Optimizing Quantized Diffusion Models via Distillation with Cross-Timestep Error Correction

Yanxi Li*, **Chengbin Du***

School of Computer Science, University of Sydney, Australia
yali0722@uni.sydney.edu.au, chdu5632@uni.sydney.edu.au

Abstract

Diffusion models (DMs) have attracted attention in generative modeling due to their ability to produce high-quality, diverse outputs by progressively adding noise to data and then denoising it. However, DMs are computationally intensive due to their iterative nature, requiring numerous forward passes and high-precision operations, making them less efficient for resource-constrained environments. Recent efforts to reduce these computational demands using quantization show promise by converting high-precision parameters to lower precision, but they face challenges unique to DMs, particularly in addressing cross-timestep error propagation in the iterative process. In this paper, we analyze cross-timestep error propagation in quantized DMs, revealing that previous methods focusing only on reducing noise estimation discrepancies are insufficient. Instead, we introduce Cross-Timestep Error Correction (CTEC), where the quantized model not only approximates the full-precision model but also corrects errors from the previous timestep. A distillation method is applied to learn this correction process effectively. We conduct extensive experiments on unconditional image generation with LSUN-Churches and LSUN-Bedrooms, as well as conditional image generation with ImageNet. Our findings demonstrate the effectiveness of our method in significantly reducing accumulated quantization errors across timesteps within the quantized diffusion process. This enhancement enables the generation of high-quality images, even when constrained by reduced bitwidths.

Introduction

Diffusion models (DMs) (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021; Ho and Salimans 2021; Rombach et al. 2022) have recently attracted significant attention in the field of generative modeling, particularly due to their success in producing high-quality outputs. The core process of DMs involves a forward diffusion process, where noise is progressively added to the data, followed by a reverse denoising process that removes this noise (Sohl-Dickstein et al. 2015), which enables the generation of realistic and diverse samples. Moreover, DMs effectively address challenges encountered by previous generative models, such as training instability and mode collapse, thereby providing a more stable and reliable alternative. As a result, DMs have become a new benchmark in the generative modeling landscape.

However, despite their remarkable performance, the generation process using DMs is computationally intensive and resource-demanding. The iterative nature of these models requires numerous forward passes, each involving deep neural network computations, which cumulatively result in significant computational overhead. Even though prior attempts have successfully decreased the number of timesteps from 1,000 to just a few dozen (Song, Meng, and Ermon 2021), sampling with DMs remains significantly slower compared to earlier non-iterative methods, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) and Variational Autoencoders (VAEs) (Kingma and Welling 2013). Additionally, the need for high-precision operations (FP32 or FP16) to maintain the quality of generated data also presents a substantial barrier to deploying DMs in resource-constrained environments.

Recently, there are some attempts (Shang et al. 2023; Li et al. 2023a; He et al. 2023; Li et al. 2023b; So et al. 2023; Huang et al. 2024; He et al. 2024) to leverage the benefits of quantization (Nagel et al. 2019, 2020; Li et al. 2021) to reduce the computational costs of DMs. Quantization is a well-established technique employed to reduce the computational burden of neural networks by converting high-precision (e.g., FP32 or FP16) parameters and activations to lower precision (e.g., INT8 or INT4). This process typically involves quantizing the weights and activations of a pre-trained model without requiring extensive retraining, thus enabling more efficient inference on hardware with limited computational resources.

One remaining challenge is that quantization methods for non-iterative models are not directly applicable to DMs, as they fail to account for the unique characteristics of DMs, particularly the cross-timestep error propagation inherent in their iterative generation process. Although there are existing methods for temporal-aware quantization (Li et al. 2023b; So et al. 2023; Huang et al. 2024) and error-aware calibration (Tang et al. 2023), they still only focus on minimizing errors within individual timesteps rather than directly correcting cross-timestep error propagation.

In this paper, we conduct a thorough analysis of cross-timestep error propagation in quantized DMs. Our findings

*These authors contributed equally.

reveal that previous methods, which focus solely on minimizing the discrepancy between the noise estimation of the quantized model ($\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t)$) and that of the full-precision model ($\epsilon_{\theta}^{(t)}(\mathbf{x}_t)$), are insufficient to minimize the discrepancy between the generated samples ($\mathbf{x}'_{t-1}$ and $\mathbf{x}_{t-1}$). When generating samples for a given timestep $t-1$, the quantized noise estimation model $\epsilon_{\theta'}$ not only approximates the full-precision noise estimation model ϵ_θ but also learns to correct the quantization error propagated from the previous timestep t. Since errors at timestep t are corrected in the following timestep $t-1$, this process is named as ***cross-timestep error correction*** (CTEC). Considering that the traditional calibration method is insufficient for learning the error correction term, an efficient distillation method is applied to learn this correction.

