ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2025.0081

TP-ViT: truncated uniform-log2 quantizer and progressive bit-decline reconstruction for vision Transformer quantization

Xichuan ZHOU, Sihuan ZHAO, Rui DING, Jiayu SHI, Jing NIE, Lihui CHEN, Haijun LIU✉

School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Abstract: Vision Transformers (ViTs) have achieved remarkable success across various artificial intelligence-based computer vision applications. However, their demanding computational and memory requirements pose significant challenges for deployment on resource-constrained edge devices. Although post-training quantization (PTQ) provides a promising solution by reducing model precision with minimal calibration data, aggressive low-bit quantization typically leads to substantial performance degradation. To address this challenge, we present the truncated uniform-log2 quantizer and progressive bit-decline reconstruction method for vision Transformer quantization (TP-ViT). It is an innovative PTQ framework specifically designed for ViTs, featuring two key technical contributions: (1) truncated uniform-log2 quantizer, a novel quantization approach which effectively handles outlier values in post-Softmax activations, significantly reducing quantization errors; (2) bit-decline optimization strategy, which employs transition weights to gradually reduce bit precision while maintaining model performance under extreme quantization conditions. Comprehensive experiments on image classification, object detection, and instance segmentation tasks demonstrate TP-ViT's superior performance compared to state-of-the-art PTQ methods, particularly in challenging 3-bit quantization scenarios. Our framework achieves a notable 6.18 percentage points improvement in top-1 accuracy for ViT-small under 3-bit quantization. These results validate TP-ViT's robustness and general applicability, paving the way for more efficient deployment of ViT models in computer vision applications on edge hardware.

Key words: Vision Transformers; Post-training quantization; Block reconstruction; Image classification; Object detection; Instance segmentation

1 Introduction

The vision Transformer (ViT) architecture (Dosovitskiy et al., 2021) has revolutionized computer vision by leveraging its innovative self-attention mechanism, establishing itself as a formidable alternative to traditional convolutional neural networks (CNNs). ViT-based models have achieved state-of-the-art (SOTA) performance across diverse artificial intelligence (AI)-driven engineering applications, including but not limited to image classification (Mahmood et al., 2024; Zhang ZC et al., 2024), object detection (Chen et al., 2024; Gao et al., 2024), semantic segmentation (Li MH et al., 2024; Zheng et al., 2025), and image processing (Tian et al., 2024; Xia et al., 2025). The key advantage of ViT lies in its ability to capture global contextual information, which has significantly advanced vision-based automation and the development of intelligent systems.

Despite their remarkable performance, Transformer-based models like ViT face significant deployment challenges in real-world engineering applications due to their large model size and high computational demands. The self-attention mechanism, while highly effective in capturing global dependencies, introduces substantial memory and processing overhead, particularly when processing high-resolution images, where computational complexity grows quadratically with input size (Zamir et al., 2022). These limitations pose critical obstacles for real-time deployment on resource-constrained platforms, which often operate under strict power and latency constraints. To address these challenges, neural network quantization has emerged as a promising solution, effectively reducing model size and computational requirements while maintaining acceptable accuracy. By compressing model parameters into lower-bit representations, quantization significantly enhances the feasibility of deploying ViT models in practical and large-scale applications.

✉ Haijun LIU, haijun_liu@126.com
Xichuan ZHOU, https://orcid.org/0000-0002-3304-3045
Haijun LIU, https://orcid.org/0000-0001-5782-4543
CLC number: TP391.4

Received: Oct. 13, 2025; Revision accepted: Jan. 15, 2026;
Crosschecked: Jan. 16, 2026

There are two primary types of quantization in neural networks: quantization-aware training (QAT) (Kim et al., 2022; Nagel et al., 2022; Zhong et al., 2022; Liu SY et al., 2023) and post-training quantization (PTQ) (Li YH et al., 2021; Jiang YF et al., 2025; Zhong et al., 2026). QAT incorporates quantization constraints during the training phase, requiring retraining on labeled datasets to mitigate accuracy loss. Although effective, this approach imposes significant computational burdens, particularly for large Transformer models, making it resource-intensive and time-consuming. In contrast, PTQ operates on pre-trained models, requiring only a small amount of unlabeled calibration data to perform quantization. This method provides a computationally efficient alternative, enabling rapid compression of Transformer-based models without demanding additional training resources or extensive datasets.

Quantizing ViTs using PTQ to low-bit widths presents significant challenges, often resulting in substantial accuracy degradation. A primary obstacle lies in quantizing the activations of specialized layers such as LayerNorm and Softmax, whose unique data distributions render conventional quantization techniques ineffective. To mitigate these issues, recent advances have introduced layer-specific quantization strategies tailored to preserve model performance. For post-LayerNorm activations, RepQ-ViT (Li ZK et al., 2023) employs a channel-wise quantization approach to capture inter-channel variations, followed by reparameterizing into a layer-wise format. This method maintains accuracy while optimizing inference efficiency. For post-Softmax activations, PTQ4ViT (Yuan et al., 2022) employs a twin uniform quantizer (UQ) to accommodate the non-Gaussian distribution of attention maps better, while FQ-ViT (Lin Y et al., 2023) leverages a log2 quantizer (LQ) to adapt to the long-tailed distribution characteristic of attention outputs. Furthermore, I&S-ViT (Zhong et al., 2026) proposes the shift-uniform-log2 quantizer (SULQ) to enhance the quantization efficiency. However, SULQ does not account for the presence of outliers, which may introduce quantization errors.

Despite these advancements, current layer-specific activation quantization techniques alone are insufficient for maintaining model accuracy at extremely low-bit widths. As evidenced by empirical results, RepQ-ViT (Li ZK et al., 2023) achieves 65.05% top-1 accuracy on ViT-S with 4-bit quantization but suffers catastrophic performance degradation (0.43%) when compressed to 3-bit. The dramatic accuracy collapse is primarily caused by excessive numerical precision loss during aggressive quantization, leading to substantial quantization errors in both activations and weights. To address this, reconstruction optimization-based PTQ methods have been explored to minimize quantization errors and enhance accuracy.

Among these methods, BRECQ (Li YH et al., 2021) and I&S-ViT (Zhong et al., 2026) both employ multi-stage block reconstruction strategies. BRECQ has two stages: the first quantizes weights to the target bit-width and then performs block reconstruction using full-precision activations and the quantized weights; the second quantizes activations to the target bit-width and then performs block reconstruction using the quantized activations and weights. I&S-ViT follows a similar approach: it first quantizes the activations to the target bit-width, then uses full-precision weights and quantized activations to block reconstruction, and finally performs block reconstruction after quantizing the weights to the target bit-width. However, traditional block reconstruction methods directly quantize both activations and weights to the target bit-width, but these methods may bring more quantization errors.

To address the challenges, we propose TP-ViT, a novel PTQ framework that effectively reduces quantization errors by introducing two key innovations: the truncated uniform-log2 quantizer (TULQ) and the bit-decline optimization strategy (BDOS).

First, we observe that SULQ (Zhong et al., 2026) improves post-Softmax activation quantization efficiency; however, its log2 transformation introduces outliers that are not properly handled in the subsequent uniform quantization, potentially leading to increased quantization errors. The conventional UQ (Li RD et al., 2019) employs percentile truncation to mitigate outlier effects by identifying and removing potential outliers through a percentile coefficient. However, as demonstrated in Fig. 1, directly applying this approach to SULQ proves inflexible (inappropriately addressing some outliers) and yields suboptimal performance. Therefore, we propose TULQ, which introduces two adjustable coefficients to correctly control the truncated range. This dual-coefficient mechanism could effectively reduce outlier impact in activation quantization, thereby minimizing quantization errors and enhancing overall quantized model performance.

