SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

DHE: A Semantic-Preserving Framework for Robust Post-Training Quantization of Vision Transformers

Xuze Mao and Yu Chen *

College of Information and Network Security, Yunnan Police College, Kunming 650223, China

* Correspondence: luckyrabbbit450@sohu.com

Received: 25 November 2025; **Revised:** 13 January 2026; **Accepted:** 23 January 2026; **Published:** 3 March 2026

Abstract: Although Vision Transformers (ViTs) have demonstrated impressive achievement in computer vision, they suffer from considerable flaws in deployment due to high computational and memory costs. Post-training quantization (PTQ) is an effective compression method but can cause severe accuracy failure of ViTs, which is mainly caused by the disturbance of attention mechanisms. Based on the scheme of fixed threshold suggested by Zhenhua Liu et al., this paper addresses this drawback by introducing a Dynamic Hybrid Enhancement (DHE) scheme, which changes the quantization paradigm of numerical reconstruction to that of semantic preservation. The main innovations are: the dynamic mechanism of adjusting the loss of rankings by dynamically moving the attention through distribution; the sensitivity of the weight matrix to differences, which prioritizes the semantically important attention connections; and the multi-head normalization strategy, which optimizes attention heads. Numerous experiments on CIFAR-10 and CIFAR-100 show that DHE has accuracy rates of 67.22% and 37.62% compared to the baseline PTQ model (i.e., the fixed-threshold method by Zhenhua Liu et al.), which is 1.82 and 2.23. The role of every single component is confirmed with the help of ablation studies, and the performance of semantic preservation through attention visualization and quantitative measures (e.g., Attention Ranking Preservation Rate, ARPR = 94.7% on CIFAR-10, ARPR = 90.1% on CIFAR-100) proves the superiority of the suggested pattern to the traditional ones.

Keywords: Image Classification; Vision Transformers; Post-Training Quantization; Precision Quantization

1. Introduction

Over the past few years, the Transformers architecture has not only been the subject of a revolutionary success in the field of natural language processing, but also, Visio Transformers (ViT), a variation of the former, has shown promise of fundamentally outperforming the conventional convolutional models in the area of computer vision. Nevertheless, such outstanding performance is accompanied by a high level of both computational and storage overheads that may bring serious challenges to its implementation in embedded devices and edge computing applications. The Transformers is a very sophisticated machine learning system, composed of the self-attention mechanism, sequence-to-sequence problems, and especially used in the domain of Natural Language Processing (NLP) [1]. ViT should prove that a pure attention-based network can yield more successful results in comparison to a convolutional neural network, particularly in the case of huge datasets, like JFT-300 or ImageNet-21K [2]. To have an understanding of visual things, it is necessary to understand a complex visual relationship among objects in a scene [3]. Although deep neural networks are well-suited when dealing with these more complex tasks, they also have large computational and storage requirements that introduce bottlenecks to real-time inference and make the

development of ViT models even more challenging. The PTQ techniques may be used to trim down the model size in order to assist in the development and deployment of efficient models. Nevertheless, PTQ remains constrained, even in the case of uncomplicated or simple PTQ algorithms in general; one of the most frequent effects of using simple PTQ algorithms is the drastic loss of accuracy, to the point that the quantized model becomes incapable of doing anything useful.

The study is aimed at discussing the degradation of performance and the decline in the accuracy of ViT due to PTQ. Extending the exact post-training quantization algorithm of Zhenhua Liu et al. [4]. We find that the fixed-threshold ranking-preserving loss is poor at accommodating the distributional differences between attention maps between different attention heads and input samples. So, we assume that the tolerance level of the quantization process must be dynamic. The essence of this study, however, unlike many other pieces of research (e.g., the study of Liu et al. [4]), is the expression of a Dynamic Hybrid Enhancement (DHE) approach. It not only embraces a dynamic threshold but also initiates a difference-sensitive weight matrix to give more importance to retain the attention relations that the model is so sure of being true, and to make use of a multi-head normalization to balance the optimization process. Being the new systematic approach to the central optimization problem of ViT quantization, preservation of value ranking within attention maps is placed as a central outcome of the overall approach to value preservation. This addition is a great boost towards the model in terms of accuracy and stability.

Altogether, this publication introduces new insights and methodological proposals on the practical implementation of ViT models and thus allows their use in broader areas of application, e.g., autonomous driving, robotics, and imaging in medicine. Though it should also be mentioned that the variability of attention distributions of various visual tasks and data domains may be problematic to generalisation performance of quantization techniques. To use an example, anatomy in medical imagery may have dense local representations but sparse global semantics, whereas dynamic objects in autonomous driving may be mixed in with complicated backgrounds, which will demand a dynamic mechanism of attention to do effective modeling both in time and space. Despite the fact that the dynamic boundary adjustment and disparity-sensitive weighting mechanisms explored in the current work are both conceptually aimed at being flexible to the diversity of attention distributions, their validity on larger-scale datasets (such as ImageNet) and cross-domain tasks is still to be proved. Thus, the approach to the methodology of the given research is centered on structural universality and adaptability in the distribution, which are the basis of the further generalization to more complex visual situations. In turn, the given study has some theoretical value and can be used to build smarter and more effective ViT models.