We conduct a series of experiments utilizing Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2021) and Latent Diffusion Models (LDMs) (Rombach et al. 2022) for both unconditional image generation with LSUN-Churches and LSUN-Bedrooms, as well as conditional image generation with ImageNet, all at a resolution of 256×256 pixels. Our experimental results indicate that the proposed error correction term effectively mitigates the accumulated quantization error across timesteps within the quantized diffusion process. This enhancement enables the generation of high-quality images even under constraints of limited bitwidths. Importantly, in scenarios with low bitwidths (e.g., W4/A4 for the LSUN datasets and W2/A8 for ImageNet), where quantization errors are more pronounced, our method demonstrates a good capacity to correct these accumulated errors, thereby preserving the overall quality of image generation.

Related Works

Diffusion Models

Diffusion Models (DMs) (Sohl-Dickstein et al. 2015) introduce a framework where data distributions are modeled through iterative forward and reverse diffusion processes, inspired by non-equilibrium thermodynamics. Building on this, Denoising Diffusion Probabilistic Models (DDPM) (Ho, Jain, and Abbeel 2020) achieve state-of-the-art image synthesis but required many steps for sampling. To enhance efficiency, Denoising Diffusion Implicit Models (DDIM) (Song, Meng, and Ermon 2021) allow for faster sampling without sacrificing quality by introducing non-Markovian diffusion processes. Further advancements included classifier-free guidance (Ho and Salimans 2021), which eliminated the need for external classifiers while maintaining a balance between sample diversity and fidelity. Improved Diffusion (Nichol and Dhariwal 2021) optimizes the reverse diffusion process, enabling faster sampling and competitive performance. Latent Diffusion Models (LDMs) (Rombach et al. 2022) apply diffusion in the latent space of pretrained autoencoders. This approach significantly reduced computational requirements while maintaining high visual fidelity, enabling state-of-the-art performance in high-resolution image synthesis and other tasks.

Model Quantization

Gupta et al. (2015) showed that deep networks could be trained with 16-bit precision using stochastic rounding, with minimal accuracy loss. Jacob et al. (2018) extended this by enabling integer-only inference, significantly reducing memory use and enhancing performance on ARM CPUs. Learned Step Size Quantization (LSQ) (Esser et al. 2020) introduces a method to learn the quantization step size during training and maintains accuracy even at extreme low-bit levels. Nagel et al. (2019) propose a data-free method that achieves near-original performance with 8-bit quantization. AdaRound (Nagel et al. 2020) further improves post-training quantization (PTQ) accuracy by adapting rounding methods to minimize task loss. BRECQ (Li et al. 2021) enables 2-bit quantization by reconstructing the basic building blocks of neural networks one by one.

Quantized Diffusion Models

Several efforts have been made to harness the potential of quantization to enhance the efficiency of DMs. PTQ4DM (Shang et al. 2023) introduces a DM-specific PTQ method that leverages insights into the quantized operations, calibration dataset, and metric selection, enabling the quantization of full-precision DMs to 8-bit while maintaining or improving performance. Q-Diffusion (Li et al. 2023a) tackles the bimodal activation distributions and changing output distributions over time by introducing timestep-aware calibration and split shortcut quantization. PTQD (He et al. 2023) proposes a unified approach to correct quantization noise in DMs, utilizing a mixed-precision scheme to optimize bitwidths during different denoising steps. Q-DM (Li et al. 2023b) identifies oscillation and error accumulation issues in low-bit quantized DMs and introduces Timestep-aware Quantization (TaQ) and Noise-estimating Mimicking (NeM) to mitigate these effects. TDQ (So et al. 2023) innovates by dynamically adjusting the quantization interval based on timestep information. TFMQ-DM (Huang et al. 2024) proposes Temporal Feature Maintenance Quantization, which optimizes temporal features separately, ensuring high compression efficiency and generation quality. EfficientDM (He et al. 2024) presents a hybrid approach combining PTQ and QAT techniques to achieve QAT-level performance with PTQ-like efficiency.

Methodology

Preliminaries

Diffusion Models. Diffusion models represent a sophisticated class of generative models characterized by their iterative noise introduction and denoising processes. These models operate through a two-phase mechanism: the forward diffusion process and the reverse denoising process (Sohl-Dickstein et al. 2015).

In the forward process, Gaussian noise is incrementally added to a data sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, forming a sequence of noisy latents $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. This can be mathematically modeled as a Markov chain where each step t involves sampling from a normal distribution:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where $\beta_t \in (0, 1)$ defines the variance schedule that controls the noise intensity at each step.

The reverse process aims to reconstruct the original data $\mathbf{x}_0$ from the noisy latent $\mathbf{x}_T$, which approximates an isotropic Gaussian distribution as T approaches infinity. Since the true reverse distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is intractable, it is approximated using a neural network parameterized distribution:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \quad (2)$$

where μ_θ and Σ_θ are mean and variance functions. Typically, μ_θ is a learnable function parameterized by deep neural networks. In contrast, $\Sigma_\theta(\mathbf{x}_t, t)$ is set to $\sigma_t^2\mathbf{I}$ to simplify the model, based on the assumption of isotropic Gaussian noise (Ho, Jain, and Abbeel 2020).