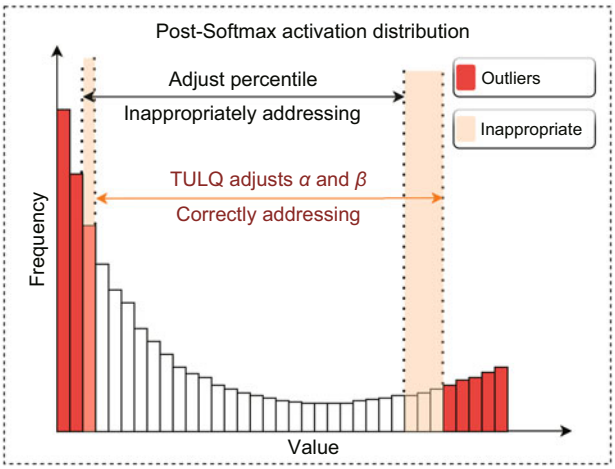

Fig. 1 Illustration of the proposed TULQ

Furthermore, existing PTQ methods employing multi-stage block reconstruction optimization, including BRECQ (Li YH et al., 2021) and the smooth optimization strategy (SOS) (Zhong et al., 2026), typically adopt a direct quantization approach where both weights and activations are immediately quantized to the target bit-width, as demonstrated in Fig. 2a. However, this abrupt quantization leads to substantial error accumulation, significantly degrading model accuracy. Therefore, we introduce BDOS, which gradually quantizes the model to the target bit to reduce quantization errors. As illustrated in Fig. 2b, BDOS introduces transition weights to quantize model weights to the target bit level progressively. This gradual quantization process effectively minimizes quantization errors, enabling the model to preserve its classification

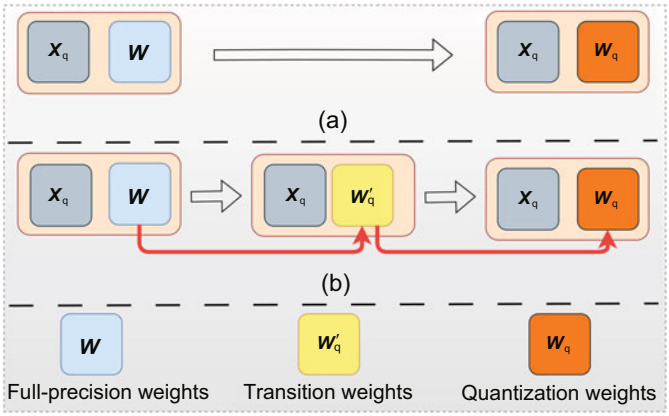

Fig. 2 Illustration of the proposed BDOS: (a) traditional block reconstruction; (b) our proposed BDOS. The description of each parameter is shown in Section 3

accuracy better than approaches that apply direct quantization to the final bit level. By integrating TULQ and BDOS, TP-ViT effectively enhances the robustness of ViTs under low-bit quantization, bridging the performance gap between full-precision and quantized models.

2 Related works

2.1 Vision Transformers

ViT and its variants have emerged as pivotal backbone networks in the computer vision community. ViT (Dosovitskiy et al., 2021) first introduces the self-attention mechanism into image classification, replacing convolutional layers with Transformer blocks, thereby achieving competitive results. Afterward, several variants are proposed to improve the performance further. DeiT (Touvron et al., 2021) alleviates dependence on large-scale training datasets through data augmentation and knowledge distillation techniques. Meanwhile, the Swin Transformer (Liu Z et al., 2021) introduces a shifted windows mechanism to enhance local attention through hierarchical feature maps and information exchange between windows. However, the high performance of ViTs is built on a high computational load, which hinders their application in edge devices (Mehta and Rastegari, 2022; Zhang JN et al., 2022).

2.2 ViT quantization

Model quantization, a crucial technique for model compression, involves converting floating-point weights and activations into lower-bit representations. It is primarily divided into QAT and PTQ. QAT (Zhong et al., 2022; Liu SY et al., 2023; Zhou et al., 2023) requires retraining of models on the entire training dataset to minimize performance degradation. In contrast, PTQ methods only require calibration on a small-scale dataset, making the methods suitable for rapid deployment, but maintaining accuracy advantages at low-bit widths remains challenging.

First, in the ViT architecture, key components such as Softmax and LayerNorm have significant impacts on the quantization performance of the model. This indicates that traditional UQs may not be suitable for such layers. Given this, some researchers have proposed a series of more targeted quantitative paradigms to improve the quantitative performance of ViT models. PTQ4ViT (Yuan et al., 2022) is proposed using dual-grained quantization to address data distribution disparities. In contrast, FQ-ViT (Lin Y et al., 2023) employs powers of two quantization for LayerNorm and symmetric quantization for Softmax. APQ-ViT (Ding et al., 2022) focuses on preserving the Matthew effect in Softmax to maintain stability in attention computations. To enhance quantization efficiency, I&S-ViT (Zhong et al., 2026) proposes SULQ, effectively addressing the limitations of standard log2 quantization. TSPTQ-ViT (Tai et al., 2023) proposes a two-scaled quantization scheme for post-Softmax and post-GELU activations, mitigating the high outlier sensitivity in ViT activation distributions by assigning separate quantization ranges to different activation regions. Meanwhile, ADFQ-ViT (Jiang YF et al., 2025) introduces a per-patch outlier-aware quantizer, designed to handle outliers and irregular distributions, particularly in post-LayerNorm activations. Additionally, AIQViT (Jiang RQ et al., 2025) leverages a low-rank compensation mechanism and a dynamic focusing quantizer, allowing ViTs to better adapt to different tasks. Additionally, several recent works adopt mixed-precision strategies to preserve accuracy under low-bit constraints; for example, AMP-ViT (Tai and Wu, 2025) automatically assigns higher bit-widths to sensitive layers, achieving improved performance without increasing overall computational cost significantly. Among them, the SULQ quantizer of I&S-ViT (Zhong et al., 2026) improves post-Softmax activation quantization efficiency; however, the log2 transformation would introduce outliers, which may affect the performance of the quantization model. To tackle this, we propose a new quantization scheme specifically designed to reduce outlier effects and enhance performance.

Furthermore, the reconstruction-based PTQ methods can mitigate accuracy degradation. BRECQ (Li YH et al., 2021) introduces block-wise reconstruction to refine quantized weights. Building upon this, QDrop (Wei et al., 2023) improves low-bit quantization by integrating activation quantization with dropout during reconstruction. Expanding on block-wise reconstruction, PD-Quant (Liu JW et al., 2023) leverages global information from the full-precision model for quantization parameter optimization. Furthermore, outlier-aware (Ma et al., 2024) introduces reconstruction granularity, revealing its impact on outlier mitigation. However, existing reconstruction-based methods directly quantize weights and activations to the target bit-width in a single step, potentially leading to quantization errors. To address this, we explore a progressive quantization optimization approach, aiming to reduce quantization errors and enhance model performance.

3 Preliminaries

3.1 Structure of ViTs

The input image I is first divided into N flattened two-dimensional (2D) patches. Each patch is then projected into a vector space $\mathbb{R}^D$ using an embedding layer, resulting in the embedding matrix $\boldsymbol{X}_0 \in \mathbb{R}^{N\times D}$. The embedding $\boldsymbol{X}_0$ is then fed into L stacked Transformer blocks. Each block consists of the multi-head self-attention (MHSA) module, the LayerNorm

module, and the multi-layer perceptron (MLP) module. For the l^{th} Transformer block, when the input is $\boldsymbol{X}_l$ and the output is $\boldsymbol{Y}_l$, the computations are as follows:

$$\begin{cases} \boldsymbol{A}_l = \text{MHSA}(\text{LayerNorm}(\boldsymbol{X}_l)) + \boldsymbol{X}_l, \\ \boldsymbol{Y}_l = \text{MLP}(\text{LayerNorm}(\boldsymbol{A}_l)) + \boldsymbol{A}_l. \end{cases} \tag{1}$$

MHSA allows the model to compute attention across different parts of the input sequence in parallel. Given the input $\boldsymbol{X}_l^{'}$, it operates as follows:

$$\begin{cases} [\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i] = \boldsymbol{X}_l^{'}\boldsymbol{W}^{qkv} + \boldsymbol{b}^{qkv}, \quad i = 1, 2, \ldots, h, \\ \text{Attn}_i(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \text{Softmax}\left(\frac{\boldsymbol{Q}_i \cdot \boldsymbol{K}_i^{\text{T}}}{\sqrt{D_h}}\right)\boldsymbol{V}_i, \\ \text{MHSA}(\boldsymbol{X}_l^{'}) = [\text{Attn}_1, \text{Attn}_2, \ldots, \text{Attn}_h]\boldsymbol{W}^{\text{O}} + \boldsymbol{b}^{\text{O}}, \end{cases} \tag{2}$$

where $\boldsymbol{W}^{qkv} \in \mathbb{R}^{D\times(3D_h)}$, $\boldsymbol{b}^{qkv} \in \mathbb{R}^{3D_h}$, $\boldsymbol{W}^{\text{O}} \in \mathbb{R}^{(hD_h)\times D}$, and $\boldsymbol{b}^{\text{O}} \in \mathbb{R}^{D}$. $\boldsymbol{Q}_i$, $\boldsymbol{K}_i$, and $\boldsymbol{V}_i$ are the query, key, and value matrices for the i^{th} attention head, respectively. "O" denotes the output. "qkv" denotes their concatenated linear projections. Attn_i denotes scaled dot-product attention for head i. h is the number of attention heads and D_h is the feature size of each head.