2. Related Work

The popularity of PTQ as a research topic in the model compression domain has increased thanks to its benefits in terms of eliminating retraining, not depending on large-scale data, and being able to run efficiently, deployed on hardware and edge devices.

To reduce the quantization error, the Bit-Split [5] approach uses an unsplit approach to break values of high precision into several low-precision segments using binary decomposition and a dynamic adaptation process that is hierarchical. ADFQ-ViT [6] proposes a post-training quantization technique for Vision Transformers. It is successful in solving the problems of outliers and non-uniform activations distributions by developing an outlier-sensitive block-wise quantizer and a Shift-Log2 quantizer. Q-Drop [7] improves the model robustness by either randomly quantizing activations or weights. This randomness permits threshold and quantization parameter optimization with very sparse unlabeled data, without considering fine-tuning accuracy at extreme bitwidth. MBQuant [8] proposes a new multi-branch architecture of arbitrary bit-width network quantization. It develops parallel quantization branches of the individual weight parameters during training, modeling a variety of precision levels. Knowledge distillation is learned by the target bit-width branch to improve the learning of low-bit models that are difficult to learn. The design allows the creation of multiple versions with quantized arbitration by the same training process, with high efficiency and flexibility compared to accuracy. Drawing on the earlier ideas presented by Fang et al. [9], proposed a post-training procedure of quantization of weights in a piecewise linear format that breaks down weight quantization into various optimizable linear parts. It can low-bit compress without retraining on only a small calibration dataset by controlling segment parameters. Nagel et al. [10] proposed an entirely data-free post-training quantization, skilled with weight equalization and bias rehabilitation to compensate for the oversights. This algorithm can be used to achieve high-efficiency, low-bit compression using the model weights alone. Moreover, Vision Transformers (ViT)

have also been applied to fields like field reconstruction and spatial interpolation in recent works, and an example of this is the VITAE [11]. and its efficient variants [12]. Although similar to these works, there are no direct instances of model quantization; they prove the architectural benefits and ability to preserve features of ViT in the complex visual modeling tasks. By engaging the autoencoder structures and spatial interpolation learning, they support the ability of ViT in long-range dependencies and spatial continuity, as it holds substantial references in the facilitation of ViT in diverse visual tasks. These research works, again, patchily also support the significance of ensuring the inner semantic frameworks of ViT (i.e., attention relationships), which resonates with the semantic-preserving paradigm suggested in this paper.

Nevertheless, the above techniques do not have adaptive tolerance when performing quantization that can interfere with the mechanisms of attention. Conventional methods usually use global optimization that involves the maximum optimization of the whole layers or networks. Nonetheless, with Vision transformers, the attentions of different heads tend to focus on different relationships with the image, with very different numerical values. An integrated objective of maximization can seriously impair the performance in some heads. Today, various papers have discussed the adjustments in quantization strategy through dynamic means to improve the performance of models.

FedDDO [13] uses dynamic bit-width technology to scale the quantization precision on resource status or data sensitivity which is generally applied during federated learning in which clients dynamically set bit-widths based on real-time resource conditions. NROWAN-DQN [14] is an exploration technique that aims at maximizing the stability and efficiency of deep reinforcement learning. It promotes exploration by placing noise on the weights of the neural network, and finds a more ingenious way of dynamically managing the noise level by adding noise decay and an online weight adjustment technique. This method balances in a better manner between coverage of exploration and the stability of training, which is effective in eliminating the controversies of traditional noisy networks. The quantization complexity of both weights and activations is traded off to a dynamic difficulty moving average of both to a dynamic difficulty megaparameter with a migration strength megaparameter to quantitatively migrate difficulty of challenging activations to more manageable weights (introducing outliers), and vice versa [15]. Although AIQViT [16] presents a structural-sensitive low-rank compensation using a Dynamic Focus Quantizer (DFQ) as a specific aim of imbalanced Softmax distribution, its gain maximizing objective is still efficient bit allocation with numerical approximation of distributions. Importantly, FedDDO, NROWAN-DQN and Smoothquant are all oriented towards federated learning [13], reinforcement learning (RL) [14] and large language models (LLM) [15] respectively. Being mismatched by nature with Vision Transformers is due to architectural discrepancies with self-attention mechanisms. Although Dynamic Focus Quantizer (DFQ) of AIQViT dynamically adjusts ranges, it is aimed at the efficient distribution of bit-width to numbers that are distributed in an approximation. Conversely, our difference-driven weight matrix gives semantic significance in order to actively maintain such relationships that are very confident in the model. The SpectFormer [17] suggests a hybrid spectral and multi-head attention architecture that balances the interaction of the features of high frequencies and the interaction of tokens. Nevertheless, it does not take into account the compatibility of quantization and the additional spectral layers of its use, which present novel challenges to Post-Training Quantization (PTQ).

