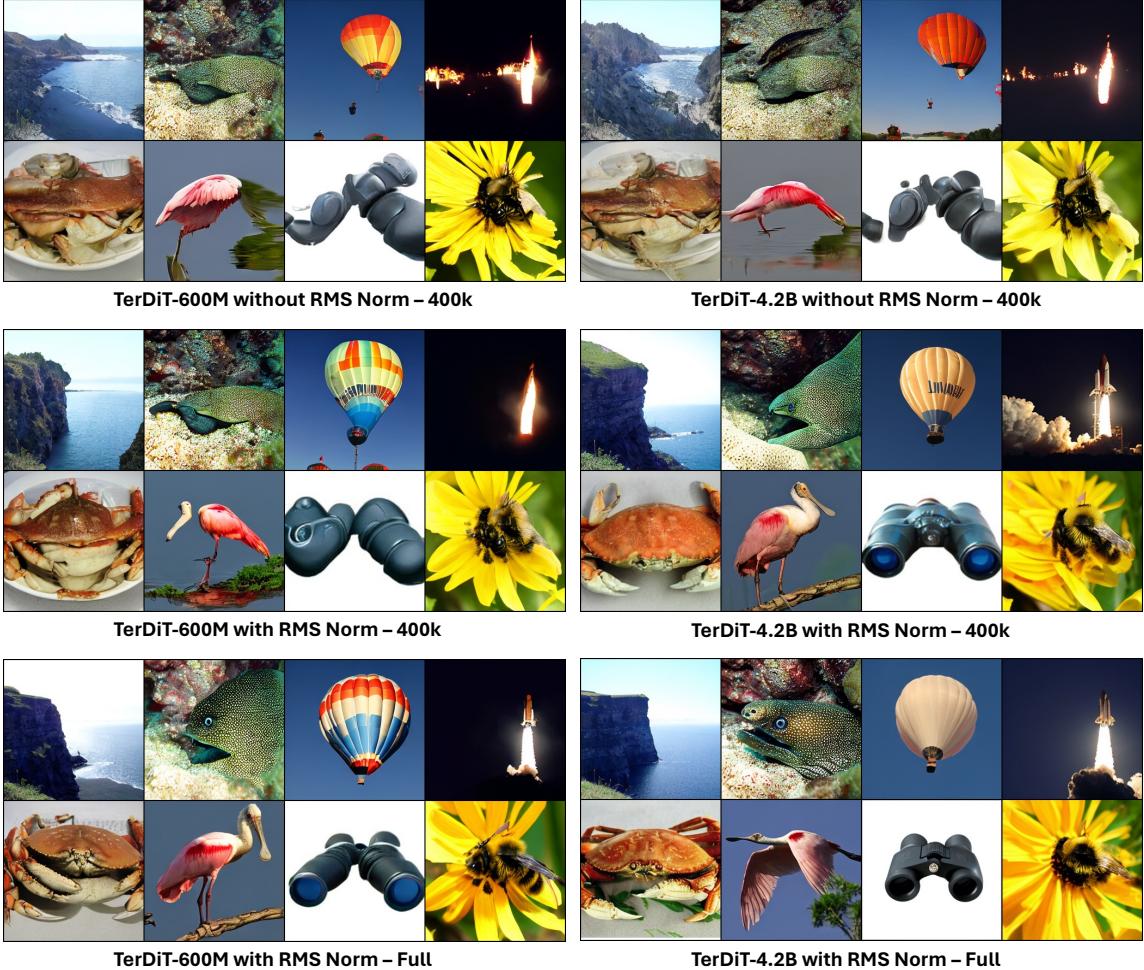


Fig. 10. Qualitative results comparison for 600M and 4.2B ternary DiT models with/without RMS Norm in the adaLN module. We choose class labels [972, 390, 417, 812, 118, 129, 447, 309] with cfg=4.



Fig. 11. Qualitative comparison with/without lr reduction using TerDiT-600M. We choose class label [207, 360, 387, 974, 88, 979, 417, 279] and cfg=4.

we provide an ablation study on this learning rate reduction.

We choose the TerDiT-600M model (256×256) for convenience. Following the setting of Sec. IV-A, we train a TerDiT-600M model with the RMS Normalized adaLN module and

5e-4 learning rate for 1550k steps. We then continue training this model with 1e-4/5e-4 for another 200k steps.