As demonstrated above, large-scale matrix multiplications of activations $\boldsymbol{A}$ (can also be $\boldsymbol{X}$ or $\boldsymbol{Y}$) and weights $\boldsymbol{W}$ are the primary contributors to computational overhead.

3.2 Quantizers

1. UQ. UQ is one of the most widely used quantization methods for activations and weights, defined as

$$\begin{cases} x_{\text{q}} = \text{UQ}(x, b) = \text{clamp}(\lfloor \frac{x}{s} \rceil + z, 0, 2^b - 1), \\ x_{\text{dq}} = \text{D-UQ}(x_{\text{q}}) = s(x_{\text{q}} - z) \approx x, \end{cases} \tag{3}$$

where x_{q} and x_{dq} represent the quantized value and the dequantized value, respectively, $\lfloor \cdot \rceil$ is the rounding operation, b is the quantization bit-width, and $\text{clamp}(\cdot)$ constrains the output between 0 and $2^b - 1$. Importantly, s is the quantization scale and z is the offset coefficient, both of which are determined by the lower and upper bounds of x as follows:

$$\begin{cases} s = \frac{\max(x) - \min(x)}{2^b - 1}, \\ z = \left\lfloor -\frac{\min(x)}{s} \right\rceil. \end{cases} \tag{4}$$

2. SULQ. SULQ is a novel quantizer (Zhong et al., 2026) for post-Softmax activations, which could improve the quantization efficiency. It can be represented as

$$\begin{cases} x_{\text{q}} = \text{SULQ}(x, b) = \text{UQ}(-\log_2(x + \eta), b), \\ x_{\text{dq}} = \text{D-SULQ}(x_{\text{q}}) = 2^{\lfloor -\text{D-UQ}(x_{\text{q}}) \rceil} - \eta \approx x, \end{cases} \tag{5}$$

where η is the shift parameter that can be adjusted based on the distribution of each layer. However, SULQ applies UQ without addressing outliers introduced by the log2 transformation, which may potentially lead to significant quantization errors.

3.3 Block reconstruction

Reconstruction techniques focus on bridging the performance gap between quantized model and the full-precision model by reducing the output deviation. The reconstruction objective can be defined as follows:

$$\min \left\| \boldsymbol{B}^l(\boldsymbol{W}, \boldsymbol{A}) - \boldsymbol{B}_{\text{q}}^l(\boldsymbol{W}_{\text{q}}, \boldsymbol{A}_{\text{q}}) \right\|^2, \tag{6}$$

where $\boldsymbol{B}^l(\cdot)$ and $\boldsymbol{B}_{\text{q}}^l(\cdot)$ denote the outputs of the l^{th} full-precision and quantized Transformer blocks, respectively. $\boldsymbol{W}_{\text{q}}$ and $\boldsymbol{A}_{\text{q}}$ correspond to the quantized weights and activations, respectively.

4 Methods

This section presents our proposed TP-ViT method for PTQ of ViTs, which introduces two key innovations to address current limitations in low-bit quantization. First, we develop TULQ, a novel approach that dynamically clips outliers during post-Softmax activation quantization to minimize the quantization errors. Second, we propose BDOS, an iterative quantization framework that progressively reduces bit-widths through intermediate precision stages while performing block-wise reconstruction. This strategy significantly reduces quantization errors and enhances quantized model performance, making TP-ViT highly effective for low-bit quantization scenarios. The process of our method is shown in Fig. 3.

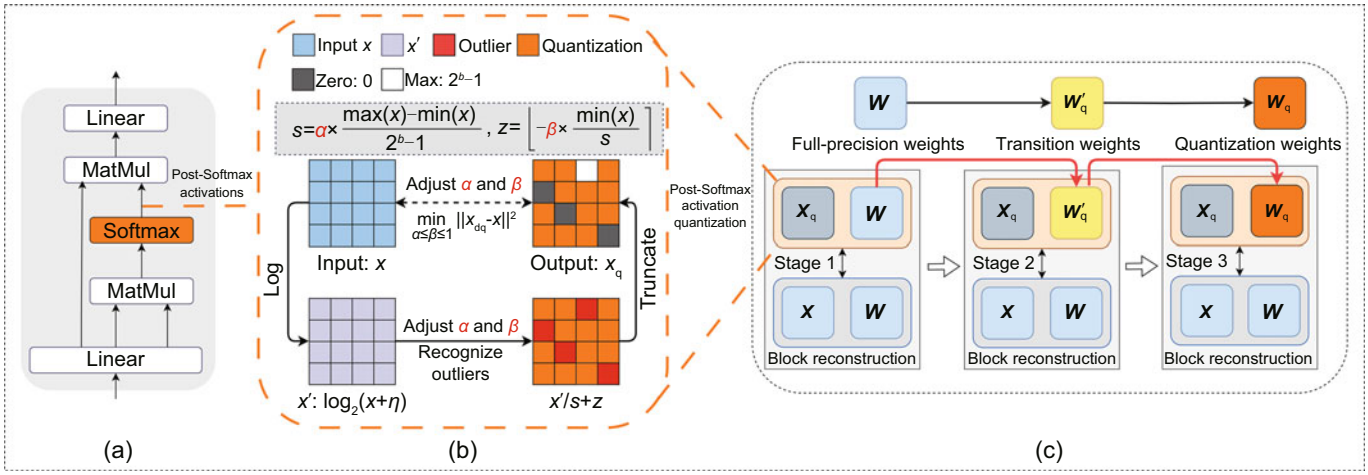

Fig. 3 Illustration of multi-head self-attention (a) and the proposed TP-ViT, which consists of TULQ (b) and BDOS (c). TULQ focuses on post-Softmax activation quantization in (a)

4.1 Truncated uniform-log2 quantizer

Although SULQ (Zhong et al., 2026) improves quantization efficiency by employing fine-grained quantization levels near zero, it fails to account for a critical limitation: it fails to account for the outliers induced by log2 transformation, which is a critical limitation for quantization performance. Although post-Softmax activations are bounded in $(0, 1]$, their logarithms span a wide negative range, creating numerical outliers that dominate the quantization error. These extreme values can substantially amplify quantization errors, ultimately degrading model performance. To overcome this issue, we try to mitigate the influence of outliers and stabilize the quantization process.

One commonly used technique for outlier suppression in activation quantization is percentile search truncated (PST) quantization (Li RD et al., 2019). This method confines the model's attention to the dominant activation distribution by truncating extreme values beyond empirically determined percentile thresholds, thereby improving quantization robustness. PST achieves this by adjusting the quantization scale s and the offset coefficient z as follows:

$$\begin{cases} s = \frac{\mathrm{pct}(x,\gamma)-\mathrm{pct}(x,1-\gamma)}{2^b-1}, \\ z = \left\lfloor -\frac{\mathrm{pct}(x,1-\gamma)}{s} \right\rceil, \end{cases} \tag{7}$$

where $\mathrm{pct}(\cdot)$ is a statistical function used to determine the relative position of a value in data. Specifically, the $\mathrm{pct}(x, \gamma)$ function returns the γ percentile value in the input x.

Although PST improves quantization performance, we identify its suboptimal efficacy for post-Softmax activations. This limitation stems from its reliance on a single coefficient to jointly control both the quantization scale s and offset coefficient z parameters, resulting in inflexible truncated boundaries that often lead to suboptimal value clipping. To empirically validate this limitation, we integrate PST into SULQ and evaluate its effectiveness on post-Softmax activation quantization (Fig. 4). Our experiments reveal that although SULQ+PST reduces the quantization error compared to standalone SULQ, the performance gain is marginal (+0.5 percentage points (PPs)). Strikingly, when we adjust the offset coefficient z after applying SULQ+PST, we observe a substantially larger improvement (+1.55 PPs). This significant discrepancy demonstrates the following: (1) The PST current truncated strategy is not fully optimized for post-Softmax distributions. (2) The control of s and z is critical for effective outlier handling. These findings strongly suggest that post-Softmax activations require a more flexible truncated mechanism.

To address the inherent limitations of PST in post-Softmax activation quantization, we propose TULQ. Unlike conventional approaches that independently constrain extreme values, which often lead to suboptimal coefficient selection, TULQ introduces a dual-parameter mechanism for more effective dynamic range control as follows:

1. The coefficient α adaptively determines the quantization scale s by optimizing the retention ratio of the activation distribution.

2. The coefficient β adjusts the offset coefficient z to selectively eliminate small-value regions while preserving critical information, thereby robustly mitigating outlier effects.