To summarize, a thorough discussion demonstrates that post-training quantization of Vision Transformers (ViT) remains in several areas of major shortage: First, a lack of flexibility. The majority of approaches apply the one-size-fits-all approach, overlooking the different distributions of attention among heads and samples, hence the errors become uneven. Second, unobserved semantic priority. Attention is paid to the reduction of numerical errors, without making a direct distinction between semantically important and unimportant relationships of attention. Third, narrow optimization. This quantization is commonly done layer-by-layer, and head-wise balance is not considered during multi-head attention, and such can be debilitating to particular heads. Fourth inadequate metrics. The final accuracy is essential as an evaluation tool, and it lacks the direct measurements to evaluate the preservation of attention and semantic loss diagnostics.

These gaps have been directly filled in our proposed DHE approach: it scales up to distributions through dynamic boundary adjustment, focuses on semantics through disparity-sensitive weighting, and scales heads through multi-head normalization, and measures preservation by metrics such as attention ranking retention.

3. Vision Transformers Architecture and Principles

This section describes the ViT architecture, quantization basics, and the mathematics behind precise post-training quantization.

3.1. ViT Structure and Self-Attention

As shown in **Figure 1**, ViT splits an image into fixed-size patches, linearly embeds each of them, adds position embeddings, and feeds the resulting sequence of vectors to a standard Transformers encoder [18]. A learnable [CLS] token is prepended to the sequence as a global representation. The token sequence is processed by a Transformers encoder, which models contextual relationships through self-attention. The Transformer captures long-range dependencies at every layer through its stacked self-attention architecture in the encoder and decoder. It applies channel attention on different network branches to capture cross-feature interactions and learn diverse representations [19]. The output representation of the [CLS] token is then fed into an MLP head for classification.

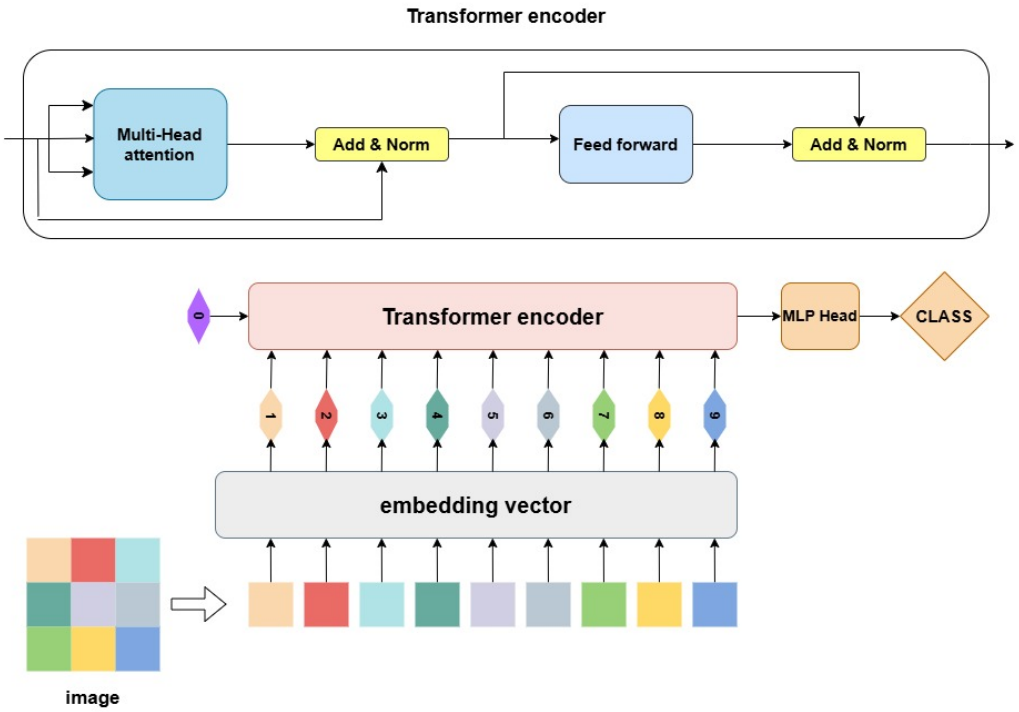

Figure 1. ViT structure.

The core of the ViT model lies in two key steps: Patch Embedding and the Self-Attention Mechanism.