Quantization Calibration. In quantization, a pre-trained model is quantized without further training. The quantization parameters are determined using a calibration dataset that approximates the model's data distribution. The objective is to minimize the quantization error, typically measured by metrics ℓ like Mean Squared Error (MSE), cosine distance, or Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{quant}} = \mathbb{E}_{\mathbf{y}',\mathbf{y}}\left[\ell(\mathbf{y}', \mathbf{y})\right], \quad (3)$$

where $\mathbf{y}'$ and $\mathbf{y}$ are the outputs from the quantized model and the full-precision model, respectively. The optimization process aims to find the quantization parameters that minimize $\mathcal{L}_{\text{quant}}$.

Quantization Error Propagation in DM

Firstly of all, we conduct a thorough analysis of the quantization error propagation in DMs. Due to the advancements that DDIM (Song, Meng, and Ermon 2021) introduces over DDPM (Ho, Jain, and Abbeel 2020) by optimizing the sampling process through non-Markovian diffusion processes, DDIM has gained increasing popularity in recent research (Rombach et al. 2022; He et al. 2024). Therefore, our analysis focuses on DDIM. As DDIM generalizes DDPM, it is feasible to extend the analysis conducted for DDIM to DDPM, and vice versa.

DDIM employs two major components for predicting the denoised observation. The first component is responsible for predicting the original, clean data sample x_0 from the noisy observation x_t at timestep t, which is formulated as. The prediction of x_0 is given by:

$$f_\theta^{(t)}(\mathbf{x}_t) := \frac{\mathbf{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\mathbf{x}_t)}{\sqrt{\alpha_t}}, \quad (4)$$

where $\epsilon_\theta^{(t)}(\cdot)$ is a noise prediction model parameterized by θ and α_t is the noise schedule parameter at the current timestep t. The second component focuses on determining the direction in which the next timestep x_{t-1} should move relative to the current timestep x_t. This is done by calculating a directional vector that points toward x_t, which is essential for guiding the reverse diffusion process. The direction is given by:

$$\Delta_\theta^{(t)}(\mathbf{x}_t) := \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(\mathbf{x}_t), \quad (5)$$

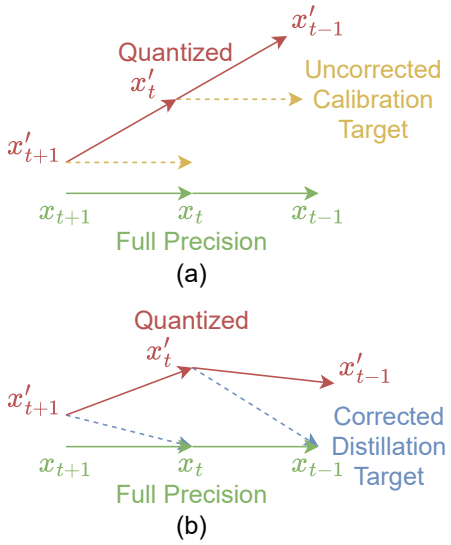


Figure 1: The uncorrected and our corrected distillation target. The solid arrows represent denoising steps. Since only $\epsilon_{\theta'}$ is optimized, in the uncorrected case, the target is the yellow dashed arrow, which is paralleled to the denoising step performed by the full precision model. In our corrected case, we define the target as an vector pointed from $\mathbf{x}'_t$ to $\mathbf{x}_{t-1}$ represented by the blue dashed arrow.

where σ_t is the standard deviation of the noise added at each step and α_{t-1} is the noise schedule parameter for the previous timestep $t-1$. DDIM combines these two components to generate the denoised observation $\mathbf{x}_{t-1}$ from the current noisy observation $\mathbf{x}_t$:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \cdot f_\theta^{(t)}(\mathbf{x}_t) + \Delta_\theta^{(t)}(\mathbf{x}_t) + \sigma_t\epsilon, \quad (6)$$

where ϵ is random Gaussian noise.