This coordinated parameterization enables TULQ to achieve superior quantization precision compared to rigid truncated methods, particularly for the heavy-tailed distribution characteristic of post-Softmax activations. The updated quantization formulae are as follows:

$$\begin{cases} s = \alpha \times \frac{\max(x)-\min(x)}{2^b-1}, \\ z = \left\lfloor -\beta \times \frac{\min(x)}{s} \right\rceil, \end{cases} \tag{8}$$

where α and β satisfy $\alpha \le \beta \le 1$. It could correctly address the outliers by adjusting α and β, as shown in Fig. 1. Notably, our proposed quantizer degenerates into SULQ when $\alpha = \beta = 1$, ensuring that it remains flexible across different activation distributions.

In summary, we propose TULQ to enhance the performance of SULQ by effectively mitigating the influence of outliers and reducing quantization errors. The TULQ quantization process is defined as follows:

$$\begin{cases} x_{\mathrm{q}} = \mathrm{clamp}\Big(\left\lfloor \frac{\log_2(x+\eta)}{s} \right\rceil + z, 0, 2^b - 1 \Big), \\ x_{\mathrm{dq}} = 2^{\lfloor -s(x_{\mathrm{q}}-z) \rceil} - \eta \approx x, \end{cases} \tag{9}$$

where s and z are obtained from Eq. (8).

As shown in Fig. 3b, in the quantization process for the post-Softmax activations of different layers, these two coefficients, α and β, can be searched to recognize the outliers and then truncated to minimize the quantization errors. The expression is as follows:

$$\min_{\alpha \le \beta \le 1} \left\| x_{\mathrm{dq}} - x \right\|^2. \tag{10}$$

As illustrated in Fig. 4, TULQ achieves the lowest quantization loss across all the blocks, demonstrating its effectiveness in preserving activation information. Furthermore, TULQ attains a top-1 accuracy of 56.61% on 3-bit DeiT-S with only quantized activations, significantly outperforming both SULQ (+6.18 PPs) and SULQ+PST (+5.68 PPs). These results

Fig. 4 The quantitative comparison of post-Softmax activation quantization errors in 3-bit DeiT-S (12-block architecture) across different quantizers, evaluated without reconstruction optimization techniques. "SULQ+PST" means that we add the percentile search truncated to SULQ, while "SULQ+PST with shift" means that we shift the offset coefficient z after applying SULQ with PST. Our proposed method named TULQ (red line) achieves the smallest loss compared to other methods and attains the best performance of 56.61% top-1 accuracy

highlight the superiority of our proposed TULQ in reducing quantization errors and improving model performance, making it a promising solution for low-bit quantization in ViTs. Notably, this gain comes with only a modest increase in calibration time (approximately 17 s more than SULQ), making it a practical and effective solution for low-bit quantization in ViTs.

4.2 Bit-decline optimization strategy

Multi-stage block reconstruction techniques have demonstrated remarkable success in PTQ for ViTs, as exemplified by BRECQ (Li YH et al., 2021) and I&S-ViT (Zhong et al., 2026). Although these methods achieve SOTA performance in 4-bit quantization scenarios, they exhibit significant performance degradation when applied to more aggressive 3-bit quantization regimes. Current approaches typically employ a sequential quantization strategy: first, quantize either weights or activations while keeping the other in full precision, followed by quantization of the remaining component. At each stage of this process, block reconstruction is performed at the end to update the weights. Our empirical analysis reveals that although the initial quantization stage (of either weights or activations alone) maintains satisfactory accuracy, the subsequent joint quantization of both weights and activations introduces substantial quantization errors, leading to severe performance deterioration.

To overcome the limitation, we propose a progressive quantization framework that employs a multi-stage BDOS for block reconstruction. The intermediate 8-bit model, reconstructed under the same full-precision supervision, provides a high-quality initialization for the final 3-bit stage. By reducing the bit-width (e.g., FP32→8-bit→3-bit), BDOS greatly reduces the optimization difficulty of ultra-low-bit quantization and decreases the error.

The conventional procedure of block reconstruction, as exemplified by I&S-ViT (Zhong et al., 2026), follows a sequential approach. Initially, the activations are quantized to the target bit-width, and block reconstruction is performed to optimize the weight to adapt to the quantization of activations. Subsequently, the quantized activations are held constant, and the weights are quantized. Block reconstruction is then applied again to optimize the weight quantization, thereby restoring the model's performance. This procedure offers two benefits. First, specialized activations (e.g., post-Softmax and post-LayerNorm) require customized quantization schemes, making fixed quantized activations preferable to avoid re-quantization overhead. Second, the strategy of quantizing weights while keeping activations fixed demonstrates lower optimization complexity compared to simultaneous weight and activation quantization.

Therefore, following this criterion, we only focus on progressively quantizing the weights during block reconstruction. The multi-stage BDOS can be divided into three stages, as shown in Fig. 3c.

1. Stage 1: activation quantization and full-precision weight learning. We first quantize the activation $\boldsymbol{A}$ to the target bit-width $\boldsymbol{A}_{\mathrm{q}}$ while keeping the weights learning $\boldsymbol{W}$ in full precision. The quantized activation will be fixed in the following process. To mitigate the quantization errors introduced by activation quantization, we apply block reconstruction to update the weights $\boldsymbol{W}^{\mathrm{f}}$ to $\boldsymbol{W}^{\mathrm{f}*}$. For the l^{th} Transformer block $\boldsymbol{B}^l$, the learning procedure of block reconstruction at this stage is formulated as follows:

$$\begin{cases} \boldsymbol{A}_{\mathrm{q}} = \mathrm{UQ}(\boldsymbol{A}), \\ \boldsymbol{W}^{\mathrm{f}*} = \underset{\boldsymbol{W}^{\mathrm{f}}}{\arg\min} \left\| \boldsymbol{B}^l(\boldsymbol{W}, \boldsymbol{A}) - \boldsymbol{B}^l_{\mathrm{q}}(\boldsymbol{W}^{\mathrm{f}}, \boldsymbol{A}_{\mathrm{q}}) \right\|^2, \end{cases} \tag{11}$$

where UQ is the uniform quantization (Eq. (3)) used for activations. Notably, the post-Softmax activations adopt our proposed TULQ (Eq. (9)).

As discussed in Zhong et al. (2026), this process could provide a smoother and more stable loss landscape, which alleviates the training challenges associated with reconstruction.

2. Stage 2: transitional bit-width weight quantization. To quantize the weights to the target low-bit representation, we first perform a transitional quantization on the weights $\boldsymbol{W}^{\mathrm{f}*}$ to obtain an intermediate bit-width weight $\boldsymbol{W}^{\mathrm{t}}_{\mathrm{q}}$. For example, we can use 8-bit as the transitional bit-width for lower-bit quantization, because 8-bit quantization is nearly lossless. The optimization procedure for this stage is

$$\begin{cases} \boldsymbol{W}^{\mathrm{t}}_{\mathrm{q}} = \mathrm{UQ}(\boldsymbol{W}^{\mathrm{f}*}), \\ \boldsymbol{W}^{\mathrm{t}*}_{\mathrm{q}} = \underset{\boldsymbol{W}^{\mathrm{t}}_{\mathrm{q}}}{\arg\min} \left\| \boldsymbol{B}^l(\boldsymbol{W}, \boldsymbol{A}) - \boldsymbol{B}^l_{\mathrm{q}}(\boldsymbol{W}^{\mathrm{t}}_{\mathrm{q}}, \boldsymbol{A}_{\mathrm{q}}) \right\|^2. \end{cases} \tag{12}$$

3. Stage 3: target weight quantization. At the final stage, the weights $\boldsymbol{W}^{\mathrm{t}*}_{\mathrm{q}}$ obtained from the previous stages are quantized to the target bit-width $\boldsymbol{W}_{\mathrm{q}}$. Block reconstruction is then applied to restore the performance of the quantized model, with the reconstruction procedure given as follows:

$$\begin{cases} \boldsymbol{W}_{\mathrm{q}} = \mathrm{UQ}(\boldsymbol{W}^{\mathrm{t}*}_{\mathrm{q}}), \\ \boldsymbol{W}^{*}_{\mathrm{q}} = \underset{\boldsymbol{W}_{\mathrm{q}}}{\arg\min} \left\| \boldsymbol{B}^l(\boldsymbol{W}, \boldsymbol{A}) - \boldsymbol{B}^l_{\mathrm{q}}(\boldsymbol{W}_{\mathrm{q}}, \boldsymbol{A}_{\mathrm{q}}) \right\|^2. \end{cases} \tag{13}$$

Note that throughout the stages, the reconstruction target $\boldsymbol{B}^l(\boldsymbol{W}, \boldsymbol{A})$ remains the original full-precision block output; the intermediate bit-width models serve only as optimized initializations for the subsequent lower-bit stage, not as new supervision signals.