- Patch Embedding maps 2D images into 1D sequences in ViT. It divides the image into patches (**Figure 2**), adds learnable position encodings to encode spatial structure, and prepends a [CLS] token for global representation. The output is a token sequence with semantic and positional data self-attention processing.
- The Self-Attention Mechanism captures global contextual relationships through linear projections and a specialized scoring function, which lets the model decompose an image into many smaller patches and arrange them into a sequence [20]. It dynamically weights relationships between all sequence positions using Query (Q), Key (K), and Value (V) matrices. Attention weights are normalized using the Softmax function, and the output is a weighted sum of values based on these weights. Each word embedding vector in the input sequence, denoted as $X \in \mathbb{R}^{n \times d_{model}}$ (where n is the sequence length and d_{model} is the embedding dimension), is transformed into three distinct representations—Query (Q), Key (K), and Value (V)—via multiplication with three trainable weight matrices: $Q = XW^Q, \quad K = XW^K, \quad V = XW^V$ where $W^Q, W^K \in \mathbb{R}^{d_{model} \times d_k}$, $W^V \in \mathbb{R}^{d_{model} \times d_v}$. The correlation between tokens is then computed as the dot product of Q and K, followed by scaling according to Equation (1) to prevent gradient vanishing issues during training.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

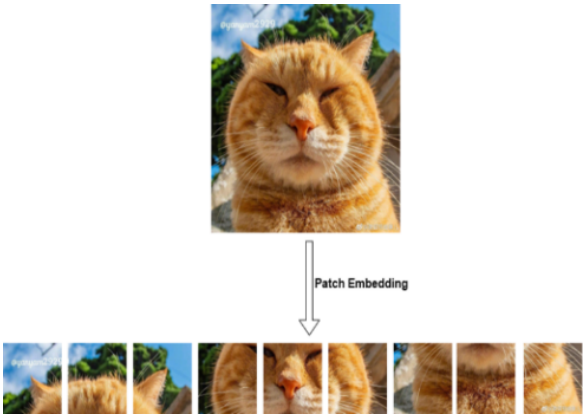

Figure 2. Image tiles.

In Equation (1), $QK^T \in \mathbb{R}^{n \times n}$ represents the attention weights from each position to all other positions; the term $\sqrt{d_k}$ is utilized to scale the dot product, preventing excessively large values that may lead to vanishing gradients in the $Softmax$ function under high-dimensional scenarios; the $Softmax$ function normalizes the values row-wise into a probability distribution where the sum of weights equals 1.

The self-attention mechanism empowers Transformers to handle long-range dependencies, overcoming a key limitation of traditional sequence models like RNNs and CNNs. However, transformers consume significantly higher computational costs compared to CNNs, limiting their feasibility in resource-constrained devices [21].

3.2. Multi-Head Self-Attention

The multi-head attention mechanism performs different learnable linear projections to transform K, V, and Q sets into sub-spaces in parallel. Transformers are a type of neural network with stacked attention and multi-layer perceptron (MLP) blocks. In each layer, the transformer first utilizes multi-head attention Attn to process the input sequence [22]. **Figure 3** illustrates the architecture of Multi-Head Attention.

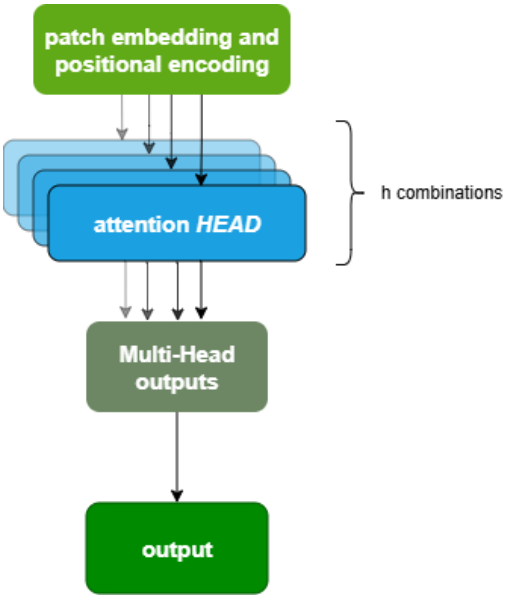

Figure 3. Multi-head attention architecture.

As shown in **Figure 3**, the multi-head attention mechanism enables efficient global context modeling in images through parallelized multi-subspace feature interactions. After patch embedding and positional encoding, multiple independent linear projections map features into different semantic subspaces. The network pays attention to the feature of different parts, and cascades them at last [23] where each attention head independently computes association weights between patches, with attention scores dynamically reflecting inter-region correlations. All head outputs are concatenated and linearly fused to form enhanced features integrating global semantics, according to Equation (2) to introduce the feature fusion of Multi-Head outputs.

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_H)W^o \tag{2}$$

In Equation (2), the $Concat()$ operation concatenates the output vectors of all heads along the feature dimension. The matrix W^o is a learnable projection matrix that linearly fuses the concatenated high-dimensional features into enhanced features and projects them back to the desired output dimension.

This architecture allows Vision Transformers to overcome traditional convolutional receptive field limitations while establishing direct long-range dependencies. The parallelization of attention heads additionally reduces reliance on single attention patterns. To better accommodate visual task requirements, subsequent enhancements like Swin Transformers can be incorporated, employing local window attention mechanisms to maintain performance while improving computational efficiency.

These optimization strategies reduce computational complexity while enhancing robustness to translation and scale variations.

