Accurate Binarized Diffusion Model

Haotong Qin, Xianglong Liu, *Senior Member, IEEE*, Xingyu Zheng, Haoran Chu, Jinyang Guo, Yulun Zhang, Yawei Li, Bo Li, Luca Benini, *Fellow, IEEE*, Michele Magno, *Senior Member, IEEE*

Abstract—Diffusion models have demonstrated remarkable success in generating high-quality images and videos, but entail heavy computational demands. Binarization offers a promising route to compact and fast models by reducing parameters to a 1-bit representation and enabling efficient bitwise operations. However, the binarized diffusion models traditionally result in significant performance loss due to limited representation capacity and optimization difficulties. To overcome these challenges, we propose **BiDM+**, a highly accurate binarized diffusion model that binarizes both weights and activations to enhance computational efficiency while maintaining high generative quality. Informed by empirical analysis for representations from statistical, temporal, and spatial perspectives, BiDM+ introduces innovations at three distinct levels: (1) Operator Level: We design *Flexible Projection Binarization* (FPB), which learns independent quantization intervals and scales, thereby mapping representations into a versatile 1-bit space and ensuring accurate binarization of both weights and activations. (2) Structural Level: We propose the *Timestep-aware Evoluted Structure* (TES) that promotes feature reuse across timesteps and leverages temporal evolution to boost the quality of activation features. TES dynamically balances stability during early diffusion steps with precision during later stages. (3) Optimization Level: We introduce *Salient-guided Binarization-aware Mimicking* (SBM), a distillation approach that exploits full-precision models and intrinsic quantization properties to accelerate convergence while mitigating minor quantization errors. Comprehensive experiments validate that our BiDM+ achieves state-of-the-art performance across multiple diffusion models and benchmarks. For instance, with only 3.6% storage and 1.9% FLOPs compared to 32-bit diffusion models, BiDM+ first records a FID low to 25.69 for LDM-4 on FFHQ 256×256, outperforming the previous best method by a significant margin of 17.73. The results underscore the practical potential of BiDM+ in resource-constrained scenarios.

1 Introduction

Diffusion models [6], [15], [20], [29], [46], [52], [68] have emerged as a leading paradigm of generation [62], [73], [74], demonstrating exceptional performance across diverse applications such as image synthesis [60], video generation [19], [42], neural network generation [63], image restoration [34]. Their ability to produce high-quality, diverse outputs has garnered significant attention, leading to widespread adoption. However, the iterative denoising process, often spanning up to one thousand steps [20], incurs substantial computational and memory overhead during inference. While various acceleration techniques have been proposed to reduce the number of required steps [35], [58], each step remains heavily reliant on expensive floating-point operations. Consequently, reducing the computational burden per step has become a crucial research focus, driving the development of diffusion model compression techniques.

Among these compression methods, quantization [31], [49], [56], distillation [11], [40], [43], [55], [77], and pruning [8] have been explored to mitigate inference costs while preserving generative quality. Quantization, in particular, stands out as a promising and versatile approach [9], [10], [22], [70], [75], mapping weights and activations to lower bit-width representations (*e.g.*, 2∼8-bit) to reduce storage and computation overhead. Recent advances in quantized diffusion models demonstrate strong performance retention at integer bit-widths such as 8-bit [16], [30], [31], [56], making them broadly applicable across different architectures. As the most extreme form of quantization, binarization compresses model parameters to 1-bit, leveraging efficient bitwise operations such as XNOR and POPCNT to significantly enhance computational efficiency [78]. Binarization has been successfully applied to convolutional neural networks and various vision tasks [37], [71], [72], showcasing its potential to reduce inference costs while maintaining practical performance. However, its application to diffusion models remains largely unexplored due to the severe representational limitations and optimization challenges that often lead to performance degradation or model collapse.

Binarizing diffusion models with existing methods remains highly challenging. In fully binarized diffusion models, where both weights and activations are constrained to 1-bit, generative performance often deteriorates significantly or even collapses. Several key challenges hinder effective diffusion model binarization: (1) Statistical distribution of representations: diffusion models exhibit broad, imbalanced activation distributions that also vary significantly across time steps. Such highly dynamic, uneven distributions make preserving essential information difficult for fixed binarization strategies. (2) Temporal variation of representations: Although adjacent timesteps in full-precision diffusion models share high similarity and stabilize generation, binarization accentuates the transition from continuous to discrete representations. This disrupts inter-step correlation and leads to instability in updates. (3) Spatial salience of representations: The magnitude of representational elements is spatially dispersed across different locations, with high-salient elements playing a dominant role in preserving the performance of binarization. The direct optimization struggles to be aware of and retain this spatially distributed salience for binarized diffusion models. Collectively, the limitations in representational capacity, the instability introduced into temporal dynamics, and the disruption of spatial alignment impede the accurate training and inference

- *H. Qin, M. Magno, Y. Li, and L. Benini are from ETH Zurich, Switzerland. X. Liu (corresponding author, email: xlliu@buaa.edu.cn), X. Zheng, H. Chu, J. Guo, and B. Li are from Beihang University, China. Y. Zhang is from Shanghai Jiaotong University, China.*

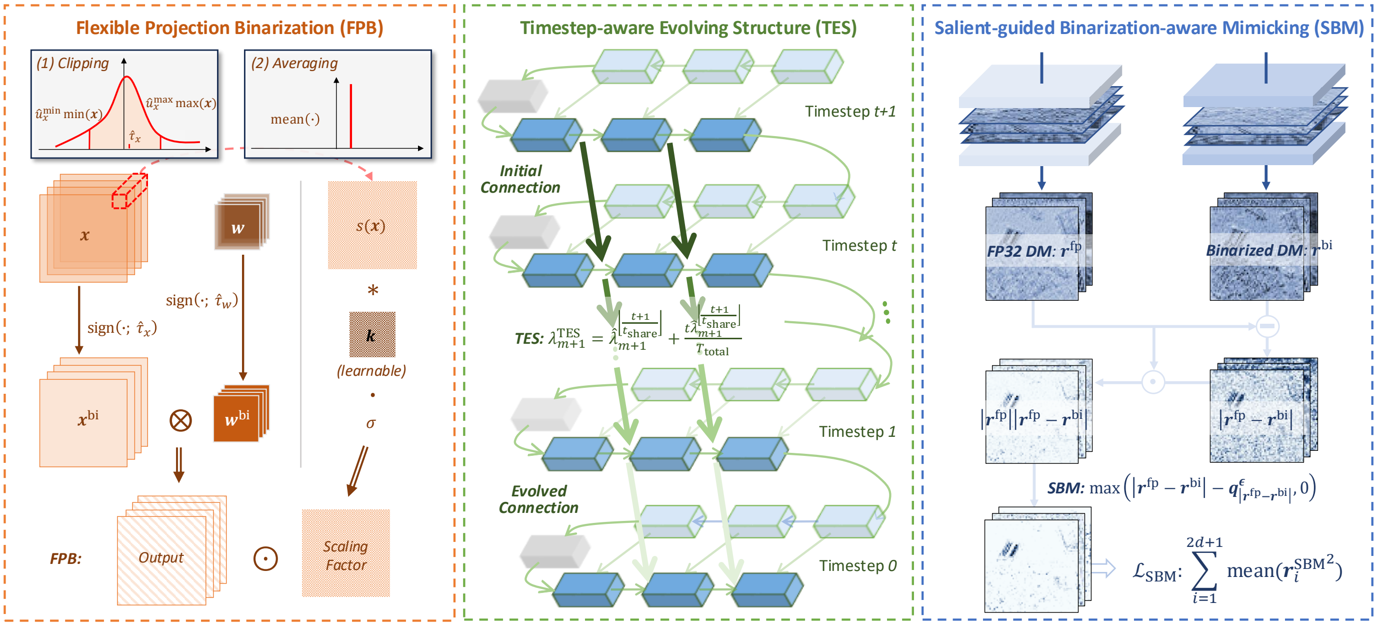

Flexible Projection Binarization (FPB)
(1) Clipping
$\hat{\alpha}_x^{\min}$ min(x)
$\hat{\alpha}_x^{\max}$ max(x)
$\hat{\tau}_x$
(2) Averaging
mean(·)
x
w
$s(x)$
sign(·; $\hat{\tau}_w$)
sign(·; $\hat{\tau}_x$)
*
k
(learnable)
·
σ
x^{bi}
⊗
w^{bi}
FPB:
Output
⊙
Scaling Factor
Timestep-aware Evolving Structure (TES)
Timestep t+1
Initial Connection
Timestep t
TES: $\lambda_{m+1}^{\text{TES}} = \hat{\lambda}_{m+1}^{\left\lfloor \frac{t+1}{t_{\text{share}}} \right\rfloor} + \frac{t\hat{\lambda}_{m+1}^{\left\lfloor \frac{t+1}{t_{\text{share}}} \right\rfloor}}{T_{\text{total}}}$
Timestep 1
Evolved Connection
Timestep 0
Salient-guided Binarization-aware Mimicking (SBM)
FP32 DM: r^{fp}
Binarized DM: r^{bi}
$|r^{\text{fp}}||r^{\text{fp}} - r^{\text{bi}}|$
$|r^{\text{fp}} - r^{\text{bi}}|$
SBM: $\max\left(|r^{\text{fp}} - r^{\text{bi}}| - q^{\epsilon}_{|r^{\text{fp}} - r^{\text{bi}}|}, 0\right)$
$\mathcal{L}_{\text{SBM}}$: $\sum_{i=1}^{2d+1} \text{mean}(r_i^{\text{SBM}^2})$

Fig. 1: Overview of BiDM+, applying Flexible Projection Binarization (FPB), Timestep-aware Evolving Structure (TES), and Salient-guided Binarization-aware Mimicking (SBM) as the operator, structure, and optimization, respectively.

of binarized diffusion models.