To improve the block reconstruction performance, the quantization process can be progressively refined through iterative bit-width reduction with multiple rounds. Specifically, when quantizing to a target 3-bit model, we implement a multi-stage approach: transition weights are first quantized to 8-bit, then progressively reduced to 4-bit, before achieving the final 3-bit precision. Block reconstruction is performed after each quantization step to minimize the reconstruction loss. Although this multi-round BDOS consistently improves final model performance compared to direct quantization, the performance gains are relatively marginal compared to a single round. Given the significant increase in computational time required for each additional round, the single-round BDOS is sufficient.

Compared to other multi-stage block reconstruction approaches (Li YH et al., 2021; Zhong et al., 2026), our proposed BDOS could significantly reduce the quantization errors during block reconstruction. This improvement cannot be achieved simply by increasing the number of iterations in block reconstruction.

5 Experiments

5.1 Experimental settings

1. Models and tasks. To validate the effectiveness of our proposed TP-ViT framework and its core components (TULQ and BDOS), we conduct comprehensive experiments on two benchmark datasets: ImageNet (Deng et al., 2009) for image classification and COCO (Lin TY et al., 2014) for object detection and instance segmentation. Our evaluation encompasses multiple vision Transformer architectures, including ViT (Dosovitskiy et al., 2021), DeiT (Touvron et al., 2021), and Swin Transformer (Liu Z et al., 2021) with small and base scales. For the ImageNet classification task, we evaluate the models under standard protocols. To demonstrate broader applicability, we further assess performance on COCO using cascade mask R-CNN (Cai and Vasconcelos, 2018) with Swin Transformer backbone for object detection and instance segmentation tasks.

2. Implementation details. The pre-trained full-precision models used in our experiments are sourced from the Timm library. For the quantization process, we employ a channel-wise uniform quantization for weights, while layer-wise uniform quantization is used for activations, except post-Softmax activations that are quantized using our proposed TULQ. Consistent with preceding studies (Zhong et al., 2026), we randomly sample 1024 images from both the ImageNet and COCO datasets for optimization. The optimization process uses the Adam optimizer (Kingma and Ba, 2017) with an initial learning rate of 4×10^{-5} and a weight decay of 0.2. For the image classification tasks on ImageNet, we set the batch size to 64 and conduct 1000 iterations for all cases, except the 6-bit quantization scenario using 200 iterations. For the object detection and instance segmentation tasks on COCO, we optimize only the backbone, set the batch size to 1, and perform 1000 iterations. The TULQ parameters η, α, and β are determined via grid search during the calibration phase. The parameter η is set following the same setting as in SULQ. Meanwhile, both α and β are searched in the range [0.7, 1.0] with a step size of 0.01 under the constraint $\alpha \leq \beta \leq 1$, ensuring fine-grained adaptation to the activation distribution of each layer. Additionally, we do not employ our proposed BDOS for 6-bit quantization. For 3-bit and 4-bit quantization, BDOS is set to one round, with the transition weights quantized to 8-bit precision. Experiments are conducted with PyTorch on a single NVIDIA 3090 GPU.

5.2 Image classification on the ImageNet dataset

This subsection evaluates the performance of different quantization methods on the ImageNet dataset for image classification tasks. We select various vision Transformers and their variants, including ViT-S, ViT-B, DeiT-S, DeiT-B, Swin-S, and Swin-B, to comprehensively assess the effectiveness of these methods under 3-bit, 4-bit, and 6-bit quantization settings for both weights and activations. Here, S denotes small and B denotes the base version. Table 1 presents the top-1 accuracy of each method under different quantization settings. We can find the following:

1. In the 3-bit quantization setting, our TP-ViT method achieves superior PTQ performance across different backbones. Specifically, it improves the top-1 accuracy of 3-bit ViT-S (+6.18 PPs) and 3-bit ViT-B (+3.25 PPs) over the SOTA method I&S-ViT, respectively. Additionally, our method achieves the highest accuracy on DeiT-S and DeiT-B, with 58.52% and 73.56%, respectively. These results demonstrate the adaptability of our quantization strategy to different model architectures and its effectiveness in preserving classification accuracy under extremely low-bit quantization scenarios.

2. In the 4-bit quantization setting, our method consistently outperforms SOTA methods, including I&S-ViT (Zhong et al., 2026) and outlier-aware (Ma et al., 2024). For example, on ViT-S, our method achieves a top-1 accuracy of 75.95%, surpassing that of I&S-ViT (74.87%) and recent mixed-precision method AMP-ViT (55.88%). It also exceeds the 4-bit quantization result of ORQ-ViT (68.49%) by a large margin. Our method also narrows the gap between quantized and full-precision accuracy to 1.53 PPs on ViT-B and 1.32 PPs on DeiT-B. These results highlight the robustness of our quantization method across different models and settings. The combination of TULQ and BDOS effectively mitigates the impact of quantization, ensuring high classification accuracy.

3. In the 6-bit quantization setting, the performance gap between our method and SOTA methods narrows, including the mixed-precision method. For instance, on 6-bit quantized DeiT-S, our method achieves a top-1 accuracy of 79.41%, slightly lower than that of outlier-aware (Ma et al., 2024) (79.50%). This suggests that as the quantization bit-width increases, the gap between the quantized and full-precision models diminishes, making it more challenging to achieve optimal results across all models. However, our method still achieves commendable results on certain models. Notably, the 6-bit quantized ViT-B attains a top-1 accuracy of 84.79%, which is even 0.25 PPs higher than that of the full-precision model. This highlights the potential of our method in achieving near or even super full-precision performance in higher bit-width quantization settings.

5.3 Object detection and instance segmentation on the COCO dataset

To demonstrate the applicability of our proposed quantization method in real-world computer vision tasks, we evaluate its performance on the COCO dataset using the 4-bit quantized cascade mask R-CNN model with Swin-T and Swin-S backbones. The results are summarized in Table 2. Our method achieves SOTA performance in both object detection and instance segmentation tasks, highlighting its effectiveness in maintaining high accuracy under low-bit quantization settings. From Table 2, TP-ViT achieves the best overall performance across all settings. Compared to I&S-ViT, our method improves AP^{box} (+0.1 PPs) and AP^{mask} (+0.1 PPs) on Swin-T. On Swin-S, our method achieves an improvement over I&S-ViT, reaching the highest 50.4% AP^{box} and 43.7% AP^{mask}. These results indicate the strong generalization ability of our method across different vision tasks and its potential for practical applications.

Table 1 Quantization results for image classification on the ImageNet dataset

Method	Bit-width (W/A)	Top-1 accuracy (%)					
		ViT-S	ViT-B	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-precision	32/32	81.39	84.54	79.85	81.80	83.23	85.27
PTQ4ViT (Yuan et al., 2022)	3/3	0.01	0.01	0.01	0.27	0.35	0.29
BRECQ (Li YH et al., 2021)	3/3	0.42	0.59	14.63	46.29	11.67	1.70
QDrop (Wei et al., 2023)	3/3	4.44	8.00	22.67	24.37	60.89	54.76
PD-Quant (Liu JW et al., 2023)	3/3	1.77	13.09	29.33	0.94	69.67	64.32
RepQ-ViT (Li ZK et al., 2023)	3/3	0.43	0.14	4.37	4.84	8.84	1.34
I&S-ViT (Zhong et al., 2026)	3/3	45.16	63.77	55.78	73.30	74.20	69.30
AIQViT (Jiang RQ et al., 2025)	3/3	41.32	43.68	55.36	66.15	71.42	63.01
TP-ViT (ours)	3/3	**51.34**	**67.02**	**58.52**	**73.56**	**74.65**	**70.44**
PTQ4ViT (Yuan et al., 2022)	4/4	42.57	30.69	34.08	64.39	76.09	74.02
APQ-ViT (Ding et al., 2022)	4/4	47.95	41.41	43.55	67.48	77.15	76.48
BRECQ (Li YH et al., 2021)	4/4	12.36	9.68	63.73	72.31	72.74	58.24
PD-Quant (Liu JW et al., 2023)	4/4	1.51	32.45	71.21	73.76	79.87	81.12
RepQ-ViT (Li ZK et al., 2023)	4/4	65.05	68.48	69.03	75.61	79.45	78.32
I&S-ViT (Zhong et al., 2026)	4/4	74.87	80.07	75.81	79.97	81.17	82.60
ADFQ-ViT (Jiang YF et al., 2025)	4/4	72.14	78.71	75.06	78.75	80.63	82.33
Outlier-aware (Ma et al., 2024)	4/4	72.88	76.59	76.00	78.83	81.02	82.46
AIQViT (Jiang RQ et al., 2025)	4/4	70.63	74.15	72.75	79.19	80.93	81.22
AMP-ViT (Tai and Wu, 2025)	4*/4*	55.88	61.84	68.43	76.14	77.20	76.51
ORQ-ViT (He et al., 2025)	4/4	68.49	73.22	70.13	76.66	80.06	81.60
TP-ViT (Ours)	4/4	**75.95**	**83.01**	**76.08**	**80.48**	**81.52**	**82.86**
PTQ4ViT (Yuan et al., 2022)	6/6	78.63	81.65	76.28	80.25	82.38	84.01
APQ-ViT (Ding et al., 2022)	6/6	79.10	82.21	77.76	80.42	82.67	84.18
BRECQ (Li YH et al., 2021)	6/6	54.51	68.33	78.46	80.85	82.02	83.94
PD-Quant (Liu JW et al., 2023)	6/6	70.84	75.82	78.40	80.52	82.51	84.32
TSPTQ-ViT (Tai et al., 2023)	6/6	79.34	82.01	77.27	80.25	82.41	84.12
I&S-ViT (Zhong et al., 2026)	6/6	80.43	83.82	79.15	81.68	82.89	84.94
ADFQ-ViT (Jiang YF et al., 2025)	6/6	80.54	83.92	79.04	81.53	82.81	84.82
Outlier-aware (Ma et al., 2024)	6/6	80.60	83.81	**79.50**	**81.72**	82.76	84.91
AIQViT (Jiang RQ et al., 2025)	6/6	80.21	83.68	78.98	81.40	82.81	84.39
AMP-ViT (Tai and Wu, 2025)	6*/6*	79.98	82.70	78.70	81.25	82.62	84.50
ORQ-ViT (He et al., 2025)	6/6	80.48	83.77	79.14	81.35	82.81	84.98
TP-ViT (ours)	6/6	**80.65**	**84.79**	79.41	**81.72**	**82.93**	**85.05**