A common approach to quantization is uniform quantization, where a floating-point vector $\mathbf{x}$ is quantized using a scale factor s and a zero point z and clamping within specified bounds $[l, u]$:

$$\hat{\mathbf{x}} = \text{Quant}(\mathbf{x}) = \text{clamp}\left(\left\lfloor \frac{\mathbf{x}}{s} \right\rceil - z, l, u\right), \quad (7)$$

where $\lfloor\cdot\rceil$ denotes the rounding operation, and $\text{clamp}(\cdot, l, u)$ ensures the values lie within the target range. The dequantization process is

$$\mathbf{x}' = \text{Deq}(\hat{\mathbf{x}}) = (\hat{\mathbf{x}} + z) \cdot s. \quad (8)$$

Such transformations introduce quantization errors, which need to be minimized to preserve the model's performance. Upon quantization, the process in Eq. 6 is modified to:

$$\mathbf{x}'_{t-1} = \sqrt{\alpha_{t-1}} \cdot f_{\theta'}^{(t)}(\mathbf{x}'_t) + \Delta_{\theta'}^{(t)}(\mathbf{x}'_t) + \sigma_t\epsilon, \quad (9)$$

where θ' represents the quantized network weights, and $\mathbf{x}'_t$ and $\mathbf{x}'_{t-1}$ denote the quantized denoised outputs from the previous timestep and the current timestep, respectively. The

transition from $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ and the introduction of quantization lead to discrepancies due to the approximations introduced by quantization. These discrepancies manifest as errors, which can accumulate over multiple timesteps, affecting the model's performance and the accuracy of the denoised outputs. Understanding and mitigating these quantization errors are crucial for improving the robustness and fidelity of the DDIM framework.

Based on Eqs. 6 and 9, the discrepancy between the full precision value $\mathbf{x}_{t-1}$ and its quantized counterpart $\mathbf{x}'_{t-1}$ can be expressed as follows:

$$\|\mathbf{x}_{t-1} - \mathbf{x}'_{t-1}\| = \|A_t(\mathbf{x}_t - \mathbf{x}'_t) + B_t(\epsilon_\theta^{(t)}(\mathbf{x}_t) - \epsilon_{\theta'}^{(t)}(\mathbf{x}'_t))\| \quad (10)$$

where

$$A_t = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}, \quad (11)$$

$$B_t = \sqrt{1 - \alpha_{t-1} - \sigma_t^2} - \sqrt{\frac{\alpha_{t-1}(1-\alpha_t)}{\alpha_t}}. \quad (12)$$

The comprehensive derivation process for Eq. 10 is provided in the supplementary material.

According to Eq. 10, there are two primary sources of error. The first source is the difference between $\mathbf{x}_t$ and $\mathbf{x}'_t$, while the second source is the difference between $\epsilon_\theta^{(t)}(\mathbf{x}_t)$ and $\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t)$. This implies that even if θ' is meticulously calibrated to ensure that $\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t)$ closely approximates $\epsilon_\theta^{(t)}(\mathbf{x}_t)$, the error originating from the previous step, specifically the term $\mathbf{x}_t - \mathbf{x}'_t$, still exists. Additionally, calibration alone cannot guarantee that $\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t)$ and $\epsilon_\theta^{(t)}(\mathbf{x}_t)$ will be perfectly identical, thus leaving a residual error. This residual error consequently results in a divergence between $\mathbf{x}'_{t-1}$ and $\mathbf{x}_{t-1}$. When such errors are accumulated over multiple timesteps, they contribute to a substantial cumulative quantization error, as illustrated in Fig. 1 (a).

Cross-Timestep Error Correction (CTEC)

Building upon the preceding analysis, rather than directly optimizing the calibration target in Eq. 3, which in this context can be expressed as $\ell(\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t), \epsilon_\theta^{(t)}(\mathbf{x}_t))$, we propose a corrected target. By setting the discrepancy in Eq. 10 to zero and reformulating the equation $\left\|A_t(\mathbf{x}_t - \mathbf{x}'_t) + B_t(\epsilon_\theta^{(t)}(\mathbf{x}_t) - \epsilon_{\theta'}^{(t)}(\mathbf{x}'_t))\right\| = 0$, the following relationship emerges:

$$\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t) = \epsilon_\theta^{(t)}(\mathbf{x}_t) + \frac{A_t}{B_t}(\mathbf{x}_t - \mathbf{x}'_t). \quad (13)$$

The comprehensive derivation process for Eq. 13 and theoretical analysis are provided in the supplementary material. In this context, the term $\frac{A_t}{B_t}(\mathbf{x}_t - \mathbf{x}'_t)$ is introduced as a correction term.

Given that the error in $\mathbf{x}'_t$ relative to $\mathbf{x}_t$ is corrected during the subsequent generation of x'_{t-1}, we term this process as ***cross-timestep error correction*** (CTEC). Consequently, the corrected target is formulated as:

$$\mathcal{L}_{\text{CTEC}} = \ell\left(\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t), \epsilon_\theta^{(t)}(\mathbf{x}_t) + \gamma \cdot \frac{A_t}{B_t}(\mathbf{x}_t - \mathbf{x}'_t)\right), \quad (14)$$

where γ is a correction weight used to control the strength of correction. By adjusting γ, the balance between calibration and error correction can be modified, ensuring stable optimization. This approach allows for the incorporation of the correction term, thereby refining the target function to better align with the intended calibration objectives, as shown in Fig. 1 (b).