To address these challenges, this paper presents **BiDM+**, a highly accurate binarization framework for efficient diffusion models. BiDM+ significantly improves computational efficiency by binarizing both weights and activations without sacrificing model performance. Its success is underpinned by several key innovations in operator design, network architecture, and optimization strategy. First, we introduce the *Flexible Projection Binarization* (FPB) operator, which customizes both the quantization range and scaling factor to map diffusion model representations into a 1-bit numerical space. This tailored projection minimizes information loss during the compression of weights and activations to 1-bit, thereby ensuring a precise binary representation. Next, we propose the *Timestep-aware Evoluted Structure* (TES). TES enhances the performance of binarized diffusion models by reusing features across time steps and incorporating temporal evolution to refine activation features. This structure not only stabilizes the early diffusion steps but also progressively boosts accuracy in later stages by continuously adapting and reusing features. Finally, we introduce the *Salient-guided Binarization-aware Mimicking* (SBM) optimization strategy. SBM improves the training of binarized diffusion models by prioritizing the most salient components during quantization while allowing for minor quantization errors. By leveraging both the guidance from a high-precision reference model and the unique quantization properties of diffusion models, SBM facilitates faster convergence and more robust performance.

Comprehensive experiments demonstrate that **BiDM+** consistently achieves state-of-the-art (SOTA) performance when compared with existing binarized diffusion models utilizing advanced methodologies. In terms of accuracy, BiDM+ significantly outperforms all current binarized diffusion models across multiple evaluation metrics without imposing an additional inference burden. Within pixel-space diffusion models (DDIM), BiDM+ is uniquely capable of enhancing the Inception Score (IS) to 5.53 on the CIFAR-10 dataset, outperforming the previous best method by 0.35 and approaching the performance of full-precision models. In the context of Latent Diffusion Models (LDM), BiDM+ reduces the Fréchet Inception Distance (FID) on FFHQ from the previous SOTA value of 43.42 to an impressive 25.69. In terms of efficiency, BiDM+ achieves up to a 28.0× reduction in storage requirements and a 52.7× decrease in computational FLOPs. As a fully binarized model (both weights and activations are 1-bit), extensive visualizations of generated samples confirm that BiDM+ is currently the most advanced approach available for generating acceptable images with fully binarized diffusion models, paving the way for practical deployment on resource-constrained edge hardware.

Please note that this manuscript extends our original conference paper [81]. In what follows, we detail the contributions of this work compared to the conference version:

- **Motivation:** We provide a deeper analysis and explanation of the empirical behavior of representations in binarized diffusion models. In particular, we analyze these representations with respect to their statistical distributions, temporal update dynamics, and spatial characteristics. Such an in-depth investigation not only clarifies the challenges associated with diffusion model binarization but also motivates the design of more accurate and efficient binarized diffusion models.
- **Methodologies:** We introduce a highly accurate binarization framework for diffusion models, termed *BiDM+*. This framework incorporates innovations at the operator, structure, and optimization levels. The techniques presented in the conference version serve as a specific baseline, while our contributions are as follows:

1) **Operator Level:** We propose the *Flexible Projection Binarization (FPB)* operator. Unlike the operator from

the conference version, FPB learns the quantization interval and scale independently to project representations into a 1-bit numerical space with enhanced flexibility. The FPB operator is designed to accurately compress both weights and activations.

2) **Structural Level:** We design the *Timestep-aware Evoluted Structure (TES)*. TES employs feature reuse across time steps, complemented by temporal evolution, to refine activation features in binarized diffusion models. Its key innovation lies in achieving a balance between the stability required during early time steps and the accuracy needed in later steps.
3) **Optimization Level:** We develop the *Salient-guided Binarization-aware Mimicking (SBM)* method. SBM enhances the distillation process for binarized diffusion models by taking into account quantization salience and tolerating minor inevitable errors. It leverages both the external full-precision counterparts and the intrinsic quantization properties of diffusion models to accelerate optimization convergence.

- **Experiments:** We conduct extensive evaluations of BiDM+ across a wider range of practical scenarios, assessing both accuracy and efficiency. Our experiments include comprehensive quantitative and visual analyses on the image generation benchmark, as well as a detailed visual examination of generation. The results demonstrate substantial improvements in both quantitative metrics and intuitive visualization compared with the conference version.

2 Related Work

Diffusion Models have demonstrated remarkable prowess in generating high-quality samples in diverse domains, including images, audio, and text [20], [24], [44], [45], [47], [60], [61]. However, these models typically feature complex architectures and require substantial computational resources due to the iterative denoising process, which may hinder their deployment in resource-constrained environments. To address these challenges, recent research has focused on improving the efficiency of diffusion models. Several techniques, such as non-retraining sampling acceleration methods [35], [39], [58], have been proposed to reduce the number of timesteps required during inference. In particular, DPM-OT reformulates the denoising trajectory as an optimal transport problem, effectively mitigating mode collapse in few-step sampling while balancing efficiency and quality [32]. Moreover, architectural innovations have been introduced to optimize performance further. For example, SVDiff achieves parameter-efficient fine-tuning by applying singular value decomposition to weight matrices, resulting in a model size reduction by a factor of 2200 [13]. DeepCache offers memory optimization through feature caching mechanisms [41]. In addition, model distillation approaches [40], [43] have been extensively studied to decrease the timestep requirements, and consistency models enable single-step generation by leveraging direct latent-space mapping [59]. Despite these advancements, these efficient diffusion model architectures still rely on computationally expensive floating-point operations. This reliance indicates potential for further optimization through low-bit quantization techniques.

Quantization reduces the precision of weights and activations from standard 32-bit floating point representations to low-bit widths, typically ranging from 1 to 8 bits, to compress and accelerate neural networks [7], [79], [83], [84]. Earlier pioneering efforts [12] demonstrated that employing limited numerical precision can yield efficient deep networks without severely compromising accuracy. Studies such as [14] introduced quantization strategies integrating other compression techniques, such as pruning and Huffman coding. Recent studies have extended these quantization methods to generative diffusion models, driven by the need to achieve substantial computational savings while maintaining acceptable quality [4], [16]. This line of work carefully adapts the quantization strategies to account for the unique timestep dependencies and spatial architectures that these models exhibit. However, the complex nature of generative tasks imposes significant challenges; for example, post-training quantization methods struggle to reduce precision below 4 bits without inducing a notable degradation in fidelity [30], [56]. Quantization-aware training has also been pursued as an alternative, yet it faces performance bottlenecks when pushing bit-widths under 3 bits, often leading to unacceptable losses in generative quality [31], [57].

Binarization represents the most extreme form of quantization, in which both weights and activations are constrained to take on binary values (typically +1 or -1), thereby maximizing potential gains in compression and computational speed [33], [64], [66]. In the domain of computer vision, binarization has predominantly been investigated within discriminative models such as convolutional neural networks and Vision Transformers [17], [27], [37], [48], [50], [51]. Although promising results have been obtained for tasks like image classification, the application of binarization to generative models remains relatively underexplored. For instance, approaches implemented in architectures such as ResNet-based variational autoencoders and Flow++ have attempted full binarization with limited success, falling short of the performance levels offered by more refined quantization schemes [1]. Furthermore, the recent development of Binary Latent Diffusion has aimed at binarizing the latent representations in latent diffusion models; however, this method has not yet achieved significant enhancements in spatial or computational efficiency [65]. The latest advancement, BinaryDM [82], approaches near-binary quantization at the W1A4 level, but full extension to activation binarization without performance collapse is still an open challenge.

3 Method

3.1 Preliminaries

Diffusion Models: In the setting of generative modeling, consider a data distribution $\boldsymbol{x}_0$ sampled from $q(\boldsymbol{x}_0)$. A forward diffusion process is then defined that transforms this initial data distribution through a sequence of random variables $\boldsymbol{x}_t$ for $t \in \{1, \ldots, T\}$. The transformation at each step employs a transition kernel $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, typically a Gaussian perturbation, which progressively adds noise to the data. This process can be formally described as:

$$q\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \boldsymbol{x}_0\right) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}), \tag{1}$$

where each transition is defined by:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right)=\mathcal{N}\left(\boldsymbol{x}_t ; \sqrt{1-\beta_t} \boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right), \tag{2}$$

and $\beta_t \in (0,1)$ specifies a noise schedule that controls the variance of the added noise at each step. The use of a Gaussian transition kernel simplifies the marginalization of the joint distribution. By defining $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^{t} \alpha_i$, the state $\boldsymbol{x}_t$ can be obtained directly without sequentially processing all previous states. Specifically, a sample from $\boldsymbol{x}_t$ can be generated using:

$$\boldsymbol{x}_t=\sqrt{\bar{\alpha}_t} \boldsymbol{x}_{\mathbf{0}}+\sqrt{1-\bar{\alpha}_t} \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

The reverse process is crucial as it aims to reconstruct the original data distribution by systematically removing noise added during the forward process. This is achieved by approximating the conditional distribution $q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$, which is typically not analytically tractable, with a parameterized transition kernel $p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)$. This approximation is modeled as:

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)=\mathcal{N}\left(\boldsymbol{x}_{t-1} ; \tilde{\boldsymbol{\mu}}_\theta\left(\boldsymbol{x}_t, t\right), \tilde{\beta}_t \boldsymbol{I}\right), \tag{4}$$

where the mean $\tilde{\boldsymbol{\mu}}_\theta\left(\boldsymbol{x}_t, t\right)$ and the variance $\tilde{\beta}_t$ are derived using reparameterization techniques. These parameters are calculated as follows:

$$\tilde{\boldsymbol{\mu}}_\theta\left(\boldsymbol{x}_t, t\right)=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right), \tag{5}$$

$$\tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t, \tag{6}$$

where $\boldsymbol{\epsilon}_\theta$ represents a neural network function approximation. This network, parameterized by θ, is trained to predict the noise component $\boldsymbol{\epsilon}$ based on the noised data $\boldsymbol{x}_t$ at each step t. By learning to accurately predict and subsequently subtract this noise, the model can effectively reverse the diffusion process, gradually denoising the data to reconstruct the original distribution $\boldsymbol{x}_0$.