"Bit-width (W/A)" denotes the quantization bit-width of weights and activations as W bits and A bits, respectively. * Mixed-precision quantization. The best results are in bold

Table 2 Performance of the 4/4 quantized cascade mask R-CNN model on the COCO dataset

Method	AP (%)			
	Swin-T		Swin-S	
	Box	Mask	Box	Mask
Full-precision	50.4	43.7	51.9	45.0
PTQ4ViT (Yuan et al., 2022)	14.7	13.5	0.5	0.5
APQ-ViT (Ding et al., 2022)	27.2	24.4	47.7	41.1
BRECQ (Li YH et al., 2021)	41.2	37.0	44.5	39.2
QDrop (Wei et al., 2023)	23.9	21.2	24.1	21.4
PD-Quant (Liu JW et al., 2023)	35.5	31.0	41.6	36.3
RepQ-ViT (Li ZK et al., 2023)	47.0	41.4	49.3	43.1
I&S-ViT (Zhong et al., 2026)	48.2	42.0	50.3	43.6
AIQViT (Jiang RQ et al., 2025)	47.1	41.4	49.8	43.4
TP-ViT (ours)	**48.3**	**42.1**	**50.4**	**43.7**

AP denotes the average precision. The box is AP^{box} referring to the AP of bounding boxes for object detection tasks, while the mask is AP^{mask} referring to the AP of segmentation masks for instance segmentation tasks. The best results are in bold

5.4 Ablation studies

To further investigate the contributions of the proposed TULQ and BDOS to the overall performance, we conduct ablation studies on 3-bit quantized ViT-S for image classification on the ImageNet dataset. The I&S-ViT (Zhong et al., 2026) method is adopted as the baseline. The results are summarized in Table 3. The baseline I&S-ViT model without TULQ or BDOS components achieves 45.16% top-1 accuracy. Implementing TULQ alone increases accuracy to 46.53%, confirming its capability to mitigate outlier effects and reduce quantization errors. The independent application of BDOS yields further improvement to 50.42%, demonstrating the efficacy of our bit-decline strategy in block reconstruction optimization. The combined implementation of both TULQ and BDOS achieves optimal performance with 51.34% top-1 accuracy, surpassing all other configurations. These experimental results conclusively validate the effectiveness of TULQ and BDOS in enhancing quantized model performance.

5.4.1 Effect of TULQ

We implement a UQ for 3-bit activation quantization in the full-precision ViT-S model while employing specialized quantizers exclusively for post-Softmax activations. Four quantization methods are evaluated: UQ, LQ, SULQ, and our proposed TULQ. As shown in Table 4, UQ yields the poorest performance. LQ demonstrates superior results compared to UQ by better accommodating the long-tail distribution of post-Softmax activations. SULQ builds upon LQ to further enhance quantization efficiency, achieving suboptimal

Table 3 Ablation studies on the ImageNet classification of TULQ and BDOS of 3-bit quantized ViT-S

Model	TULQ	BDOS	Top-1 acc. (%)
ViT-S	Full-precision		81.39
	×	×	45.16
	✓	×	46.53
	×	✓	50.42
	✓	✓	**51.34**

Acc. refers to accuracy. The best result is in bold

Table 4 Ablation studies on the ImageNet classification of different quantizers applied to post-Softmax activations of ViT-S

Model	W/A	Method	Top-1 acc. (%)
ViT-S	32/32	Full-precision	81.39
	32/3	UQ	41.39
	32/3	LQ	50.19
	32/3	SULQ	52.29
	32/3	TULQ (ours)	**54.09**

Acc. refers to accuracy. W/A denotes the quantization bit-width of weights and activations as W bits and A bits, respectively. The best result is in bold

performance. Our TULQ method effectively mitigates outlier effects, ultimately delivering the best performance with 54.09% top-1 accuracy. As shown in Fig. 5, TULQ achieves the smallest reconstruction loss among all quantizers, highlighting its effectiveness.

5.4.2 Effect of BDOS

To explore the potential of BDOS, we first examine the impact of different quantization bits for the transition weights in a single round of BDOS. In these experiments, we employ a UQ for all weights and activations, except post-Softmax activations, which are quantized using our proposed TULQ quantizer. As shown in Table 5, the model using BDOS with 8-bit transition weights achieves the best top-1 accuracy of 51.34%, which is significantly higher than the accuracy obtained without BDOS (46.53%). Meanwhile, we analyze the reconstruction loss from the final stage with varying transition weights, as shown in Fig. 6. The results indicate that using 8-bit transition weights yields the smallest reconstruction loss. This is because 8-bit quantization is nearly lossless, allowing 8-bit transition weights to provide a better initialization for low-bit quantization models by minimizing the quantization error.

In addition, we examine the impact of varying the number of BDOS rounds on the quantization results. Table 6 shows the top-1 accuracy of the ViT-S model under a final quantization bit of 3-bit, with different numbers of rounds in the BDOS process. Specifically, we observe that using three rounds of BDOS achieves a top-1 accuracy of 51.49%, which is 0.15 PPs higher than that of a single round. The performance gain from increasing the number of rounds is relatively limited, suggesting that the improvement is marginal compared to the substantial gains achieved with just one round.

Moreover, Fig. 7 illustrates the block reconstruction loss as the number of BDOS rounds increases. The losses between each block generally decrease with more rounds, indicating a

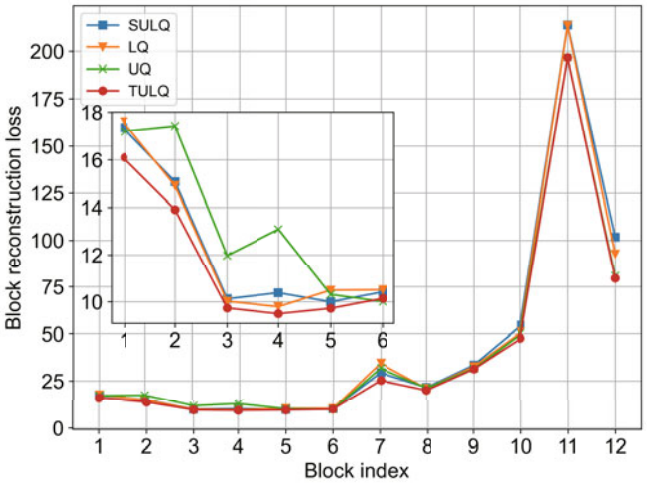

Fig. 5 Comparison of the reconstruction loss in the 12-block 32/3 quantized ViT-S using different quantizers

Table 5 The effectiveness of the transition weights with ViT-S as the quantized model

Model	N	W	Top-1 acc. (%)
ViT-S		Full-precision	81.39
	0	$32 \rightarrow 3$	46.53
	1	$32 \rightarrow 8 \rightarrow 3$	**51.34**
	1	$32 \rightarrow 6 \rightarrow 3$	50.95
	1	$32 \rightarrow 4 \rightarrow 3$	51.12