Quantitative Results. We evaluate FID, sFID, Inception Score, Precision, and Recall of these two models. As can be



Fig. 12. TerDiT-4.2B (256×256) with $\text{cfg}=4$ (left) and $\text{cfg}=10$ (right), class label [89, 475, 978, 971, 508, 32, 963, 235].



Fig. 13. TerDiT-600M (left) and TerDiT-4.2B (right), class label [257, 300, 975, 973, 478, 388, 427, 809], $\text{cfg}=4$.



Fig. 14. TerDiT-4.2B, class label [89, 475, 978, 971, 508, 32, 963, 235] (left) and [257, 300, 975, 973, 478, 388, 427, 809] (right), $\text{cfg}=4$.

seen in Tab. III, the reduction in the learning rate will lead to better evaluation results.

Qualitative Results. We also provide qualitative results comparison on TerDiT-600M (256×256) with/without learning rate reduction in Fig. 11.

The quality comparison highlights the importance of learning rate reduction in the later stages of training.

F. Image Quality with High CFG Values

In our experiments above, we use a cfg scale of 4 for qualitative evaluations. Setting the cfg scale too high can result in images that are less creative and of lower quality. In this section, we test the robustness of TerDiT when handling high cfg scale. We provide sample images with $\text{cfg}=4$ and $\text{cfg}=10$ in Fig. 12. Although there is slight distortion, quantization has not severely impacted the generation quality at high cfg values, demonstrating the robustness of TerDiT to extreme cases.

G. More Qualitative TerDiT Results

In this subsection, we provide more generation results of TerDiT with 256×256 resolution in Fig. 13, with 512×512 resolution in Fig. 14. TerDiT is capable of generating images with good visual quality.

V. DISCUSSIONS AND FUTURE WORKS

In this paper, we propose quantization-aware training (QAT) and efficient deployment methods for large-scale ternary DiT models. Competitive evaluation results on the ImageNet generation task (256×256 and 512×512) compared with full-precision models and baseline quantization methods demonstrate the feasibility of training a large ternary DiT from scratch while achieving promising generation results. To our best knowledge, this work is also the first study concerning the extremely low-bit quantization of DiT models. In this section, we give more explanations and discussions.

Firstly, training ternary DiT models is less stable and more time-consuming than training full-precision networks. Although we discuss methods to enhance training stability by adding norms, it still requires more time than training full-precision networks, leading to increased carbon dioxide emissions during model training in a broader context. We will explore more DiT structures [58, 59] and hardware-software co-development to increase training speed.

Secondly, in our paper, we do not make a complete comparison between ternary quantization and INT8/FP16 quantization, as INT8 and FP16 do not fall under the category of extremely low-bit quantization, which is outside the comparison

scope of our paper. On the one hand, hardware and software support for INT8 and FP16 is mature, with many hardware chips and software libraries readily available [47, 48, 60, 61]. In contrast, research on extremely low-bit quantization for LLMs or Stable Diffusion remains in its early stages. On the other hand, while FP16 or INT8 can achieve strong performance, they reduce memory usage by only up to 75%. In theory, ternary quantization can reduce memory usage by up to 16 \times . Therefore, it is undeniable that ternary quantization holds greater potential compared to FP16 or INT8. In fact, in addition to deploying LLMs on CPUs [23], hardware and software accelerations for ternary (binary) CNNs have already been implemented on FPGA [62, 63, 64]. We will continue to explore the hardware acceleration implementation of DiT in future work.