In the training of diffusion models, practitioners often employ a variant of the variational lower bound optimized to enhance the quality of generated samples. This modified loss function, focusing on the minimization of noise discrepancies at various diffusion steps, is described by:

$$\mathcal{L}_{\mathrm{DM}}=\mathbb{E}_{t \sim[1, T], \boldsymbol{x}_0 \sim q\left(\boldsymbol{x}_0\right), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_t-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t, t\right)\right\|^2\right]. \tag{7}$$

The U-Net architecture, first introduced by [53] for medical image segmentation, has become pivotal in the field of diffusion models due to its robust feature integration capabilities across different levels of representation. U-Net's design features a series of encoding (D_m) and decoding (U_m) blocks, indexed from 1 through d, which process features at progressively higher semantic levels as m increases. The architecture ensures that every decoding step is enhanced by direct feature transfers from corresponding encoding layers, fostering rich detail preservation and integration. This connectivity is mathematically represented as:

$$\operatorname{concat}\left(D_m(\cdot), U_{m+1}(\cdot)\right). \tag{8}$$

The synergy between detailed low-level features and high-dimensional semantic information, facilitated by skip connections in U-Net, significantly contributes to the effectiveness of diffusion models in generating high-fidelity outputs.

Binarization Framework: Quantization methods like binarization are not only instrumental in reducing model size but also in speeding up inference by leveraging hardware-efficient operations. For weights and activations converted to 1-bit representations, the calculations become more about bitwise operations, which are inherently faster on hardware architectures.

A common approach in binarization involves converting the weights $\boldsymbol{w}$ into 1-bit values, a process highlighted in seminal works such as those by [5], [51] and further refined by [21] in the context of diffusion models. The binarization function for weights can be described mathematically as

$$\sigma \boldsymbol{w}^{\mathrm{bi}}= \begin{cases}\sigma, & \text { if } \boldsymbol{w} \geq 0, \\ -\sigma, & \text { otherwise, }\end{cases} \tag{9}$$

where the sign function maps the weight $\boldsymbol{w}$ to either $+1$ or -1 with a zero threshold. The scaling factor σ, a floating-point scalar, is initially set as $\frac{\sum\left|\boldsymbol{w}_i\right|}{w \times h}$ and is adaptable during the training phase. Similarly, activations are typically quantized using basic binarizers, which is supported by research from [23], [36]. The quantization function for activations is

$$\boldsymbol{x}^{\mathrm{bi}}= \begin{cases}1, & \text { if } \boldsymbol{x} \geq 0, \\ -1, & \text { otherwise. }\end{cases} \tag{10}$$

By quantizing both weights and activations to 1-bit, the computational procedures within the denoising model can be streamlined using bitwise XNOR and POPCNT operations, resulting in significant reductions in memory and latency.

Moreover, when combining binarized weights with binarized activations, the model employs an efficient matrix multiplication substitute using bitwise operations:

$$\sigma\left(\boldsymbol{w}^{\mathrm{bi}} \otimes \boldsymbol{x}^{\mathrm{bi}}\right), \tag{11}$$

where $\otimes$ denotes the MatMul operation performed by element-wise XNOR and POPCNT instructions. This approach is especially effective in layers such as convolutions and fully connected layers where high-dimensional dot products are common.

By employing these techniques, the binarized neural network models can achieve considerable improvements in speed and efficiency, making them particularly well-suited for deployment on edge devices with limited computational resources and power constraints.

3.2 Observations for Binarization of Diffusion Models

Intuitively, the fundamental impact of binarization on diffusion models manifests in the representations generated by quantized layers. To systematically investigate this effect, we conduct a series of empirical studies on the representations of diffusion models, analyzing them from three key perspectives: statistical distribution, temporal updates, and internal spatial. These observations provide critical insights into the underlying reasons for the degradation of diffusion models caused by binarization.

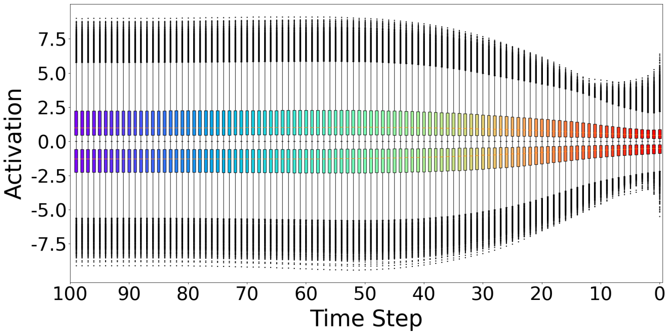


Fig. 2: The activation range of one certain convolution of the 32-bit DDIM on CIFAR-10 varies with denoising time steps.

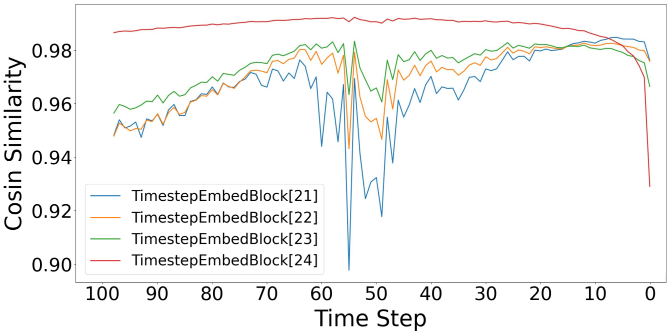


Fig. 3: The similarity of features at each step of full-precision LDM-4 on LSUN-Bedrooms compared to the previous step.

Observation 1: *The representations in diffusion models exhibit a highly dynamic and imbalanced statistical distribution.*

As shown in Fig. 2, the statistical properties of activation representations in full-precision diffusion models exhibit significant diversity. On the one hand, activations within a single step have a broad and uneven distribution range, often accompanied by outliers, making it challenging for binarizers to accurately retain the original information after quantization. On the other hand, even within the same layer, the activation range of diffusion models varies substantially across different time steps (even differing by nearly 70%), posing a challenge for existing binarization with fixed mapping strategies to handle such highly dynamic variations.

Observation 2: *The representations in diffusion models exhibit high similarity between adjacent time steps, while the similarity gradually decreases as the time steps progress.*

As shown in Fig. 3, the activation representations in full-precision diffusion models exhibit high similarity between adjacent time steps, which ensures the stability of the generated outputs. The similarity diminishes in the final time steps and the deeper blocks, promoting diversity in the generated results. Since binarization is applied at each computational unit and at every time step of diffusion models, the resulting quantization errors and high discretization significantly reduce the correlation between adjacent time-step representations, leading to unstable updates.

Observation 3: *The magnitude of representation elements in diffusion models exhibits spatial dispersion, which exists consistently among different time steps and locations.*

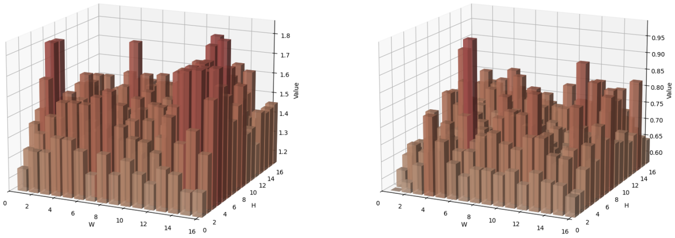


Fig. 4: The salience (magnitude) of quantization for different layers and time steps in LDM-4 on LSUN-Bedrooms.

Magnitude serves as a crucial metric for evaluating the salience of elements within model representations during the quantization process [69], [80]. We observe that the highly salient elements in the diffusion model are dispersed across different spatial locations within the representation, which plays a dominant role in determining the final quantization performance. This spatial dispersion of salience remains consistently present across different layers and time steps of diffusion models, but the specific distributions are significantly variant. These facts imply that naive error-driven binarization methods struggle to accommodate the spatially distributed salience within the representation space, thereby hindering the optimization of binarized diffusion models.

The above observations provide an analysis of the characteristics of features in diffusion models, including the dynamism of distributions, temporal similarity of representations, and the dispersion of salience. These insights help in understanding the performance degradation of diffusion models in the context of binarization. Based on these observations, we propose a series of techniques at the operator, structural, and optimization levels to achieve accurate diffusion model binarization systematically.

3.3 Flexible Projection Binarization for Accurate Feature Numeric

We first investigate the fundamental computational units in binarized diffusion models to mitigate the degradation of numerical diversity in their representations.