N represents the number of rounds. Acc. refers to accuracy. W denotes the quantization bit-width of weights as W bits. The best result is in bold. When the target quantization bit is 3-bit, the transition weights in one round of BDOS are quantized to 4-, 6-, and 8-bit, respectively

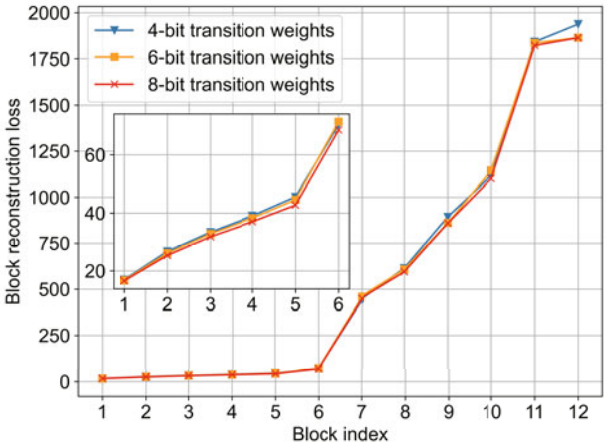

Fig. 6 Comparison of the reconstruction loss in the 12-block 3/3 quantized ViT-S using different quantization bits for the transition weights in one round of BDOS. Adopting 8-bit transition BDOS (red line) achieves the smallest loss

positive correlation between the number of BDOS rounds and the reduction in reconstruction loss. This suggests that multiple rounds of BDOS allow for a more nuanced quantization process, potentially reducing quantization errors and improving the model's ability to generalize. However, increasing the number of BDOS rounds also significantly increases the time required for quantization. For example, each additional BDOS round for ViT-S increases the quantization time by approximately 12 min, but the performance increment is subtle, which

Table 6 Quantitative results of BDOS testing at different rounds when the target quantization bit is 3-bit and the quantized model is ViT-S

Model	N	W	Top-1 acc. (%)
ViT-S		Full-precision	81.39
	0	$32 \to 3$	46.53
	1	$32 \to 8 \to 3$	51.34
	2	$32 \to 8 \to 4 \to 3$	51.42
	3	$32 \to 8 \to 6 \to 4 \to 3$	**51.49**

N represents the number of rounds. Acc. refers to accuracy. W denotes the quantization bit-width of weights as W bits. The best result is in bold

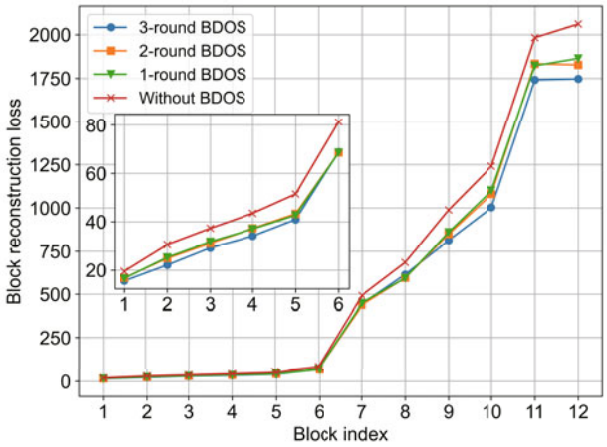

Fig. 7 Comparison of the reconstruction loss in the 12-block 3/3 quantized ViT-S using different rounds of BDOS. As the number of BDOS rounds increases, the reconstruction loss gradually decreases. A trade off between time cost and accuracy should be considered

indicates that one round of BDOS is enough.

5.5 Time efficiency

Fig. 8 illustrates the quantization time and accuracy of various PTQ methods on the 3/3 DeiT-S model. Our proposed method demonstrates relatively good time efficiency, requiring only 42 min. Additionally, our method achieves outstanding performance in terms of accuracy, reaching a top-1 accuracy of 58.52%, which is the highest among all PTQ methods. In comparison, other methods without block reconstruction optimization, such as PTQ4ViT (Yuan et al., 2022) and RepQ-ViT (Li ZK et al., 2023), while having their own advantages in quantization time, fail to match our method's accuracy. These results indicate that our method maintains competitive accuracy levels while operating within acceptable quantization time, achieving an optimal balance between efficiency and performance.

5.6 Comparison of visualization results

To visually demonstrate the effectiveness of our proposed method, we select six images from the ImageNet validation set and apply Grad-CAM (Selvaraju et al., 2020) to compare the visualization results with those of I&S-ViT (Zhong et al., 2026), both based on 3-bit quantized ViT-S, as shown in Fig. 9. The visualizations indicate that our method more effectively highlights discriminative features, as seen in images 2 and 3. Additionally, images 5 and 6 show that our 3-bit quantized ViT-S

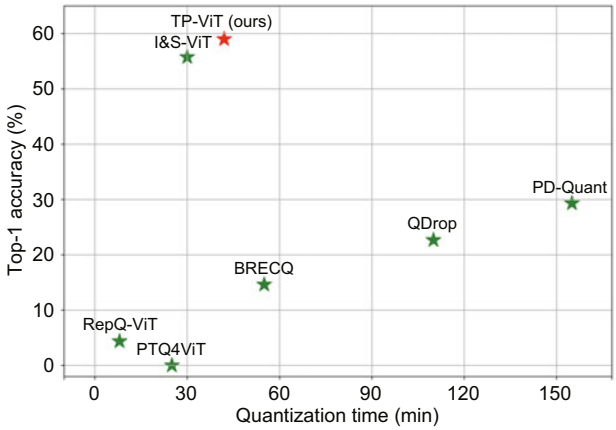

Fig. 8 The quantization time and accuracy using different PTQ methods to perform 3-bit quantization on DeiT-S

maintains a stronger focus on informative objects without being distracted by class-independent elements. By employing TULQ and BDOS, our method achieves a lower quantization error and superior classification performance, resulting in a 6.18 PPs improvement in top-1 accuracy for 3-bit quantization compared to I&S-ViT.

6 Conclusions

In this study, we propose a novel PTQ framework, TP-ViT, tailored for ViTs to address significant performance degradation during low-bit quantization. Our approach introduces TULQ and BDOS. TULQ mitigates the impact of outliers in activation quantization, reducing quantization errors, while BDOS enhances model performance through progressive quantization. Experiments on ImageNet and COCO datasets demonstrate the efficacy of TP-ViT, achieving substantial improvements in top-1 accuracy and outperforming SOTA methods in object detection and instance segmentation tasks. Ablation studies validate the contributions of both TULQ and BDOS, showing that our method achieves high accuracy while maintaining time efficiency. This work provides a robust PTQ framework for efficient deployment of ViTs on resource-constrained hardware, with potential for further application to other Transformer-based architectures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62301092 and 62301093).

Author contributions

Xichuan ZHOU supervised the project, conceived the research idea, and developed the methodology. Sihuan ZHAO contributed to the methodology, validated the experiments, implemented the software, and drafted the paper. Rui DING performed validation, developed the software, and co-drafted the paper. Jiayu SHI assisted in experimental validation and software implementation. Jing NIE contributed to software development, participated in conceptual discussions, and revised the paper. Lihui CHEN supervised the work, contributed to the conceptualization, and revised the paper. Haijun LIU supervised the project, secured funding, and revised and finalized the paper.

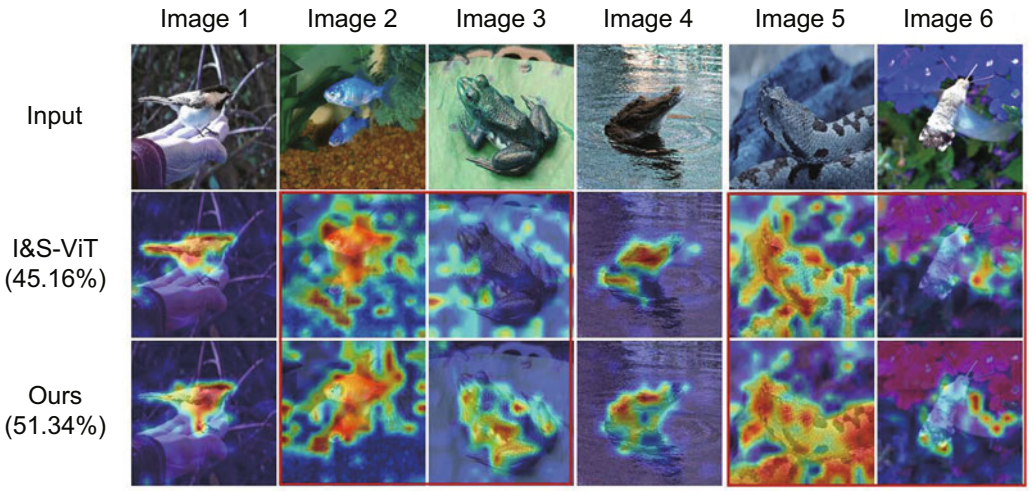

Fig. 9 Qualitative visualization image classification results of 3-bit quantized ViT-S using the I&S-ViT method and our proposed TP-ViT

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this paper, the authors used a generative AI tool (Qwen) solely to assist with language editing and improve clarity and fluency. The authors retain full responsibility for the scientific content, accuracy, and integrity of the final paper.