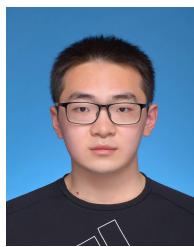
REFERENCES

- [1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [2] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, “Cascaded diffusion models for high fidelity image generation,” *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 2249–2281, 2022.
- [3] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents, 2022,” URL <https://arxiv.org/abs/2204.06125>, vol. 7, 2022.
- [4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [5] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *arXiv preprint arXiv:2205.11487*, 2022.
- [6] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, “Diffusion models in vision: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [7] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
- [8] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, “Scaling rectified flow transformers for high-resolution image synthesis,” *arXiv preprint arXiv:2403.03206*, 2024.
- [9] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, “Video generation models as world simulators,” 2024. [Online]. Available: <https://openai.com/research/video-generation-models-as-world-simulators>
- [10] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao *et al.*, “Sora: A review on background, technology, limitations, and opportunities of large vision models,” *arXiv preprint arXiv:2402.17177*, 2024.
- [11] X. Li, Y. Liu, L. Lian, H. Yang, Z. Dong, D. Kang, S. Zhang, and K. Keutzer, “Q-diffusion: Quantizing diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 535–17 545.
- [12] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, Y. Wang, S. Tulyakov, and J. Ren, “Snapfusion: Text-to-image diffusion model on mobile devices within two seconds,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [13] Y. He, L. Liu, J. Liu, W. Wu, H. Zhou, and B. Zhuang, “Ptqd: Accurate post-training quantization for diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [14] H. Wang, Y. Shang, Z. Yuan, J. Wu, and Y. Yan, “Quest: Low-bit diffusion model quantization via efficient selective finetuning,” *arXiv preprint arXiv:2402.03666*, 2024.
- [15] L. Chen, Y. Meng, C. Tang, X. Ma, J. Jiang, X. Wang, Z. Wang, and W. Zhu, “Q-dit: Accurate post-training quantization for diffusion transformers,” *arXiv preprint arXiv:2406.17343*, 2024.
- [16] J. Deng, S. Li, Z. Wang, H. Gu, K. Xu, and K. Huang, “Vq4dit: Efficient post-training vector quantization for diffusion transformers,” *arXiv preprint arXiv:2408.17131*, 2024.
- [17] Y. Shang, Z. Yuan, B. Xie, B. Wu, and Y. Yan, “Post-training quantization on diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1972–1981.
- [18] Y. Yang, X. Dai, J. Wang, P. Zhang, and H. Zhang, “Efficient quantization strategies for latent diffusion models,” *arXiv preprint arXiv:2312.05431*, 2023.
- [19] C. Wang, Z. Wang, X. Xu, Y. Tang, J. Zhou, and J. Lu, “Towards accurate data-free quantization for diffusion models,” *arXiv preprint arXiv:2305.18723*, vol. 2, no. 5, 2023.
- [20] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, “Efficient processing of deep neural networks: A tutorial and survey,” *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.
- [21] H. Wang, S. Ma, L. Dong, S. Huang, H. Wang, L. Ma, F. Yang, R. Wang, Y. Wu, and F. Wei, “Bitnet: Scaling 1-bit transformers for large language models,” *arXiv preprint arXiv:2310.11453*, 2023.
- [22] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, “The era of 1-bit llms: All large language models are in 1.58 bits,” *arXiv preprint arXiv:2402.17764*, 2024.
- [23] J. Wang, H. Zhou, T. Song, S. Mao, S. Ma, H. Wang, Y. Xia, and F. Wei, “1-bit ai infra: Part 1.1, fast and lossless bitnet b1. 58 inference on cpus,” *arXiv preprint arXiv:2410.16144*, 2024.
- [24] F. Li, B. Liu, X. Wang, B. Zhang, and J. Yan, “Ternary weight networks,” *arXiv preprint arXiv:1605.04711*, 2016.