Existing binarizers [51], [84] usually compress weights and activations to {+1, -1} with statistics-based scaling factor:

$$\left(\boldsymbol{x}^{\text{bi}} \otimes \boldsymbol{w}^{\text{bi}}\right) \odot \left(\sigma \cdot \boldsymbol{\mathcal{A}} * \boldsymbol{k}\right), \tag{12}$$

where $\boldsymbol{x} \in \mathbb{R}^{c \times w_{\text{in}} \times h_{\text{in}}}$, $\boldsymbol{w} \in \mathbb{R}^{c \times w \times h}$, $\boldsymbol{\mathcal{A}} = \frac{\sum |\boldsymbol{x}_{i,:,:}|}{c}$, and $\sigma = \frac{1}{w \times h} \|\boldsymbol{w}\|_{\ell 1}$. Here, $\boldsymbol{k} \in \mathbb{R}^{1 \times 1 \times w \times h}$ represents a two-dimensional filter, where $\forall i, j$, $k_{i,j} = \frac{1}{w \times h}$. The operators $*$ and $\otimes$ denote convolution composed of floating-point multiply-add and bitwise XNOR and POPCNT operations, respectively. However, the scaling factors $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{k}$, which determine the numerical scalings, are derived from direct and static sampling of original weights and activations (such as average absolute values), making it difficult to adapt to the variable expressions of diffusion models in inference. Some improved binarizers, such as XNOR++ [2], use trainable scaling factors. However, these binarizers remain constant during inference, making it hard to produce representations with varying suitable numerical scalings for dynamic feature inputs of diffusion models.

We propose a Flexible Projection Binarization (FPB) operator to enhance the numerical diversity of computational units in binarized diffusion models. FPB includes learnable thresholds to learn appropriate discretized mappings for weights and activations specific to a particular binarizer. It also introduces a learnable method for scaling factors that is capable of handling dynamic inputs to address the impact of outliers while considering dynamic information during inference. Specifically, for weights $\boldsymbol{w}$, we first introduce a learnable mean offset $\hat{\tau}_w$ in the binarized weights:

$$\boldsymbol{w}^{\text{bi}} = \operatorname{sign}(\boldsymbol{w} - \hat{\tau}_w). \tag{13}$$

During inference, $\hat{\tau}_w$ can be merged into the binarized weights without any additional overhead. For the scaling factor of weights, we allow $\boldsymbol{k}$ to be learned during training to adaptively match changes in the range of activations before and after, and set the elements during initialization to $\frac{\alpha}{w \times h}$ as defined in Eq. (12) to approximate the numerical value of the original weights.

For activations $\boldsymbol{x}$, we also apply the channel-wise learnable mean offset $\hat{\tau}_x$ in the binarized activations. Considering the numerical representation capability of binarized activation is significantly determined by scaling factors, we further design a scaling factor projection for the binarized activations that can control the impact of numerical outliers. Revisiting the design of scaling factors in binarized units, we find that in the original binarizers, the scaling factor for activations is usually sampled directly from features at a specified granularity, such as the average of absolute values [51]. However, this method leads to a high consistency in the activation scaling factor mapping across the entire network, making it difficult to dynamically handle the unconventional effects of outliers during inference, which has been proven crucial for the performance of generative models [28], [69]. Therefore, we construct learnable numerical sampling thresholds for each channel, aimed at learning the channel-wise numerical boundaries for the dynamic sampling process of the activation scaling factor. The computation of the activation scaling factor in FPB is as follows:

$$s\left(\boldsymbol{x}\right) = \operatorname{mean}\left(\operatorname{clip}\left(\boldsymbol{x}, \hat{\mu}_x^{\text{min}} \min\left(\boldsymbol{x}\right), \hat{\mu}_x^{\text{max}} \max\left(\boldsymbol{x}\right)\right)\right), \tag{14}$$

where $\hat{\mu}_x^{\text{min}}$ and $\hat{\mu}_x^{\text{max}}$ represent learnable scaling parameters of boundaries initialized to 1 for activation sampling, and the $\operatorname{clip}\left(\cdot, \hat{\mu}_x^{\text{min}} \min\left(\boldsymbol{x}\right), \hat{\mu}_x^{\text{max}} \max\left(\boldsymbol{x}\right)\right)$ denotes truncating the input to the range. By adjusting the numerical range through learnable scaling parameters, FPB controls the impact of outliers outside the range while obtaining scaling factors from activation sampling.

The forward propagation of the proposed FPB unit during inference can be represented as:

$$\text{FPB:}\ \left(\operatorname{sign}\left(\boldsymbol{x} - \hat{\tau}_x\right) \otimes \boldsymbol{w}^{\text{bi}}\right) \odot \left(\sigma \cdot s\left(\boldsymbol{x}; \hat{\mu}_x^{\text{min}}, \hat{\mu}_x^{\text{max}}\right) * \boldsymbol{k}\right), \tag{15}$$

where β_x denotes a learnable mean offset for activation in the binarizer. Note that compared to existing binarizers, FPB is the first to separately treat the learnable parameters in the binarizer for discretization and scaling projections ($\hat{\tau}_x / \hat{\mu}_x$ *vs.* $\hat{\tau}_x$ for activations), allowing for high flexibility in learning to improve the binarized representation effect as much as possible under stringent extreme 1-bit conditions (Fig. 5). As the fundamental operator, FPB flexibly projects weights and activations into 1-bit, providing a basis for accurate representation and optimization of binarized diffusion models.

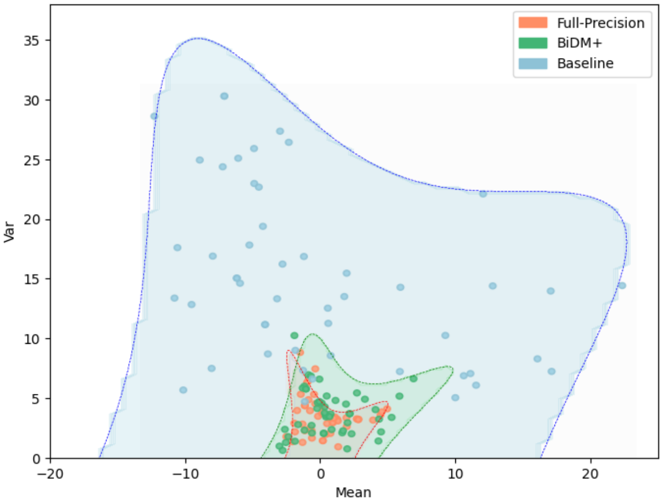


Fig. 5: Compared with the baseline binarizer, the proposed FPB can greatly restore the feature statistics in diffusion models to be close to their original full-precision counterparts.

3.4 Timestep-aware Evolving Structure for Stable Temporal Representation

With the proposed binarizer, we further design a binarization-friendly structure to alleviate the temporal instability of feature updating in binarized diffusion models.

During the denoising process in diffusion models, the output from the previous time step is used as the input for the current time step, and noise is progressively reduced at each time step to achieve well-generated synthetic images. However, the analysis in Sec. 3.2 indicates that binarization leads to more severe fluctuations in the intermediate features of the diffusion model across different time steps, making the entire propagation process difficult to stabilize and accurately execute. We find that this negative impact caused by binarization stems from the inherent nature of the diffusion model. If we unfold the propagation of the diffusion model along the time steps, we can see that features are continuously computed in a model with T times shared weights across time steps. This means that the effects of binarization, such as the inaccurate forward caused by binarization error [51] and backward caused by gradient mismatch [38], accumulate along the temporal dimension throughout the entire propagation process over T time steps, resulting in catastrophically unstable features at the same location over time in the binarized diffusion model.

We construct a Timestep-aware Evoluted Structure (TES) that enhances the stability of binarized models by establishing cross-time-step feature connections. Empirical analysis suggests that features at a specific location in diffusion models exhibit high similarity across adjacent time steps. Therefore, we introduce cross-time-step connections at the same location in binarized diffusion models to mitigate the feature distortion caused by binarization. The basic form of TES can be expressed as:

$$\operatorname{concat}(D_m^t(\cdot), (1 - \lambda_{m+1}) \cdot U_{m+1}^t(\cdot) + \lambda_{m+1} \cdot U_{m+1}^{t+1}(\cdot)), \tag{16}$$

where $\lambda \in [0, 1)$ is a coefficient used to balance the influences from the current and previous time step features. This form is designed to alleviate the severe fluctuations of features between adjacent time steps in binarized diffusion models, mitigating the attenuation of information in forward propagation and the mismatch of gradients in backward propagation due to binarization.