3.3. Quantization Principles

The high memory usage, energy consumption, and latency have hampered the application of Vision Transformers (ViT) in mobile and embedded devices because they should compute highly accurate results relying on the self-attention and the MLP modules that have large parameters. To overcome this, model quantization is a technique that quantizes high-precision parameters with post-training optimization to resolve low-precision storage and computation needs and allows the deployment of ViT to edge intelligence applications. The two popular quantization techniques are presented below: first, Quantization-Aware Training (QAT)—Adds a quantization noise signal at training to train models to run on low-precision computation. It deploys artificially quantized forward passes that round and clip weights and activations. The sensitivity of ViTs may make Softmax frequent, alternatively called dot-product attention, and calculates the similarity between all query-key pairs [24]. QAT, however, needs complete retraining using substantial resources and is hence applicable in accuracy-sensitive industrial applications. Second, Post-Training Quantization (PTQ) quantizes pre-trained models with minimal unlabeled calibration set in a static manner. Vision Transformers (ViT) are, however, very sensitive to quantization, especially in class tokens and attention matrices, and can cause massive accuracy distortions under difficult circumstances, such as low-light or small object detection. Despite being more efficient than QAT, PTQ has several serious concerns with regard to robustness, which restrict its viability: 1. When mapping high-precision post-training quantized models for conversational summarization to low-bit integers, a significant drop in model accuracy occurs. The performance of the quantized model is far inferior to that of the full-precision model. 2. The presence of numerous high-magnitude outliers in activations leads to severe degradation during the quantization of conversational summarization models. For example, under the W4/8A16 mixed-precision configuration, the GPT-3 model quantized using the PTQ scheme can still maintain reasonable performance on some question-answering tasks, but suffers significant accuracy degradation on most other tasks and completely loses the ability to generate meaningful text [25]. The core idea for reducing PTQ errors is to find an optimal mapping relationship that minimizes the information loss caused by quantization.

3.4. Precise Post-Training Quantization Model

The core techniques of precise post-training quantization primarily include the following: First, similarity-aware quantization is proposed, which optimizes the quantization intervals of linear layers by maximizing the Pearson correlation coefficient between the full-precision and quantized outputs. Second, for the self-attention mechanism, ranking-aware quantization is introduced. It preserves the relative order of elements in attention maps

through a ranking loss function, preventing functional degradation of the attention mechanism due to quantization. Third, based on nuclear norm analysis of attention maps and output features, a mixed-precision quantization scheme is proposed. It allocates different bit-widths to layers with varying sensitivity levels, further enhancing quantization efficiency. Each of these methods will be introduced in detail below.

- Quantization Task
  The quantization task is formulated as finding the optimal low-bit quantization intervals for both weights and inputs. The quantization function typically employs a uniform quantizer, which divides the data range into equal intervals. The quantization will be performed according to Equation (3).

$$\Psi_{\Delta}(Y) = Clip(Round\left(\frac{Y}{\Delta}\right), 0, 2^{b-1} - 1) \tag{3}$$

  In Equation (3), Δ is the quantization interval (scale factor); b is the quantization bit-width; Y is the tensor representing weights or inputs; $Clip$ denotes the operation that clips elements exceeding the quantization range.
- Similarity-Aware Quantization for Linear Operations
  For linear layers, the quantization intervals for weights and inputs are optimized to maximize the similarity between the feature maps before and after quantization. This similarity is typically measured using the Pearson correlation coefficient. The quantization process follows Equation (4).

$$\max_{\Delta_l^W, \Delta_l^X} \frac{1}{N} \sum_{i=1}^{N} \Gamma(O_l^i, \hat{O}_l^i) \; s.t. \; \Delta_l^W, \Delta_l^X \in R^+ \tag{4}$$

  In Equation (4), $\Gamma(O_l^i, O_l^i)$ represents the similarity between the original and quantized output feature maps. The Pearson correlation coefficient in Equation (5) is adopted as the similarity metric:

$$\Gamma\left(\hat{O}, O\right) = \frac{\sum_{j=1}^{m}\left(O_j - \overline{O}\right)\left(\hat{O}_j - \hat{O}\right)}{\sqrt{\sum_{j=1}^{m}\left(O_j - \overline{O}\right)^2}\sqrt{\sum_{j=1}^{m}\left(\hat{O}_j - \overline{\hat{O}}\right)^2}} \tag{5}$$

- Ranking-Aware Quantization for Self-Attention
  For the self-attention layer, a ranking loss is introduced to preserve the relative order of attention values. The ranking loss is computed using a hinge function, which yields a loss of zero only when the attention value pairs are correctly ordered and differ by at least a specified margin. The hinge function is defined in Equation (6) as follows:

$$\max_{\Delta_l^W, \Delta_l^x} \frac{1}{N} \sum_{i=1}^{N} \Gamma(O_l^i, \hat{O_l^i}) - \gamma \cdot L_{ranking} \; s.t. \; \Delta_l^W, \Delta_l^X \in R^+ \tag{6}$$