Algorithm 1: Distillation with CTEC

Input: The full-precision parameters θ, randomly initialized LoRA parameters θ_{LoRA}.
Output: The optimized LoRA parameters θ_{LoRA}.

for epoch $n = 1, \ldots, N$ **do**
  Sample a random $\mathbf{x}_T$
  for step $t = T, \ldots, 1$ **do**
    $\theta' \leftarrow \text{Deq}(\text{Quant}(\theta + \theta_{\text{LoRA}}))$
    Calculate $\epsilon_\theta^{(t)}(\mathbf{x}_t)$ and $\epsilon_{\theta'}^{(t)}(\mathbf{x}'_t)$
    Calculate the CTEC term $\frac{A_t}{B_t}(\mathbf{x}_t - \mathbf{x}'_t)$
    Calculate loss $\mathcal{L}_{\text{CTEC}}$ using Eq. 14.
    $\theta_{\text{LoRA}} \leftarrow \theta_{\text{LoRA}} - \eta \cdot \nabla_{\theta_{\text{LoRA}}} \mathcal{L}_{\text{CTEC}}$
    Generate $\mathbf{x}_{t-1}$ and $\mathbf{x}'_{t-1}$ with Eqs. 6 and 9
  end for
end for

Distillation with CTEC

In the traditional calibration method, only the scale factor s and the zero point z are optimized. However, this approach is insufficient for effectively learning the target in Eq. 14. Consequently, a more efficient method is necessary to accurately learn the corrected target. Low-Rank Adaptation (LoRA) (Hu et al. 2023) emerges as a promising technique for fine-tuning models in a cost-effective manner.

LoRA introduces trainable low-rank matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$ into the pre-trained model's weight matrix $W_0 \in \mathbb{R}^{d\times k}$, where the rank $r \ll \min(d, k)$. The essence of low-rank matrices lies in their ability to approximate complex transformations with fewer parameters, thus reducing computational overhead. In this context, the weight matrix W_0 is refined by a low-rank update $\Delta W = BA$, resulting in the updated weight $W_0 + \Delta W$. Consequently, the forward pass is modified as follows:

$$\mathbf{h} = (W_0 + \Delta W)\mathbf{x} = W_0\mathbf{x} + BA\mathbf{x}. \quad (15)$$

This approach allows the model to adapt effectively while keeping the pre-trained weights W_0 frozen. After fine-tuning with LoRA, the learned updates are merged into the original weights, ensuring that inference costs remain unaffected. Given these advantages, we adopt LoRA to optimize our corrected distillation target as specified in Eq. 14. The detailed procedure is presented in Algorithm 1.

Method	Bitwidth (W/A)	LSUN-Churches (LDM-8, 100 timesteps)		LUN-Bedrooms (LDM-4, 100 timesteps)	
		IS ↑	FID ↓	IS ↑	FID ↓
Full Precision	32/32	**2.70**	**4.08**	**2.29**	**3.43**
PTQ4DM (Shang et al. 2023)	8/8	2.52	5.54	2.21	4.75
Q-Diffusion (Li et al. 2023a)	8/8	2.53	4.87	2.19	4.67
LSQ (Esser et al. 2020)	8/8	2.68	4.06	2.18	3.23
ADP-DM (Wang et al. 2023)	8/8	2.69	4.02	**2.35**	3.88
EfficientDM (He et al. 2024)	8/8	**2.71**	4.01	2.27	2.98
CTEC (Ours)	8/8	2.69	**4.00**	2.34	**2.95**
PTQ4DM (Shang et al. 2023)	6/6	2.46	11.05	2.08	11.10
Q-Diffusion (Li et al. 2023a)	6/6	2.47	10.90	2.11	10.10
LSQ (Esser et al. 2020)	6/6	2.66	5.04	2.16	3.55
ADP-DM (Wang et al. 2023)	6/6	2.67	6.90	2.27	9.88
EfficientDM (He et al. 2024)	6/6	2.82	6.29	2.28	3.62
CTEC (Ours)	6/6	**2.86**	**5.02**	**2.31**	**3.52**
LSQ (Esser et al. 2020)	4/4	2.63	**9.08**	2.11	**6.17**
EfficientDM (He et al. 2024)	4/4	2.81	14.34	2.27	10.60
CTEC (Ours)	4/4	**2.89**	12.23	**2.30**	8.52

Table 1: Performance of unconditional image generation on LSUN-Churches and LSUN-Bedrooms at a 256×256 resolution. Bitwidth (W/A) refers to the bitwidth used for weight and activation quantization.

Experiments

We conduct extensive experiments to evaluate the proposed CTEC distillation method, covering both unconditional and conditional image generation. Following this, we analyze the error propagation and accumulation across timesteps within the quantized diffusion process, showing that our correction term effectively mitigates these errors. Additionally, we carry out ablation studies on the key components of the proposed method.