However, the introduction of the feature from the last time step means reducing the influence of the current time step t, thus tending to revert to the features at time step $t+1$. This hinders accurate inference during the later stages of denoising. Moreover, as shown in Fig. 3, the inherent similarity of the diffusion model varies across different blocks and time steps. Therefore, we design a timestep-aware evolutionary strategy to adjust the coefficients of feature connections. We initially set up $\frac{T_{\text{total}}}{t_{\text{share}}}$ independent learnable coefficients, with t_{share} defaulting to 10. Each coefficient balances accurate learning and compact storage use over t_{share} adjacent time-steps. Subsequently, when applying these coefficients, we consider the accompanying temporal decay:

$$\text{TES: } \text{concat}(D_m^t(\cdot), (1-\lambda_{m+1}^{\text{TES}}) \cdot U_{m+1}^t(\cdot) + \lambda_{m+1}^{\text{TES}} \cdot U_{m+1}^{t+1}(\cdot)),$$
$$\lambda_{m+1}^{\text{TES}} = \hat{\lambda}_{m+1}^{\lfloor \frac{t}{t_{\text{share}}} \rfloor} + \frac{t\hat{\lambda}_{m+1}^{\lfloor \frac{t}{t_{\text{share}}} \rfloor}}{T_{\text{total}}}, \tag{17}$$

where the learnable $\hat{\lambda}_{m+1}$ is initialized to 0.25, which will cause $\lambda_{m+1}^{\text{TES}}$ to be initialized to 0.5 at the time step when denoising starts. This means that at the initial time, $U_{m+1}^t(\cdot)$ and $U_{m+1}^{t+1}(\cdot)$ have the same impact and can be learned and adjusted during the following training process. Since t is introduced in the coefficients, the influence of the feature $U_{m+1}^{t+1}(\cdot)$ decreases with the time step increasing in TES. The structure with the established temporal connections enables stable updates in the early stages of denoising and accurate generation in the later stages for the binarized model. It also possesses the capability to adaptively adjust the coefficients through learning to suit different time-steps.

In summary, TES uses a time-step-aware strategy to adaptively average inputs and connections and leverages the high-dimensional feature similarities between adjacent time steps to enhance information representation and, as Fig. 6 shows, achieve stable propagation. The cross-time-step connections of TES enable the binarized diffusion model to use the output information from the previous step for appropriate information compensation, stabilizing the feature updates over time steps. In the later stages, TES actively reduces the influence of information from the previous time step to promote a more accurate generation of the final content by the binarized diffusion model.

3.5 Salient-guided Binarization-aware Mimicking for Spatial-focused Optimization

To improve the convergence of optimization for binarized diffusion models, we designed a strategy called Salient-guided Binarization-aware Mimicking (SBM). SBM allows binarized diffusion models to leverage precise representations from full-precision strategies and prioritizes optimization for elements in the representation that have a greater impact during quantization.

Compared to standard quantization-aware training that solely supervises output-related loss, feature-mimicking optimizes the binarized diffusion model by comparing its intermediate outputs to those of the full-precision counterpart. However, direct alignment between the representations of full-precision and binarized diffusion models is challenging. On the one hand, diffusion models generate high-dimensional features at each layer to produce high-quality outputs in generative tasks, making element-wise alignment of these features difficult. On the other hand, quantization introduces unavoidable perturbations and information degradation in all weights and activations, making it hard to align the representations of full-precision and binarized diffusion models. Additionally, although some existing studies have proposed distillation methods for diffusion models, most focus on directly reducing the number of time steps rather than compressing features at the same time step.

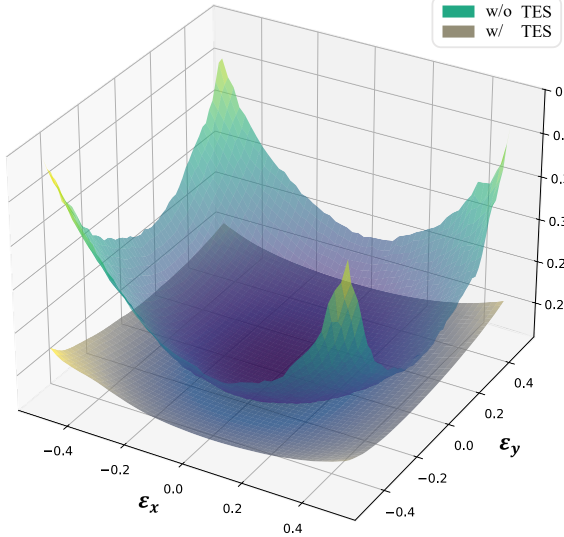


Fig. 6: The loss landscape of binarized diffusion models within one single inference timestep. We applied perturbations to the trainable weight parameters of BiDM+, *i.e.*, *w/* and *w/o* TES, along two orthogonal directions. The loss surface of the model with TES, which lies closer to the bottom, indicates that TES leads to greater robustness to latent weight variations, suggesting a more stable training process.

To address these issues, we propose the Salient-guided Binarization-aware Mimicking (SBM) strategy to facilitate the convergence of optimization for binarized diffusion models. We first construct a salient-guided representation for mimicking. Specifically, since the scaling of quantizers primarily depends on elements with greater magnitudes, they are typically considered an important indicator of salience in quantization within the representation [69]. In the representations of diffusion models, the salience of individual elements under binarization is not uniform. In mimicking the representation of the i-th block output, we focus more on the elements of the diffusion model that are more significant for quantization by optimizing the following expression:

$$\boldsymbol{r}_i^{\text{salient}} = \left|\boldsymbol{r}_i^{\text{fp}}\right| \left|\boldsymbol{r}_i^{\text{fp}} - \boldsymbol{r}_i^{\text{bi}}\right|, \tag{18}$$

where $\boldsymbol{r}_i^{\text{fp}}$ and $\boldsymbol{r}_i^{\text{bi}}$ represent the representations of the full-precision and binarized diffusion models, respectively, and $|\cdot|$ denotes taking the absolute value. Since the compared outputs will be quantized as activations for the next computation, a larger $\left|\boldsymbol{r}_i^{\text{bi}}\right|$ means that the element has higher

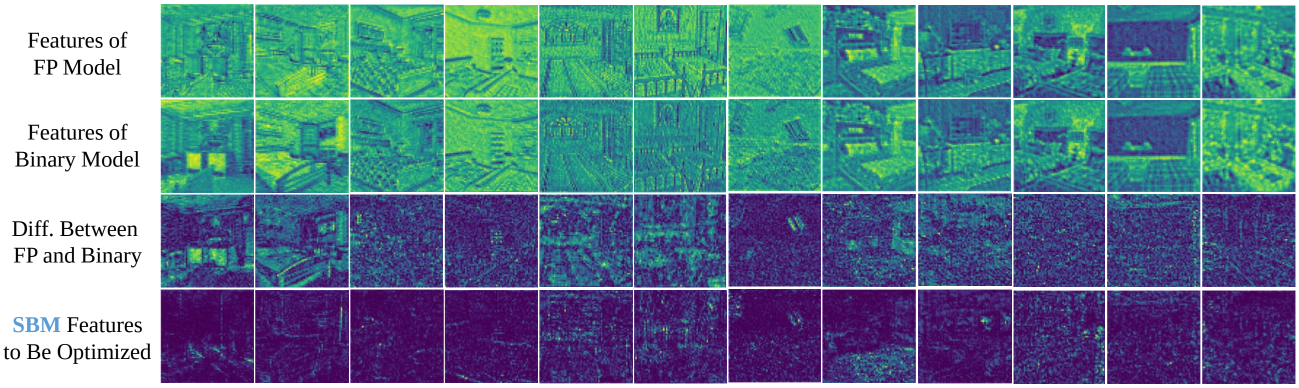


Fig. 7: Visualization of the last TimeStepBlock's output of the LDM model on the LSUN-bedroom dataset.

salience in the next quantization. This implies that we not only consider the numerical impact of binarization on the binarized diffusion model compared to the full-precision counterpart but also prioritize attention to representations that significantly affect binarization.

Furthermore, due to the introduction of highly discretized binarization functions, it is nearly impossible to fully align the representations of binarized diffusion models and their full-precision counterparts. These unavoidable minor errors introduce continuous disturbances into our optimization, affecting our ability to optimize and improve truly important representations through mimicking. Therefore, we introduce a statistical-based lenient mechanism in SBM. The representations for mimicking are constructed as follows:

$$\text{SBM: } \boldsymbol{r}_i^{\text{SBM}} = \left|\boldsymbol{r}_i^{\text{fp}}\right| \max\left(\left|\boldsymbol{r}_i^{\text{fp}} - \boldsymbol{r}_i^{\text{bi}}\right| - q^{\epsilon}_{\left|\boldsymbol{r}_i^{\text{fp}} - \boldsymbol{r}_i^{\text{bi}}\right|}, 0\right), \quad (19)$$

where q^ϵ represents the ϵ-th percentile for a certain distribution, with ϵ empirically set to 10% to ignore the impact of smaller numerical perturbations. This means that if the representation elements between the binarized and full-precision diffusion models only have a minor impact, we can completely ignore them to focus more on optimizing critical locations. This lenient mechanism effectively prevents minor perturbations from cumulatively causing optimization direction disturbances in representation mimicking.

With the proposed SBM representation, we use the following training loss to optimize binarized diffusion models:

$$\begin{aligned} \mathcal{L}^{\text{total}} &= \mathcal{L}^{\text{DM}} + \frac{\gamma}{2d+1}\sum_{i=1}^{2d+1}\mathcal{L}_i^{\text{SBM}} \\ &= \mathcal{L}^{\text{DM}} + \frac{\gamma}{2d+1}\sum_{i=1}^{2d+1}\text{mean}\left({\boldsymbol{r}_i^{\text{SBM}}}^2\right), \end{aligned} \quad (20)$$

where d represents the number of up-sampling or down-sampling blocks, thus producing a total of $2d+1$ intermediate features, including intermediate blocks. γ is a hyperparameter coefficient used to balance the loss terms, set at 5e-4. We visualize the SBM representations in the strategy described above. As shown in Fig. 7, the constructed SBM representation integrates quantization errors and salience factors and benefits from the lenient strategy to exclude the impact of minor perturbations. This implies that SBM allows for an accurate and focused optimization of binarized diffusion models, thus exhibiting better convergence.