  In Equation (6), $\mathcal{L}_{ranking}$ denotes the pairwise ranking loss function; γ is a trade-off hyperparameter. $L_{ranking}$ is defined in Equation (7) as follows:

$$\mathcal{L}_{ranking} = \sum_{k=1}^{h} \sum_{i=1}^{w-1} \sum_{j=i+1}^{w} \Phi\left(\left(\hat{A}_{ki} - \hat{A}_{kj}\right) \cdot sign\left(A_{ki} - A_{kj}\right)\right) \tag{7}$$

  In Equation (7), $\Phi(p) = (\theta - p)_+$ denotes the hinge function with parameter θ, and (h, w) represents the height and width of matrix A.
- Bias Correction
  The quantization error of weights and inputs is computed, and the expectation of the quantization error is calculated. This expected error is subtracted from the layer output to correct for the accumulated quantization error. The following function (8) is adopted to subtract the expected error on the output from the biased output, ensuring that the mean value of each output unit remains consistent:

$$E[\hat{O}] = E[O] + E[\epsilon^W X] + E[\epsilon^X W] + E[\epsilon^X \epsilon^W] \tag{8}$$

In Equation (8), ϵ^X and ϵ^W denote the quantization errors of the input and weights, respectively.

- Mixed-Precision Quantization for Vision Transformers
  Mixed-precision quantization makes use of different quantization precision for different layers of a NN [26]. Different Transformers layers exhibit varying sensitivity to quantization, making it suboptimal to assign the same bit-width to all layers. The sensitivity of each layer is estimated by computing the nuclear norm of attention maps and output features, and a mixed-precision quantization scheme is employed to determine the bit-width configuration. Layers with higher nuclear norms are allocated higher bit-widths to preserve their performance. The quantization priority is determined by referencing the following function (9):

$$\Omega = \sum_{i=1}^{L} \Omega_i = \sum_{i=1}^{L} \sum_{j=1}^{m} \sigma_j(\mathbf{Y}_i) \cdot \| \hat{\mathbf{Y}}_i - \mathbf{Y}_i \|_2^2 \tag{9}$$

In Equation (9), σ_j denotes the *j* singular value, and $\mathbf{Y}_i$ represents the features of the *i* layer.

4. Innovation Work

4.1. Dynamic Hybrid Enhancement Methods

In the evaluation of image retrieval, there are normal evaluation metrics based on the score ranking [27]. The hinge loss used to quantize the attention values by the traditional post-training methods for ViT PTQ [28] has a constant threshold θ limiting the number of ranks the values can assume. The loss function is defined as follows:

$$\mathcal{L}_{ranking} = \sum_{k=1}^{h} \sum_{i=1}^{w-1} \sum_{j=i+1}^{w} \Phi\left((\hat{A}_{ki} - \hat{A}_{kj}) \cdot sign\left(A_{ki} - A_{kj}\right)\right) \tag{10}$$

Where $\hat{A}$ and A represent the quantized and original attention maps, respectively. Nevertheless, the constant threshold has trouble adjusting to the changes in the dynamic range of the various attention maps, resulting in cumulative quantization errors. To remedy this, the following innovative designs are suggested in this paper:

- Dynamic Boundary Adjustment
  Dynamically change the tolerance threshold θ depending on the distribution properties of attention values.

$$\theta_{eff} = \theta \cdot (1 + \gamma \cdot \sigma(A)) \tag{11}$$

Where $\sigma(A)$ is the standard deviation of values of attention, and γ is a scaling factor. In this process, the standard deviation becomes a modulation factor that allows the threshold to be varied adaptively by the variation in the attention map, making it more robust to outliers.
Equation (11) modulates the fixed threshold θ by a factor proportional to the standard deviation. When the distribution of the attention map is dispersed, θ_{eff} automatically increases, allowing a larger quantization tolerance margin to protect critical outliers from being compromised. When the distribution is concentrated, θ_{eff} approaches θ, enabling fine-grained order preservation. Here, γ acts like a "sensitivity knob," fine-tuning the intensity of our response to such distributional variations.
This process enables the threshold to be adaptively changed depending on the degree of variability (standard deviation $\sigma(A)$) of the attention map. Experiments show that on both CIFAR-10 and CIFAR-100 DHE is more robust and less sensitive than the default PTQ, which essentially addresses the problem of distribution mismatch due to uniform thresholds.

- Difference-Sensitive Weight Matrix
  The weight matrix is obtained and computed with the Sigmoid function applied on the differences in attention:

$$W_{ij} = \text{Sigmoid}(|A_{ki} - A_{kj}|) \tag{12}$$

The raw difference is plotted in Equation (12) to generate a weight value of 0 to 1. The greater difference means that there will be a weight nearer 1 to assign it a greater share of the total loss. This informs the maximization

process to put the rightness of those key relationships first. Under this Equation the loss function is magnified in order to pay attention to the most important comparisons. It was experimentally shown that it allows the Attention Ranking Preservation Rate (ARPR) to achieve over 90% in both datasets, CIFAR-10 and CIFAR-100.