Experimental Settings

In our experiments, we employed DDIM (Song, Meng, and Ermon 2021) and LDM (Rombach et al. 2022) for both unconditional and conditional image generation tasks. The LSUN-Churches and LSUN-Bedrooms datasets (Yu et al. 2015) are utilized for unconditional generation, while the ImageNet datasets (Deng et al. 2009) is used for conditional generation. For the LSUN-Churches, we employed the LDM-8 configuration as outlined in the original paper. Similarly, the LDM-4 configuration was applied for the LSUN-Bedrooms. In both LSUN datasets, the DDIM sampler was configured with 100 timesteps. For the conditional generation task using ImageNet, we used the LDM-4 configuration alongside a DDIM sampler with 20 timesteps. All generated images had a resolution of 256×256 pixels.

For the LSUN dataset, we evaluate the performance of our models by reporting the Inception Score (IS) (Salimans et al. 2016) and the Fréchet Inception Distance (FID) (Heusel et al. 2017). IS measures both the diversity and quality of generated images, while FID evaluates the similarity between the distributions of real and generated images in the Inception network's feature space. For the more complex ImageNet dataset, which contains a diverse set of classes, additional metrics are employed alongside IS and FID. We extend our evaluation by including the spatial Fréchet Inception Distance (sFID) (Nash et al. 2021) and Precision (Kynkäänniemi et al. 2019). sFID incorporates spatial information to capture the quality and realism of generated images, particularly regarding spatial coherence and structure. Precision focuses on the fidelity of generated images, ensuring that they are not only diverse but also indistinguishable from real images.

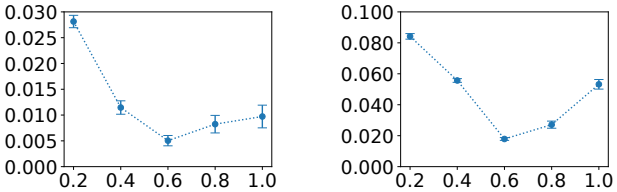
Evaluation of Unconditional Generation

We evaluate the performance of our method for unconditional image generation on the LSUN-Churches and LSUN-Bedrooms datasets. The evaluation is conducted with bitwidth configurations of W8/A8, W6/A6, and W4/A4. The results for IS and FID are presented in Table 1.

With W8/A8 bitwidth, our approach attains the lowest FID among all methods, indicating superior quality in generated images. Although our IS is not the highest, it is nearly equivalent to the best-performing method, with only marginal differences, 0.02 and 0.01, on the LSUN-Churches and LSUN-Bedrooms datasets, respectively. Notably, when the bitwidth is reduced to W6/A6, our method outperforms previous approaches in terms of both IS and FID. This demonstrates the capacity of our method to mitigate quantization errors and maintain high generation quality even under constrained bitwidth conditions. At W4/A4 bitwidth, our method achieves the highest IS. Although our FIDs are higher than those of LSQ, they are lower than those of EfficientDM. The relatively low FID of LSQ can be attributed to its optimization on the original training dataset (He et al.

Method	Bitwidth (W/A)	IS ↑	FID ↓	sFID ↓	Precision ↑ (%)
Full Precision	32/32	**364.73**	**11.28**	**7.70**	**93.66**
Q-Diffusion (Li et al. 2023a)	8/8	350.93	10.60	9.29	92.46
PTQD (He et al. 2023)	8/8	359.78	**10.05**	9.01	93.00
EfficientDM (He et al. 2024)	8/8	362.34	11.38	**8.04**	93.77
CTEC (Ours)	8/8	**362.92**	10.92	8.23	**94.01**
Q-Diffusion (Li et al. 2023a)	4/8	336.80	9.29	9.29	91.06
PTQD (He et al. 2023)	4/8	344.72	8.74	7.98	91.69
EfficientDM (He et al. 2024)	4/8	353.83	9.93	7.34	93.10
CTEC (Ours)	4/8	**355.62**	**8.52**	**7.30**	**93.81**
Q-Diffusion (Li et al. 2023a)	2/8	49.08	43.36	17.15	43.18
PTQD (He et al. 2023)	2/8	53.36	39.37	15.14	45.89
EfficientDM (He et al. 2024)	2/8	175.03	7.60	8.12	78.90
CTEC (Ours)	2/8	**176.37**	**7.43**	**7.98**	**80.20**

Table 2: Performance of conditional image generation on ImageNet at a 256×256 resolution. Bitwidth (W/A) refers to the bitwidth used for weight and activation quantization. LDM-4 and DDIM with 20 timesteps are used.



(a) LSUN-Churches (b) LSUN-Bedrooms

Figure 2: Calibration error (MSE, vertical axes) at the final diffusion timestep with different correction weights γ (horizontal axes).