4 Experiment

We evaluate BiDM+ on multiple datasets, including CIFAR-10 (32×32) [26], LSUN-Bedrooms (256×256) [76], LSUN-Churches (256×256) [76], and FFHQ (256×256) [25], using both Denoising Diffusion Implicit Models (DDIM) [20] and Latent-space Diffusion Models (LDM) [52]. Our evaluation metrics include Inception Score (IS), Fréchet Inception Distance (FID) [18], Sliding FID (sFID) [54], and Precision and Recall. We employ generic binarization techniques, *e.g.*, XNOR [51], DoReFa [84], ReActNet [37], state-of-the-art (SOTA) binarization methods, *e.g.*, ReSTE [66], and SOTA quantization approaches tailored for generative models, *e.g.*, BBCU [67], EfficientDM [16], and BiDM [81].

We utilize quantization-aware training (QAT) with pretrained diffusion models, training both quantizer parameters and latent weights. For CIFAR-10, the training configuration includes a learning rate of 6×10^{-5}, a batch size of 64, and 100K iterations with 100 sampling steps. For higher-resolution datasets, the learning rate is reduced to 2×10^{-5}, the batch size is set to 4, and the training extends to 200K iterations with 200 denoising steps. Generative performance is evaluated using IS, FID, sFID, and Precision-and-Recall metrics, with 50K randomly generated images used for computation after training. Theoretical inference efficiency is measured by estimating the number of operations (OPs), including bit-operations (BOPs) and floating-point operations (FLOPs) [3]. All experiments are conducted on an NVIDIA A100 40GB.

4.1 Accuracy Results

For DDIMs, initial experiments were conducted using the CIFAR-10 32×32 dataset. As detailed in Table 1, applying W1A1 binarization through baseline methods resulted in notable degradation. In contrast, BiDM+ exhibited significant enhancements across all metrics, including an improvement in the IS metric from 5.18 to 5.53 and a reduction in the FID metric by 7.85%.

For LDMs, our investigations included evaluations of LDM-4 on FFHQ 256×256 and LSUN-Bedrooms 256×256 datasets and LDM-8 on LSUN-Churches 256×256 dataset, using a DDIM sampler with 200 steps. The detailed results are documented in Table 2. Across these datasets, BiDM+ substantially outperformed the advanced baseline methods. For example, compared to BiDM, the existing

Model	Dataset	Method	#Bits	IS↑	FID↓	sFID↓	Precision↑
DDIM	CIFAR-10 32×32	FP	32/32	8.90	5.54	4.46	67.92
		XNOR++	1/1	2.23	251.14	60.85	44.98
		DoReFa	1/1	1.43	397.60	139.97	0.17
		ReActNet	1/1	3.35	231.55	119.80	18.37
		ReSTE	1/1	1.26	394.29	125.84	0.18
		XNOR	1/1	4.23	113.36	27.67	46.96
		BiDM	1/1	5.18	81.65	25.68	52.92
		BiDM+	1/1	**5.53**	**73.80**	**21.62**	**56.03**

TABLE 1: Binarization results for DDIM on CIFAR-10 datasets with 100 steps.

Model	Dataset	Method	#Bits	FID↓	sFID↓	Precision↑	Recall↑
LDM-4	FFHQ 256×256	FP	32/32	4.87	6.96	74.73	50.57
		XNOR++	1/1	379.49	320.64	0.00	0.00
		DoReFa	1/1	214.06	177.63	2.09	0.00
		ReActNet	1/1	147.88	141.31	3.36	0.69
		ReSTE	1/1	144.37	97.43	4.03	0.03
		XNOR	1/1	89.37	54.04	31.31	4.11
		BiDM	1/1	43.42	32.35	49.44	13.96
		BiDM+	1/1	**25.69**	**29.08**	**60.35**	**21.41**
LDM-4	LSUN-Bedrooms 256×256	FP	32/32	2.99	7.08	65.02	47.54
		XNOR++	1/1	319.66	184.75	0.00	0.00
		BBCU	1/1	236.07	89.66	0.59	5.66
		EfficientDM	1/1	194.45	113.24	0.99	9.20
		DoReFa	1/1	188.30	89.28	0.86	0.18
		ReActNet	1/1	154.74	61.50	4.63	9.30
		ReSTE	1/1	59.44	42.16	12.06	2.92
		XNOR	1/1	106.62	56.81	6.82	5.22
		BiDM	1/1	22.74	17.91	33.54	19.90
		BiDM+	1/1	**19.37**	**21.50**	**39.38**	**22.52**
LDM-8	LSUN-Churches 256×256	FP	32/32	4.36	16.00	74.64	48.98
		XNOR++	1/1	292.48	168.65	0.02	0.00
		DoReFa	1/1	162.06	95.37	7.85	0.74
		ReActNet	1/1	56.39	54.68	45.13	2.06
		ReSTE	1/1	47.88	52.44	51.98	3.34
		XNOR	1/1	42.87	49.24	51.53	4.28
		BiDM	1/1	29.70	45.14	55.75	14.80
		BiDM+	1/1	**27.19**	**32.03**	**51.12**	**14.90**

TABLE 2: Quantization results for LDM on FFHQ, LSUN-Bedrooms, and LSUN-Churches datasets.

SOTA binarization method for diffusion models, with reductions in FID by 17.73%, 3.37%, and 2.51% on the FFHQ, LSUN-Bedrooms, and LSUN-Churches datasets, respectively. Specifically, compared to XNOR++, which utilizes fixed activation scaling factors, leading to a constrained dynamic range of activations, BiDM+ enhances the flexibility of the quantization values and scalings, especially by making the tiny convolution k and learnable clipping boundaries $\mu_x^{\min}$ and $\mu_x^{\max}$. This led to significant enhancements, with the FID metric decreasing from 43.42 to 25.69 on FFHQ and from 22.74 to 19.37 on LSUN-Bedrooms. Additionally, compared to the SOTA generic binarization method, ReSTE, our BiDM+ demonstrated superior improvements across multiple metrics, particularly on the LSUN-Bedrooms dataset. In the realm of quantization-aware training for diffusion models, EfficientDM presents a relevant comparison. It incorporates TALSQ to manage variations in activation range, yet struggles under the extreme conditions of W1A1 binarization, possibly due to its quantizer failing to adapt to the binary constraints adequately. We also compare BiDM+ with BBCU, a binarization method tailored more for generative models like diffusion models rather than discriminative models. However, BBCU exhibits significant limitations when applied to diffusion models that are evidenced by a severe 236.07 FID metric, while BiDM+ is only 19.37.

The results show that the broad activation range and flexible representational capacity of diffusion models pose significant challenges to traditional binarization methods. While these existing methods' optimization strategies, often not designed for the unique demands of generative tasks in binarized diffusion models, tend to deliver only conventional and suboptimal results. The comprehensive improvements from the operator, structure, and optimization perspectives led BiDM+ to achieve SOTA performance.

4.2 Efficiency Results

Our analysis of the diffusion model's inference efficiency under full binarization illustrates that BiDM+ incurs only a marginal increase in floating-point additions for timestep connections compared to conventional binarization methods

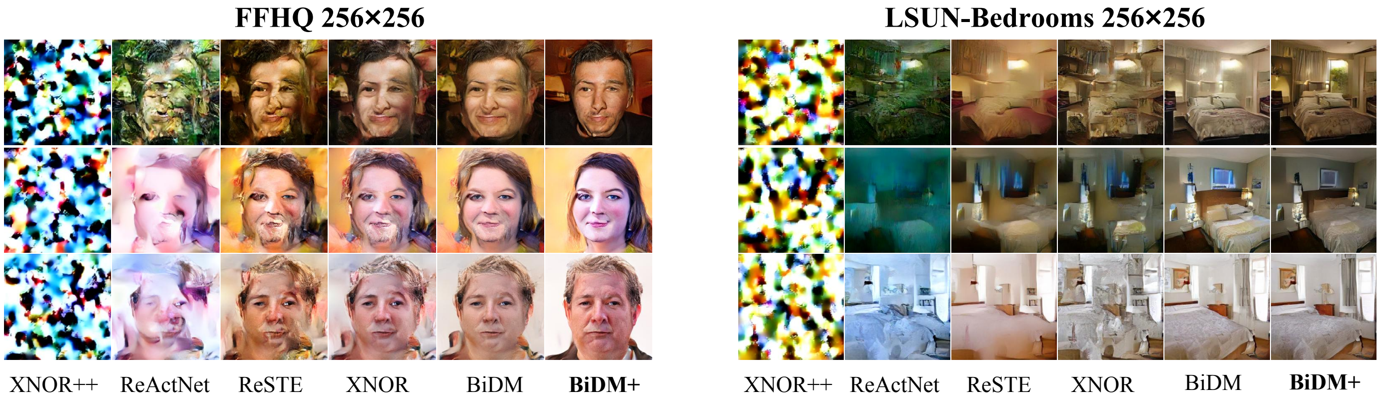


Fig. 8: Visualization of samples generated by the W1A1 baseline and our BiDM+. BiDM+ is the first fully binarized diffusion model method capable of generating viewable images, significantly surpassing advanced binarization methods.