- Multi-Head Normalization

  To prevent loss scale imbalance caused by the multi-head attention mechanism, independent normalization is applied to the loss of each attention head:

$$L_{ranking} = \frac{1}{H}\sum_{h=1}^{H}\sum_{ij} W_{ij} \cdot max\left(0, \theta_{eff} - \Delta\hat{A}_{hij} \cdot sign\left(\Delta A_{hij}\right)\right) \tag{13}$$

  Where H is the number of attention heads, $\Delta\hat{A}_{hij} = \hat{A}_{hi} - \hat{A}_{hj}$, $\Delta A_{hij} = A_{hi} - A_{hj}$.

In Equation (13), the loss per head is first calculated separately and then averaged. This guarantees that every attention head becomes equally important towards the overall loss, irrespective of its numerical value. It also ensures that the optimization process enhances the performance of all the heads in an equal measure, as opposed to having a few heads taking charge of the whole process. It has been experimentally demonstrated that the introduction of the multi-head normalization module was accompanied by a reduction in the training oscillations with the KL divergence decreasing to 0.63. This allows the quantized model to better generate an approximation of the output distribution of the full precision model. The dynamic threshold adjustment is an automatic way of controlling the tolerance value on the loss function, which depends on the data distribution (standard deviation) of the attention map. It loosens the constraint when the data has high variability to safeguard the important information, and narrows the constraints when the data is not spread out to achieve fine-grained optimization. This adaptive threshold is directly involved in computing the rank-aware loss that allows the loss function to be flexible in applying ranking restrictions based on the dynamism of the input.

These Dynamic Hybrid Enhancement (DHE) techniques are a viable remedy to the violation of global dependencies due to quantization by generation attention diversity, which has offered a conceptual assurance of the effective implementation of Vision Transformers.

4.2. Algorithm Design

This section focuses on the implementation of DHE Methods. The core function designs and workflow are outlined as follows (**Algorithm 1**).

Algorithm 1 Dynamic Hybrid Enhancement Methods

Referenced Equations: Equations (11)–(13);
Input:
original_attention_maps: List of attention maps from the full-precision model, shape [B, H, S, S];
quantized_attention_maps: List of quantized attention maps, same shape as original, representing attention weight matrices after quantization;
margin: Base tolerance threshold θ (default 0.1);
gamma: Standard deviation adjustment coefficient γ (default 0.1);
Output:
scalar_loss: A scalar loss value;

```
(1)  total_loss ← 0.0
(2)  for each orig_map in original_attention_maps and quant_map in quantized_attention_maps do
(3)    std ← std(orig_map) # Compute standard deviation of original attention
(4)    dynamic_margin ← margin × (1 + gamma × std)
(5)    orig_diff ← unsqueeze(orig_map, -1) - unsqueeze(orig_map, -2) # Shape [B, H, S, S]
(6)    quant_diff ← unsqueeze(quant_map, -1) - unsqueeze(quant_map, -2)
(7)    weight_matrix ← sigmoid(abs(orig_diff)) # Assign high weights to important differences
(8)    aligned_diff ← quant_diff × sign(orig_diff) # Maintain sign consistency
(9)    layer_loss ← weight_matrix × relu(dynamic_margin - aligned_diff)
(10)   total_loss ← total_loss + mean(layer_loss) / shape(orig_map)[1] # Normalize by number of heads H
(11) end for
(12) return scalar_loss ← total_loss / len(original_attention_maps) # Average over multiple layers
```

The core design of this algorithm lies in an adaptive dynamic boundary mechanism that ensures the relative ranking information embedded within the attention maps is preserved to the greatest extent possible after model quantization. The overall workflow can be systematically described as follows:

Initialization and Iteration Preparation (Lines 1–2): This is done by first making the global loss accumulator total loss = 0. Then the algorithm compares corresponding attention map pairs (orig_map, quant_map) of the

full-precision (quantized) model and considers them at each layer by calculating the standard deviation σ of the attention weight distribution at that layer.

Dynamic Tolerance Boundary Calculation (Lines 3–4): In each layer, the algorithm calculates the standard deviation σ of the original attention map orig_map, which is a measure of the dispersion in attention weight distribution in that layer. This standard deviation is used to adjust the base tolerance threshold θ, generating an adaptive dynamic boundary $\theta_{eff} = \theta \times (1 + \gamma \times \sigma)$. This design makes the algorithm more tolerant of quantization errors in layers with complex, high-variance distributions, while imposing stricter constraints in layers with concentrated distributions.

Relative Ranking Difference Construction (Lines 5–6): The algorithm calculates the difference matrices orig_diff and quant_diff for the original and quantized attention maps, respectively, using tensor expansion and broadcasting mechanisms. The key pairs of the relative importance relationships of all of the important pairs of the same query position are coded in these two five-dimensional tensors [B, H, S, S, S], upon which the ranking loss comparison is based.