- [25] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “Film: Visual reasoning with a general conditioning layer,” in *AAAI Conference on Artificial Intelligence*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:19119291>
- [26] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie, “Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers,” 2024.
- [27] B. Zhang and R. Sennrich, “Root mean square layer normalization,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/1e8a19426224ca89e83cef47f1e7f53b-Paper.pdf
- [28] P. Gao, L. Zhuo, Z. Lin, C. Liu, J. Chen, R. Du, E. Xie, X. Luo, L. Qiu, Y. Zhang *et al.*, “Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers,” *arXiv preprint arXiv:2405.05945*, 2024.
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [30] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” *arXiv preprint arXiv:1503.03585*, 2015.
- [31] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv preprint arXiv:2011.13456*, 2020.
- [32] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” *arXiv preprint arXiv:2104.07636*, 2021.
- [33] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” *arXiv preprint arXiv:2102.12092*, 2021.
- [34] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, “Wavegrad: Estimating gradients for waveform generation,” *arXiv preprint arXiv:2009.00713*, 2020.
- [35] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” *arXiv preprint arXiv:2204.03458*, 2022.
- [36] Z. Cao, F. Hong, T. Wu, L. Pan, and Z. Liu, “Diffusion++: 3d-aware diffusion transformer for large-vocabulary 3d generation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [37] Y. He, J. Liu, W. Wu, H. Zhou, and B. Zhuang, “Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models,” *arXiv preprint arXiv:2310.03270*, 2023.
- [38] X. Zheng, H. Qin, X. Ma, M. Zhang, H. Hao, J. Wang, Z. Zhao, J. Guo, and X. Liu, “Binarydm: Towards accurate binarization of diffusion model,” *arXiv preprint arXiv:2404.05662*, 2024.
- [39] Y. Li, S. Xu, X. Cao, X. Sun, and B. Zhang, “Q-dm: An efficient low-bit quantized diffusion model,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [40] H. Chang, H. Shen, Y. Cai, X. Ye, Z. Xu, W. Cheng, K. Lv, W. Zhang, Y. Lu, and H. Guo, “Effective quantization for diffusion models on cpus,” *arXiv preprint arXiv:2311.16133*, 2023.
- [41] C. Zhu, S. Han, H. Mao, and W. J. Dally, “Trained ternary quantization,” *arXiv preprint arXiv:1612.01064*, 2016.
- [42] H. Alemdar, V. Leroy, A. Prost-Boucle, and F. Pérot, “Ternary neural networks for resource-efficient ai applications,” in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2547–2554.
- [43] P. Wang, Q. Hu, Y. Zhang, C. Zhang, Y. Liu, and J. Cheng, “Two-step quantization for low-bit neural networks,” in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2018, pp. 4376–4384.
- [44] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
- [45] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [46] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [47] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “Gptq: Accurate post-training quantization for generative pre-trained transformers,” *arXiv preprint arXiv:2210.17323*, 2022.
- [48] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and S. Han, “Awq: Activation-aware weight quantization for llm compression and acceleration,” *arXiv preprint arXiv:2306.00978*, 2023.
- [49] Y. Bengio, N. Léonard, and A. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *arXiv preprint arXiv:1308.3432*, 2013.
- [50] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv preprint arXiv:1809.11096*, 2018.
- [51] A. Sauer, K. Schwarz, and A. Geiger, “Stylegan-xl: Scaling stylegan to large diverse datasets,” in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.
- [52] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.
- [53] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” in *European conference on computer vision*. Springer, 2016, pp. 525–542.
- [54] Z. Liu, B. Wu, W. Luo, X. Yang, W. Liu, and K.-T. Cheng, “Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced

- training algorithm,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 722–737.
- [55] H. Badri and A. Shaji, “Half-quadratic quantization of large machine learning models,” November 2023. [Online]. Available: https://mobiusml.github.io/hqq_blog/
- [56] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [57] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016.
- [58] H. Lu, G. Yang, N. Fei, Y. Huo, Z. Lu, P. Luo, and M. Ding, “Vdt: General-purpose video diffusion transformers via mask modeling,” *arXiv preprint arXiv:2305.13311*, 2023.
- [59] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu *et al.*, “Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis,” *arXiv preprint arXiv:2310.00426*, 2023.
- [60] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, “Mixed precision training,” *arXiv preprint arXiv:1710.03740*, 2017.
- [61] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, “Llm. int8 (): 8-bit matrix multiplication for transformers at scale. corr abs/2208.07339 (2022),” 2022.
- [62] S. Zhu, L. H. Duong, H. Chen, D. Liu, and W. Liu, “Fat: An in-memory accelerator with fast addition for ternary weight neural networks,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 42, no. 3, pp. 781–794, 2022.
- [63] R. Zhao, W. Song, W. Zhang, T. Xing, J.-H. Lin, M. Srivastava, R. Gupta, and Z. Zhang, “Accelerating binarized convolutional neural networks with software-programmable fpgas,” in *Proceedings of the 2017 ACM/SIGDA international symposium on field-programmable gate arrays*, 2017, pp. 15–24.
- [64] G. Rutishauser, J. Mihali, M. Scherer, and L. Bonini, “xtern: Energy-efficient ternary neural network inference on risc-v-based edge systems,” in *2024 IEEE 35th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 2024, pp. 206–213.