Method	#Bits	Storage	Computation			Accuracy
		Size (MB)	BOPs ($\times 10^9$)	FLOPs ($\times 10^9$)	OPs ($\times 10^9$)	FID↓
FP	32/32	1045.4	-	96.00	96.00	4.87
XNOR	1/1	37.3 (↓96.4%)	92.1	0.38	1.82 (↓98.1%)	89.37
BiDM	1/1	37.3 (↓96.4%)	92.1	0.38	1.82 (↓98.1%)	43.42
BiDM+	1/1	**37.3 (↓96.4%)**	**92.1**	**0.38**	**1.82 (↓98.1%)**	**25.69**

TABLE 3: Inference efficiency of our proposed BiDM+ of LDM-4 on FFHQ.

Method	#Bits	FID↓	sFID↓	Prec.↑	Recall↑
Vanilla (XNOR)	1/1	379.49	320.64	0.00	0.00
+FPB	1/1	43.85	32.82	52.70	15.78
+FPB +TES	1/1	41.90	30.95	53.36	16.28
+FPB +TES +SBM (**BiDM+**)	1/1	**25.69**	**29.08**	**60.35**	**21.41**

TABLE 4: Ablation result of each proposed component in BiDM+ on LDM-4 for FFHQ.

such as XNOR-Net. The majority of operations, including convolutions, remain unchanged. The computational requirement for executing a floating-point convolution at a depth of 1 for scaling factors is minimal. As detailed in Table 3, BiDM+ achieves the same 28.0× memory efficiency and 52.7× computational savings as the XNOR baseline while significantly enhancing generation capabilities and reducing the FID from 89.37 to 25.69 on FFHQ.

4.3 Ablation Results

We conducted a comprehensive ablation study on LDM-4 on the FFHQ 256 × 256 dataset to evaluate the effectiveness of each component proposed in BiDM+. We analyzed the effectiveness of FPB, TES, and SBM techniques and summarized the results in Table 4. Applying these three techniques to the binarization of LDM-4, the results show that each technique improves the model performance. After adding the FPB method, both FID and sFID are significantly reduced, from 379.49 and 43.85 to 320.64 and 32.82, respectively. After further integrating TES, FID is further reduced from 41.90, and other quantitative metrics are also significantly improved. Finally, applying SBM in optimization pushes the binarized diffusion models down to 25.69. These results validate the motivation of our approach, which is to achieve accurate optimization by introducing learnable and flexible projections to achieve diverse binarized representations, constructing evolvable connections between consecutive time steps to achieve stable propagation, and making binarized model features simulate full-precision features in optimization.

5 Conclusion

We introduce BiDM+, a highly accurate binarized diffusion model that successfully bridges the gap between computational efficiency and high-quality generative performance. By integrating Flexible Projection Binarization, Timestep-aware Evoluted Structure, and Salient-guided Binarization-aware Mimicking, BiDM+ overcomes the inherent limitations of traditional binarized diffusion approaches. The model not only reduces storage requirements by up to 96.4% and computational operations by 98.1% compared to 32-bit models but also achieves SOTA results, as demonstrated by a significant FID improvement. These advancements confirm that precise binarization strategies, when paired with timestep-aware architectural innovations and informed optimization techniques, can effectively preserve the representation capability for high-fidelity image synthesis. Our findings establish a solution for deploying diffusion models in resource-constrained scenarios while opening avenues for future research on ultra-efficient generative models.

References

[1] Thomas Bird, Friso H Kingma, and David Barber. Reducing the computational cost of deep generative models with binary neural networks. *arXiv:2010.13476*, 2020.

[2] Adrian Bulat and Georgios Tzimiropoulos. Xnor-net++: Improved binary neural networks. *arXiv:1909.13863*, pages 1–12, 2019.

[3] Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2020.

[4] Zheng Chen, Haotong Qin, Yong Guo, Xiongfei Su, Xin Yuan, Linghe Kong, and Yulun Zhang. Binarized diffusion model for image super-resolution. *arXiv:2406.05723*, 2024.

[5] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, pages 1–11, 2016.

[6] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[7] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. In *International Conference on Learning Representations*, pages 1–12, 2019.

[8] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *arXiv:2305.10924*, 2023.

[9] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[10] Ruihao Gong, Yifu Ding, Zining Wang, Chengtao Lv, Xingyu Zheng, Jinyang Du, Haotong Qin, Jinyang Guo, Michele Magno, and Xianglong Liu. A survey of low-bit large language models: Basics, systems, and algorithms. *arXiv:2409.16694*, 2024.

[11] Guangyu Guo, Longfei Han, Le Wang, Dingwen Zhang, and Junwei Han. Semantic-aware knowledge distillation with parameter-free feature uniformization. *Visual Intelligence*, 1(1):6, 2023.

[12] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International Conference on Machine Learning*, pages 1737–1746, 2015.

[13] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.

[14] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv:1510.00149*, pages 1–14, 2015.

[15] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

[17] Yefei He, Zhenyu Lou, Luoming Zhang, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Bivit: Extremely compressed binary vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5651–5663, 2023.

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

[19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022.

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

[21] Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms. *arXiv:2402.04291*, 2024.

[22] Wei Huang, Xudong Ma, Haotong Qin, Xingyu Zheng, Chengtao Lv, Hong Chen, Jie Luo, Xiaojuan Qi, Xianglong Liu, and Michele Magno. How good are low-bit quantized llama3 models? an empirical study. *arXiv:2404.14047*, 2024.

[23] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in Neural Information Processing Systems*, 29:1–9, 2016.

[24] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv:2104.01409*, 2021.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. pages 1–60, 2009.

[27] Phuoc-Hoan Charles Le and Xinlin Li. Binaryvit: pushing binary vision transformers towards convolutional models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4664–4673, 2023.

[28] Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdqunat: Absorbing outliers by low-rank components for 4-bit diffusion models. *arXiv preprint arXiv:2411.05007*, 2024.

[29] Xinyu Li, Danni Ai, Hong Song, Jingfan Fan, Tianyu Fu, Deqiang Xiao, Yining Wang, and Jian Yang. Stqd-det: Spatio-temporal quantum diffusion model for real-time coronary stenosis detection in x-ray angiography. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

[31] Yanjing Li, Sheng Xu, Xianbin Cao, Xiao Sun, and Baochang Zhang. Q-dm: An efficient low-bit quantized diffusion model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[32] Zezeng Li, Shenghao Li, Zhanpeng Wang, Na Lei, Zhongxuan Luo, and David Xianfeng Gu. Dpm-ot: a new diffusion probabilistic model based on optimal transport. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 22624–22633, 2023.

[33] Mingbao Lin, Rongrong Ji, Zihan Xu, Baochang Zhang, Fei Chao, Chia-Wen Lin, and Ling Shao. Siman: Sign-to-magnitude network binarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6277–6288, 2022.

[34] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *ECCV*, 2024.

[35] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv:2202.09778*, 2022.

[36] Zechun Liu, Wenhan Luo, Baoyuan Wu, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Binarizing deep network towards real-network performance. *International Journal of Computer Vision*, 128:202–219, 2020.

[37] Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In *Proceedings of the European Conference on Computer Vision*, pages 143–159. Springer, 2020.

[38] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting 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*, pages 722–737. Springer, 2018.

[39] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv:2211.01095*, 2022.

[40] Weijian Luo. A comprehensive survey on knowledge distillation of diffusion models. *arXiv:2304.04262*, 2023.

[41] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. *arXiv:2312.00858*, 2023.

[42] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9117–9125, 2023.

[43] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[44] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv:2103.16091*, 2021.

[45] Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *International Conference on Artificial Intelligence and Statistics*, pages 4474–4484. PMLR, 2020.

[46] William Peebles and Saining Xie. Scalable diffusion models with

transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[47] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[48] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2250–2259, 2020.

[49] Haotong Qin, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno. Accurate lora-finetuning quantization of llms via information retention. *arXiv:2402.05445*, 2024.

[50] Haotong Qin, Xiangguo Zhang, Ruihao Gong, Yifu Ding, Yi Xu, and Xianglong Liu. Distribution-sensitive information retention for accurate binary neural network. *International Journal of Computer Vision*, 131(1):26–47, 2023.

[51] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *Proceedings of the European Conference on Computer Vision*, pages 525–542. Springer, 2016.

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

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

[55] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv:2202.00512*, 2022.

[56] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1972–1981, 2023.

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

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

[59] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023.

[60] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv:2011.13456*, 2020.

[62] Jia-Mu Sun, Tong Wu, and Lin Gao. Recent advances in implicit representation-based 3d shape generation. *Visual Intelligence*, 2(1):9, 2024.

[63] Kai Wang, Dongwen Tang, Boya Zeng, Yida Yin, Zhaopan Xu, Yukun Zhou, Zelin Zang, Trevor Darrell, Zhuang Liu, and Yang You. Neural network diffusion. *arXiv:2402.13144*, 2024.

[64] Peisong Wang, Xiangyu He, Gang Li, Tianli Zhao, and Jian Cheng. Sparsity-inducing binarized neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12192–12199, 2020.

[65] Ze Wang, Jiang Wang, Zicheng Liu, and Qiang Qiu. Binary latent diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22576–22585, 2023.

[66] Xiao-Ming Wu, Dian Zheng, Zuhao Liu, and Wei-Shi Zheng. Estimator meets equilibrium perspective: A rectified straight through estimator for binary neural networks training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17055–17064, 2023.

[67] Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Basic binary convolution unit for binarized image restoration network. *arXiv:2210.00405*, 2022.

[68] Mengfei Xia, Yu Zhou, Ran Yi, Yong-Jin Liu, and Wenping Wang. A diffusion model translator for efficient image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[69] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR, 2023.