2024), which incurs higher training costs. In contrast, our method relies solely on the non-quantized model for distillation. These results showcase the efficacy of our approach, particularly in handling low bitwidth quantization.

Evaluation of Conditional Generation

We evaluate the performance of our method for conditional image generation on the ImageNet datasets. The evaluation is conducted with bitwidth configurations of W8/A8, W4/A8, and W2/A8. The results for IS, FID, sFID and Precision are presented in Table 2. Our method consistently outperforms previous approaches under low bitwidth settings, specifically W4/A8 and W2/A8, across all four evaluated metrics. This demonstrates its superior performance in generating high-quality and precise images for conditional image generation. With an increased bitwidth of W8/A8, our method achieves results that are competitive with existing approaches. While the FID and sFID are slightly higher than the top-performing methods under W8/A8, our method attains the highest IS and Precision.

Different Weights for Correction

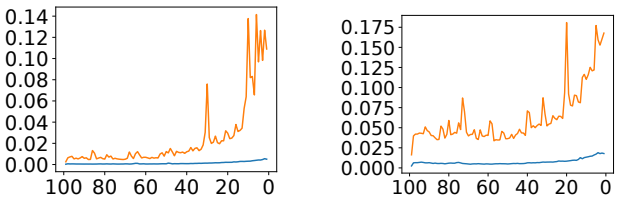
In this ablation study, we evaluate the impact of varying correction strengths by adjusting the correction weight γ. The primary metric used for evaluation is the MSE error at the final timestep. It is important to note that we do not average errors across all timesteps. This decision stems from the fact that errors in earlier timesteps can potentially be mitigated by corrections in subsequent steps. Consequently, only the errors at the final timestep have a direct influence on the quality of the generated image. The outcomes of this experiment are illustrated in Fig. 2. The experiment is conducted under both the LSUN-Churches (Fig. 2a) and LSUN-Bedrooms (Fig. 2b) settings with W8/A8.

The results across both datasets demonstrate a consistent pattern. When the correction strength is set to a low value ($\gamma = 0.2$ or 0.4), the final step error remains relatively high. This outcome is anticipated. When the strength is too low, it lacks the capacity to sufficiently correct the errors. At $\gamma = 0.6$, we observe the lowest final step error. However, it is noteworthy that increasing γ beyond this point, specifically to 0.8 and 1.0, results in a noticeable rise in errors. This trend may be attributable to the interaction between the calibration target $\epsilon_\theta^{(t)}(\mathbf{x}_t)$ and the correction terms $\gamma \cdot \frac{A_t}{B_t}(\mathbf{x}_t - \mathbf{x}'_t)$ in Eq. 14, which are optimized concurrently. An excessive correction strength might have potentially negative effects on the calibration target, which leads to suboptimal results.

A similar observation can be found in the analysis of the error bars, which reveals that when γ becomes too large, the errors exhibit instability, with significant variances. This instability may also result from the aforementioned interaction between the calibration and correction processes. Therefore, it is essential to maintain an appropriate balance between calibration accuracy and error correction. Based on the results of this experiment, we have determined that $\gamma = 0.6$ is a proper value for achieving this balance.

Error Propagation in Quantized DMs

To provide a comprehensive understanding of the proposed error correction process, we evaluate the MSE errors across



(a) LSUN-Churches (b) LSUN-Bedrooms

Figure 3: Calibration errors (MSE, vertical axes) across 100 timesteps (horizontal axes) in DDIM. Errors with our CTEC correction (orange lines) term are compared with errors without this correction (blue lines).

Method	Target	LoRA	IS ↑	FID ↓
Calibration	w/o correction	N/A	**2.57**	**5.15**
	with CTEC	N/A	2.49	8.53
Distillation	w/o correction	✗	2.55	5.04
	with CTEC	✗	**2.58**	**4.62**
Distillation	w/o correction	✓	2.62	4.22
	with CTEC	✓	**2.69**	**4.00**

Table 3: Ablation studies on components of the proposed CTEC distillation method on LSUN-Churches at a 256×256 resolution. The bitwidth is W8/A8. LDM-4 and DDIM with 100 timesteps are used.

100 timesteps in DDIM. Errors after applying our CTEC term are compared with errors without correction. The outcomes of this experiment are illustrated in Fig. 3. The experiment is conducted under both the LSUN-Churches (Fig. 3a) and LSUN-Bedrooms (Fig. 3b) settings with W8/A8.

The comparison underscores the critical impact of the CTEC correction term on error magnitude, illustrating two primary observations that highlight the effectiveness of the correction. First, the application of the correction term significantly reduces the MSE error, making it approximately 10 times smaller under both settings. Secondly, in the absence of the correction term, the error exhibits considerable perturbations. While the error may decrease at certain timesteps, it often experiences dramatic increases in the subsequent steps. In contrast, when our correction term is applied, the error shows only slight, gradual increases as the timesteps progress, without sudden spikes.