References

Cai ZW, Vasconcelos N, 2018. Cascade R-CNN: delving into high quality object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6154-6162. https://doi.org/10.1109/cvpr.2018.00644

Chen J, Zhang HY, Gong MM, et al., 2024. Collaborative compensative Transformer network for salient object detection. *Patt Recogn*, 154:110600. https://doi.org/10.1016/j.patcog.2024.110600

Deng J, Dong W, Socher R, et al., 2009. ImageNet: a large-scale hierarchical image database. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.248-255. https://doi.org/10.1109/cvpr.2009.5206848

Ding YF, Qin HT, Yan QH, et al., 2022. Towards accurate post-training quantization for vision Transformer. Proc 30th ACM Int Conf on Multimedia, p.5380-5388. https://doi.org/10.1145/3503161.3547826

Dosovitskiy A, Beyer L, Kolesnikov A, et al., 2021. An image is worth 16×16 words: Transformers for image recognition at scale. https://doi.org/10.48550/arXiv.2010.11929

Gao LN, Liu B, Fu P, et al., 2024. TSVT: token sparsification vision Transformer for robust RGB-D salient object detection. *Patt Recogn*, 148:110190. https://doi.org/10.1016/j.patcog.2023.110190

He XY, Lu Y, Liu H, et al., 2025. ORQ-ViT: outlier resilient post training quantization for vision Transformers via outlier decomposition. *J Syst Architect*, 168:103530. https://doi.org/10.1016/j.sysarc.2025.103530

Jiang RQ, Zhang Y, Wang LG, et al., 2025. AIQViT: architecture-informed post-training quantization for vision Transformers. https://doi.org/10.48550/arXiv.2502.04628

Jiang YF, Sun N, Xie XS, et al., 2025. ADFQ-ViT: activation-distribution-friendly post-training quantization for vision Transformers. *Neur Netw*, 186:107289. https://doi.org/10.1016/j.neunet.2025.107289

Kim HJ, Shin JW, Del Barrio AA, 2022. CTMQ: cyclic training of convolutional neural networks with multiple quantization steps. https://doi.org/10.48550/arXiv.2206.12794

Kingma DP, Ba J, 2017. Adam: a method for stochastic optimization. https://doi.org/10.48550/arXiv.1412.6980

Li MH, Halstead M, McCool C, 2024. Knowledge distillation for efficient instance semantic segmentation with Transformers. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.5432-5439. https://doi.org/10.1109/cvprw63382.2024.00552

Li RD, Wang Y, Liang F, et al., 2019. Fully quantized network for object detection. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2805-2814. https://doi.org/10.1109/cvpr.2019.00292

Li YH, Gong RH, Tan X, et al., 2021. BRECQ: pushing the limit of post-training quantization by block reconstruction. https://doi.org/10.48550/arXiv.2102.05426

Li ZK, Xiao JR, Yang LW, et al., 2023. RepQ-ViT: scale reparameterization for post-training quantization of vision Transformers. Proc IEEE/CVF Int Conf on Computer Vision, p.17181-17190. https://doi.org/10.1109/iccv51070.2023.01580

Lin TY, Maire M, Belongie S, et al., 2014. Microsoft COCO: common objects in context. Proc 13th European Conf on Computer Vision, p.740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Lin Y, Zhang TY, Sun PQ, et al., 2023. FQ-ViT: post-training quantization for fully quantized vision Transformer. https://doi.org/10.48550/arXiv.2111.13824

Liu JW, Niu L, Yuan ZH, et al., 2023. PD-Quant: post-training quantization based on prediction difference metric. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.24427-24437. https://doi.org/10.1109/cvpr52729.2023.02340

Liu SY, Liu ZC, Cheng KT, 2023. Oscillation-free quantization for low-bit vision Transformers. Proc 40th Int Conf on Machine Learning, p.21813-21824.

Liu Z, Lin YT, Cao Y, et al., 2021. Swin Transformer: hierarchical vision Transformer using shifted windows. Proc IEEE/CVF Int Conf on Computer Vision, p.9992-10002. https://doi.org/10.1109/iccv48922.2021.00986

Ma YX, Li HX, Zheng XW, et al., 2024. Outlier-aware slicing for post-training quantization in vision Transformer. Proc 41st Int Conf on Machine Learning, p.33811-33825.

Mahmood T, Wahid A, Hong JS, et al., 2024. A novel convolution Transformer-based network for histopathology-image classification using adaptive convolution and dynamic attention. *Eng Appl Artif Intell*, 135:108824. https://doi.org/10.1016/j.engappai.2024.108824

Mehta S, Rastegari M, 2022. MobileViT: light-weight, general-purpose, and mobile-friendly vision Transformer. https://doi.org/10.48550/arXiv.2110.02178

Nagel M, Fournarakis M, Bondarenko Y, et al., 2022. Overcoming oscillations in quantization-aware training. Proc 39th Int Conf on Machine Learning, p.16318-16330.

Selvaraju RR, Cogswell M, Das A, et al., 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*, 128:336-359. https://doi.org/10.1007/s11263-019-01228-7

Tai YS, Wu AYA, 2025. AMP-ViT: optimizing vision Transformer efficiency with adaptive mixed-precision post-training quantization. IEEE/CVF Winter Conf on Applications of Computer Vision, p.6828-6837. https://doi.org/10.1109/wacv61041.2025.00664

Tai YS, Lin MG, Wu AYA, 2023. TSPTQ-ViT: two-scaled post-training quantization for vision Transformer. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/icassp49357.2023.10096817

Tian YC, Han JH, Chen HT, et al., 2024. Instruct-IPT: all-in-one image processing Transformer via weight modulation. https://doi.org/10.48550/arXiv.2407.00676

Touvron H, Cord M, Douze M, et al., 2021. Training data-efficient image Transformers & distillation through attention. Proc 38th Int Conf on Machine Learning, p.10347-10357.

Wei XY, Gong RH, Li YH, et al., 2023. QDrop: randomly dropping quantization for extremely low-bit post-training quantization. https://doi.org/10.48550/arXiv.2203.05740

Xia ZR, Dai L, Chen ZH, et al., 2025. Multi-stage feature aggregation Transformer for image rain and haze joint removal. *Eng Appl Artif Intell*, 149:110490. https://doi.org/10.1016/j.engappai.2025.110490

Yuan ZH, Xue CH, Chen YQ, et al., 2022. PTQ4ViT: post-training quantization for vision Transformers with twin uniform quantization. 17th European Conf on Computer Vision, p.191-207. https://doi.org/10.1007/978-3-031-19775-8_12

Zamir SW, Arora A, Khan S, et al., 2022. Restormer: efficient Transformer for high-resolution image restoration. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5718-5729. https://doi.org/10.1109/cvpr52688.2022.00564

Zhang JN, Peng HW, Wu K, et al., 2022. MiniViT: compressing vision Transformers with weight multiplexing. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12135-12144. https://doi.org/10.1109/cvpr52688.2022.01183

Zhang ZC, Chen ZD, Wang YX, et al., 2024. A vision Transformer for fine-grained classification by reducing noise and enhancing discriminative information. *Patt Recogn*, 145:109979. https://doi.org/10.1016/j.patcog.2023.109979

Zheng X, Luo YH, Zhou PY, et al., 2025. Distilling efficient vision Transformers from CNNs for semantic segmentation. *Patt Recogn*, 158:111029. https://doi.org/10.1016/j.patcog.2024.111029

Zhong YS, Lin MB, Li XC, et al., 2022. Dynamic dual trainable bounds for ultra-low precision super-resolution networks. 17th European Conf on Computer Vision, p.1-18. https://doi.org/10.1007/978-3-031-19797-0_1

Zhong YS, Hu JW, Lin MB, et al., 2026. I&S-ViT: an inclusive & stable method for post-training ViTs quantization. *IEEE Trans Patt Anal Mach Intell*, 48(2):1063-1080. https://doi.org/10.1109/TPAMI.2025.3610466

Zhou XC, Ding R, Wang YX, et al., 2023. Cellular binary neural network for accurate image classification and semantic segmentation. *IEEE Trans Multim*, 25:8064-8075. https://doi.org/10.1109/tmm.2022.3233255