[70] Yisong Xiao, Aishan Liu, Tianyuan Zhang, Haotong Qin, Jinyang Guo, and Xianglong Liu. Robustmq: benchmarking robustness of quantized models. *Visual Intelligence*, 1(1):30, 2023.

[71] Yixing Xu, Kai Han, Chang Xu, Yehui Tang, Chunjing Xu, and Yunhe Wang. Learning frequency domain approximation for binary neural networks. *Advances in Neural Information Processing Systems*, 34:25553–25565, 2021.

[72] Zihan Xu, Mingbao Lin, Jianzhuang Liu, Jie Chen, Ling Shao, Yue Gao, Yonghong Tian, and Rongrong Ji. Recu: Reviving the dead weights in binary neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5198–5208, 2021.

[73] Zhekai Xu, Haohong Shang, Shaoze Yang, Ruiqi Xu, Yichao Yan, Yixuan Li, Jiawei Huang, Howard C Yang, and Jianjun Zhou. Hierarchical painter: Chinese landscape painting restoration with fine-grained styles. *Visual Intelligence*, 1(1):19, 2023.

[74] Yichao Yan, Zanwei Zhou, Zi Wang, Jingnan Gao, and Xiaokang Yang. Dialoguenerf: Towards realistic avatar face-to-face conversation video generation. *Visual Intelligence*, 2(1):24, 2024.

[75] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7308–7316, 2019.

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

[77] Zelong Zeng, Fan Yang, Hong Liu, and Shin'ichi Satoh. Improving deep metric learning via self-distillation and online batch diffusion process. *Visual Intelligence*, 2(1):18, 2024.

[78] Jianhao Zhang, Yingwei Pan, Ting Yao, He Zhao, and Tao Mei. Dabnn: A super fast inference framework for binary neural networks on arm devices. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2272–2275, 2019.

[79] Yulun Zhang, Haotong Qin, Zixiang Zhao, Xianglong Liu, Martin Danelljan, and Fisher Yu. Flexible residual binarization for image super-resolution. In *Forty-first International Conference on Machine Learning*, 2024.

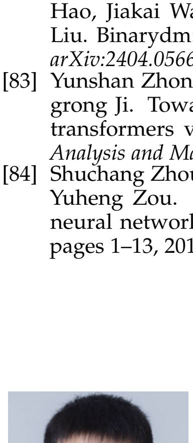
[80] Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *International Conference on Machine Learning*, pages 7543–7552. PMLR, 2019.

[81] Xingyu Zheng, Xianglong Liu, Yichen Bian, Xudong Ma, Yulun Zhang, Jiakai Wang, Jinyang Guo, and Haotong Qin. Bidm: Pushing the limit of quantization for diffusion models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[82] Xingyu Zheng, Haotong Qin, Xudong Ma, Mingyuan Zhang, Haojie Hao, Jiakai Wang, Zixiang Zhao, Jinyang Guo, and Xianglong Liu. Binarydm: Towards accurate binarization of diffusion model. *arXiv:2404.05662*, 2024.

[83] Yunshan Zhong, You Huang, Jiawei Hu, Yuxin Zhang, and Rongrong Ji. Towards accurate post-training quantization of vision transformers via error reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[84] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv:1606.06160*, pages 1–13, 2016.



Haotong Qin is a Postdoctoral Researcher at the Center for Project-Based Learning (PBL) D-ITET at ETH Zürich. He received his B.S. and Ph.D. degrees at Beihang University. His research interests include model compression and deployment toward efficient deep learning in real-world scenarios. He has published 35 papers in top-tier journals and conferences, such as IEEE TPAMI, IEEE TNNLS, IJCV, ICML, NeurIPS, ICLR, and CVPR. He serves as Guest Editor for Neural Networks, *etc.*, Area Chair in NeurIPS, ACM MM, AISTATS, *etc.*, and Reviewer for IEEE TPAMI, IEEE TIP, IEEE TNNLS, IJCV, NeurIPS, ICML, ICLR, *etc.*

Xianglong Liu (Senior Member, IEEE) is currently a Professor, serves as the vice dean of the School of Computer Science and Engineering at Beihang University, and is also the deputy director of the State Key Laboratory of Complex and Critical Software Environment. He received his B.S. and Ph.D. degrees under the supervision of Prof. Wei Li and visited the DVMM Lab at Columbia University as a joint PhD student supervised by Prof. Shih-Fu Chang. He is the recipient of the China National Excellent Youth Science Fund. He has published over 100 papers in top conferences/journals in artificial intelligence and information security, such as NeurIPS, ICLR, CVPR, ICCV, CSS, and IJCV. He serves as Associate Editor and Guest for several SCI journals like Pattern Recognition and IET Image Processing, and as Promotion Editor for journals like Frontiers of Computer Science and Acta Aeronautica et Astronautica Sinica. He serves as Area Chair in top conferences such as AAAI and ACM MM.

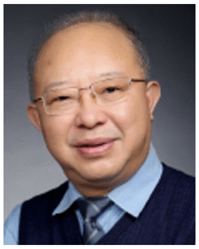
Xingyu Zheng is a PhD Candidate at Beihang University, supervised by Prof. Xianglong Liu. He is mainly focused on the practical application of AIGC models, including inference acceleration and model quantization. He obtained a BEng degree from Beihang University, and awarded the "Honor Student" and "Excellent Undergraduate Graduation Thesis". His work has been published in top-tier conferences such as ICLR, NeurIPS, ICML and ACM MM. He has organized workshops in conferences like IJCAI. He serves as reviewer for Pattern Recognition, IEEE TNNLS, NeurIPS, *etc*.

Haoran Chu is an undergraduate student majoring in Computer Science and Engineering at Beihang University. His research interests include model quantization and image generation, as well as other tasks in computer vision.

Jinyang Guo is an Assistant Professor at the State Key Laboratory of Complex & Critical Software Environment, Institute of Artificial Intelligence, Beihang University, China. Previously, he obtained his B.Eng (Hons1) degree from the School of Electrical and Telecommunication, The University of New South Wales, Australia, and later received his Ph.D. from the School of Electrical and Information Engineering, The University of Sydney. His research interests include efficient and scalable AI computing (*e.g.*, model-efficiency, data-efficiency, label-efficiency), neuromorphic computing (*e.g.*, spiking neural networks), and AI4Science (*e.g.*, AI for quantum computing).

Yulun Zhang received a B.E. degree from the School of Electronic Engineering, Xidian University, China, in 2013, an M.E. degree from the Department of Automation, Tsinghua University, China, in 2017, and a Ph.D. degree from the Department of ECE, Northeastern University, USA, in 2021. He is an associate professor at Shanghai Jiao Tong University, Shanghai, China. He was a postdoctoral researcher at Computer Vision Lab, ETH Zürich, Switzerland. His research interests include image/video restoration and synthesis, biomedical image analysis, model compression, multimodal computing, large language model, and computational imaging. He is/was an Area Chair for CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, IJCAI, ACM MM, and a Senior Program Committee (SPC) member for IJCAI and AAAI.

Yawei Li is currently a Lecturer at ETH Zürich. His research primarily focuses on efficient deep learning and AI algorithms and systems, with applications spanning vision, language, and biosignals. He is particularly interested in topics such as developing foundation models for biosignals, designing efficient deep neural networks for vision tasks, and exploring model quantization and pruning techniques that benefit both vision and language applications. Furthermore, his work involves the deployment of models on RISC-V cores and advancing DNN integration in embedded systems.



Bo Li is currently a Changjiang distinguished professor with the School of Computer Science and Engineering, Beihang University. He is a recipient of the National Science Fund for Distinguished Young Scholars. He is currently the dean of AI Research Institute, Beihang University. He is the chief scientist of National 973 Program and the principal investigator of the National Key Research and Development Program. He has published more than 100 papers in top journals and conferences and held more than 50 patents.

Luca Benini (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1997. He has served as the Chief Architect for the Platform2012/ STHORM Project with STMicroelectronics, Grenoble, France, from 2009 to 2013. Currently, he holds the Chair of digital circuits and systems with ETH Zürich, Zürich, Switzerland, and is a Full Professor with the University of Bologna, Bologna, Italy. He has published more than 1000 peer-reviewed articles and five books. His current research interest includes energy-efficient computing systems, smart sensing micro-systems and machine learning hardware. He is a fellow of ACM and a member of the Academia Europaea. He was a recipient of the 2016 IEEE CAS Mac Van Valkenburg award, the 2019 IEEE Donald O. Pederson Best Paper, the 2020 IEEE TCAS-I Darlington Best Paper, the 2020 IEEE Very Large Scale Integration Systems Best Paper and the 2020 ACM/IEEE A. Richard Newton Award.

Michele Magno (Senior Member, IEEE) is currently a Privatdozent with D-ITET, ETH Zürich, where he has been leading the Center for Project-Based Learning since 2020. He is also a senior member of IEEE. He received master's and Ph.D. degrees in electronic engineering from the University of Bologna, Bologna, Italy, in 2004 and 2010, respectively. Since 2013, he has been with ETH Zürich, Switzerland, and has become a Visiting Lecturer or a Professor at several universities, namely, the University of Nice Sophia, France; Enssat Lannion, France; the University of Bologna, Italy; and Mid University Sweden, Sweden; where is a Full Visiting Professor with the Department of Electrical Engineering. Some of his publications were awarded Best Papers at several IEEE conferences.