By examining the error trajectories with and without the CTEC term, we can clearly see the impact of the correction term in reducing errors throughout the quantized diffusion process. This reduction effectively enhances the stability of the quantized DMs. The observed improvements provide strong evidence that the proposed ***cross-timestep*** method is effective in addressing and correcting the propagation of errors across timesteps.

Ablation Studies

We perform a series of ablation studies to critically evaluate the key components of the proposed CTEC distillation method. Firstly, we evaluate the impact of including the CTEC term in the optimization target, comparing performance with and without this component to determine its contribution to the overall method. Secondly, we evaluate two optimization approaches: the conventional calibration method, which focuses solely on optimizing quantization parameters, and a distillation-based method. Finally, we evaluate the influence of applying distillation with LoRA while keeping the original model weights fixed, against directly optimizing the original weights. The results are reported in Table 3. The ablation studies are conducted using the LSUN-Churches dataset with a bitwidth of W8/A8.

The experimental results demonstrate that the full implementation (the last row of Table 3) of the proposed method outperforms other configurations.It achieves the highest IS and the lowest FID. In contrast, the results also reveal that the use of conventional calibration methods, particularly when combined with error correction, results in degraded performance. This outcome aligns with our expectations. The reason for this decline in performance is that conventional calibration updates only quantization parameters, which are likely too simplistic to handle the complexities introduced by error correction. Therefore, the introduction of a distillation-based method is necessary.

The experiment further shows that, when employing distillation-based methods, the inclusion of a correction term (the fourth and last row of Table 3) consistently improves the results, regardless of whether LoRA is used. This highlights the critical role of the proposed correction term. Correcting accumulated errors across timesteps leads to higher quality of generated images compared to merely matching the outputs of the quantized and full-precision noise prediction models. Additionally, the results suggest that utilizing LoRA for distillation yields superior outcomes compared to directly fine-tuning the original weights.

Conclusion

In this paper, we provide a comprehensive analysis of cross-timestep error propagation in quantized DMs. Our analysis reveals the limitations of previous methods that primarily focus on minimizing the discrepancy between the noise estimations of quantized and full-precision models. Our findings demonstrate that this approach alone is insufficient for reducing the overall discrepancy between the generated samples. We introduce the concept of cross-timestep error correction (CTEC), where the quantized model not only approximates the full-precision model's noise estimation but also corrects the quantization errors propagated from previous timesteps. This correction is learned through a distillation method, which effectively enhances the model's performance in generating samples with reduced error. Our experiments on unconditional image generation with two LSUN datasets and conditional image generation with ImageNet demonstrate the efficacy of our method in significantly reducing the accumulated quantization error across timesteps within the quantized diffusion process. This improvement facilitates the generation of high-quality images, even under the limitations of reduced bitwidths.

References

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

Esser, S. K.; McKinstry, J. L.; Bablani, D.; Appuswamy, R.; and Modha, D. S. 2020. Learned step size quantization. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gupta, S.; Agrawal, A.; Gopalakrishnan, K.; and Narayanan, P. 2015. Deep learning with limited numerical precision. In *International conference on machine learning*, 1737–1746. PMLR.

He, Y.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2024. EfficientDM: Efficient Quantization-Aware Fine-Tuning of Low-Bit Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

He, Y.; Liu, L.; Liu, J.; Wu, W.; Zhou, H.; and Zhuang, B. 2023. Ptqd: Accurate post-training quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2023. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

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

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

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32.

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

Li, Y.; Gong, R.; Tan, X.; Yang, Y.; Hu, P.; Zhang, Q.; Yu, F.; Wang, W.; and Gu, S. 2021. BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction. In *International Conference on Learning Representations*.

Li, Y.; Xu, S.; Cao, X.; Sun, X.; and Zhang, B. 2023b. Q-dm: An efficient low-bit quantized diffusion model. *Advances in Neural Information Processing Systems*, 36.

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

Nagel, M.; Baalen, M. v.; Blankevoort, T.; and Welling, M. 2019. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1325–1334.

Nash, C.; Menick, J.; Dieleman, S.; and Battaglia, P. 2021. Generating images with sparse representations. In *International Conference on Machine Learning*, 7958–7968. PMLR.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

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

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

Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; and Yan, Y. 2023. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1972–1981.

So, J.; Lee, J.; Ahn, D.; Kim, H.; and Park, E. 2023. Temporal dynamic quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Tang, S.; Wang, X.; Chen, H.; Guan, C.; Wu, Z.; Tang, Y.; and Zhu, W. 2023. Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models. *arXiv preprint arXiv:2311.06322*.

Wang, C.; Wang, Z.; Xu, X.; Tang, Y.; Zhou, J.; and Lu, J. 2023. Towards accurate data-free quantization for diffusion models. *arXiv preprint arXiv:2305.18723*, 2(5).

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.