VI. BIOGRAPHY SECTION



Xudong Lu is a Ph.D. student at the Multimedia Laboratory (MMLab), The Chinese University of Hong Kong. He obtained his Bachelor of Engineering degree from Shanghai Jiao Tong University. His research interests include model compression, large language models, and multimodal large language models. He has served as a reviewer for IEEE TCSVT, CVPR, NeurIPS, ICLR, ICML and ICCV.



Ajun Zhou received the M.S. degree from the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Science, Beijing, China, in 2019. He is currently working toward the Ph.D. degree with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong. His research interests include deep learning and computer vision.



Ziyi Lin is pursuing a Ph.D. degree in Electronic Engineering at the Chinese University of Hong Kong. Prior to this, he obtained a B.E. degree in Computer Science at Beihang University, Beijing, China. His research was also in close collaboration with SenseTime Group Ltd. and Shanghai AI Laboratory. His research interests include video recognition and multi-modal generative modeling.



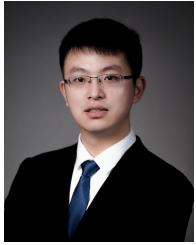
Qi Liu is a Master’s student at the ReThinkLab, Shanghai Jiao Tong University. He earned a Bachelor of Engineering in Computer Science and Technology (major) and a Bachelor of Science in Mathematics and Applied Mathematics (minor) from Shanghai Jiao Tong University. His research interests span formal verification, large language model reasoning, and efficient machine learning. Liu has served as a reviewer for Transactions on Machine Learning Research (TMLR).



Yuhui Xu is a Research Scientist at Salesforce AI Research. He earned his B.S. from Southeast University in 2016 and his Ph.D. from Shanghai Jiao Tong University in 2021. Dr. Xu’s research focuses on deep learning, with a particular emphasis on efficient deployment of deep learning models. His expertise spans model quantization, pruning, and automated machine learning. Dr. Xu was awarded the prestigious CSIG (China Society of Image and Graphics) PhD Fellowship in 2023. His work aims to bridge the gap between cutting-edge deep learning algorithms and their practical applications, driving innovation in AI efficiency and accessibility.



Renrui Zhang received the B.S. degree from the School of Electronic Engineering and Computer Science at Peking University, Beijing, China, in 2021. He is currently a Ph.D candidate in Multimedia Laboratory (MMLab) at The Chinese University of Hong Kong, Hong Kong, China, supervised by Professor Hongsheng Li and Xiaogang Wang. His research interests include large language models, multi-modality learning, and 3D vision.



Xue Yang (Member, IEEE) is currently an Assistant Professor with the Department of Automation, Shanghai Jiao Tong University, Shanghai, China. Before that, he received the Ph.D. in Computer Science from Shanghai Jiao Tong University, Shanghai, China, in 2023, the M.S. degree from Chinese Academy of Sciences University, Beijing, China, in 2019, and the B.E. degree from Automation, Central South University, Hunan, China, in 2016. His research interests are computer vision and machine learning. He published first-authored papers in IEEE TPAMI, IJCV, CVPR, ECCV, ICCV, NeurIPS, ICML, ICLR, etc., and won the most influential AAAI'21 paper. He was Area Chair for ICLR.

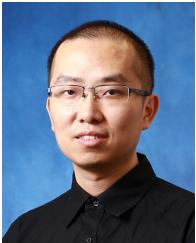


Junchi Yan (Senior Member, IEEE) is a Professor and Deputy Director with School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai, China. Before that, he was a Senior Research Staff Member with IBM Research where he started his career since April 2011. He is IAPR/IET Fellow, and his research interest includes machine learning and vision. He served Area Chair for conferences including ICLR, ICML, NeurIPS, SIGKDD, CVPR, AAAI, IJCAI, etc., and Associate Editor of IEEE TPAMI. His work was selected as CVPR 2024 best

paper candidate.



Peng Gao is currently a research scientist at Shanghai Artificial Intelligence Laboratory. His research involves AIGCs, vision language models and large language models. He has published extensively on top-tier conferences and journals including CVPR, NeurIPS, ICML and IJCV.



Hongsheng Li (Member, IEEE) is an associate professor with the Department of Electronic Engineering, Chinese University of Hong Kong and Co-affiliated to Multimedia Laboratory.