Difference Alignment and Importance Weighting (Lines 7–8): In order to make the comparison valid, the algorithm carries out two important operations: 1. Importance Weighting: Different weights associated with the original differences are produced and taken through a sigmoid function to produce a weight_matrix. In this operation, the pairs with the largest difference in significance in the original model are weighted more, giving the loss operation's center more attention to the most important pairs in the decision the model makes. 2. Sign Alignment: The quantized difference quant_diff is multiplied with sign of original difference orig_diff, which creates aligned difference. This makes the ranking changes in the quantized model rated in the same direction.

Layer Loss Calculation and Accumulation (Lines 9–10): The core loss is computed via relu(θ_{eff} − aligned_diff). This function penalizes the model only when the quantized model fails to maintain sufficient ranking distinction (i.e., aligned_diff < θ_{eff}). The weighting of the loss is then done by the weight_matrix so as to focus on the maintenance of important ranking relationships. The calculated loss of the layer is averaged both on the batch and spatial dimensions and divided by the number of attention heads H to counterattack scale bias that comes with the multi-head structure. This is finally added to the global loss.

Global Loss Normalization and Output (Lines 11–12): After processing all attention layers, the accumulated total_loss is divided by the total number of attention layers to obtain the average loss across all layers. This final scalar loss value is returned and used to guide the quantization-aware training process. Through gradient descent, it optimizes the model parameters to accurately maintain the attention ranking structure of the original model even under low-precision representation.

This dynamic ranking loss function addresses the loss of critical relative ranking information during model quantization by integrating an adaptive boundary, importance weighting, and sign alignment strategy. Therefore, it preserves the core inference performance of the model while achieving compression.

4.3. Compare with Recent Works

The quantization of the weights and activations of a model is currently considered the primary idea of mainstream post-training quantization, such as ADFQ-ViT [6], Q-Drop [7], and MBQuant [8]. Studies concerning retaining the order of attention maps in Vision Transformers are fairly few, though, and it prevents the successful preservation of the principles which form their core functionality during quantization.

The main strength of the mechanism of Dynamic Hybrid Enhancement (DHE) is that it does not depend on any particular model architecture assumptions. Rather, it gives an approximation solution to the underlying problem of maintaining relative order when quantizing.

Specifically:

Dynamic Boundary Adjustment is an adaptation system dependent on the statistical characteristics (standard deviation) of the attention map, and thus it is an information-driven approach that assumes no structure or distribution.

The essence of the Difference-Sensitive Weight Matrix mechanism is to pay attention to the value of relative differences among elements.

Thus, the DHE approach does not have certain assumptions regarding the self-attention mechanism of the Vision Transformers. Its fix, how to create successful quantization without loss of the relative ranking of the numerical

connections, is a problem that is present in all model compression. Its principles of design are applicable across any quantization problem that is based on an interest in ensuring that there is a relative significance or rank between features and, therefore, is highly indicative of generalization.

In order to conduct a systematic assessment of the features of various approaches, **Table 1** entails a comparison between mainstream post-training quantization models, among a number of dynamic quantization schemes, and against the method that is proposed in the present study.

Table 1. Compare with recent works.

Methods	Attention Order	Dynamic Range	Difference Importance	Semantic-Level	Error Accumulation
ADFQ-ViT	Y	N	N	N	N
Q-Drop	N	N	N	N	N
MBQuant	N	N	Y	N	N
VT-PTQ	Y	N	N	Y	Y
FedDDO	N	Y	N	N	N
NROWAN-DQN	N	Y	N	N	N
Smoothquant	N	Y	N	N	N
AIQViT	Y	Y	N	N	N
We	Y	Y	Y	Y	Y

The paper is a new post-training quantization paradigm for Vision Transformers that achieves the advancement of the limitations of traditional approaches, where computation is reduced to numerical reconstruction of the weights and activations. In comparison to general techniques of the PTQ, such as ADFQ-ViT, or more specific techniques such as AIQViT, our work provides a complete technical framework for attention map quantization protection: it adapts to the variations of the attention distribution by means of an adaptive boundary adjustment mechanism, it preserves the critical semantic relationships by way of a weight-based difference sensitivity mechanism, and it provides balance in optimization by resorting to multi-head normalization. Therefore, our study attains innovations in five fundamental dimensions, specifically: the preservation of attention order, the dynamic range adaptation, the awareness of the importance of a difference, semantic protection, and the control of error accumulation. The paper offers both theoretical basic building blocks and real implementation paths to effective implementation of ViT models, and has suggested a paradigm shift in quantization of models from being based on traditional numerical reconstruction to model semantic preservation.

5. Results Analysis

This section evaluates model performance on CIFAR-10 and CIFAR-100. CIFAR-10 contains 50 K training and 10 K test images across 10 categories; CIFAR-100 contains 50 K training and 10 K test images across 100 categories. Experiments include comparative studies and ablation studies: comparative experiments assess model size and accuracy across standard PTQ, fine-grained PTQ, and the proposed enhanced PTQ methods; ablation studies examine the impact of different modules on performance and validate the effectiveness of the proposed improvements. Model type abbreviations are listed in **Table 2**.

Table 2. Model type mapping.

Model ID	Model Type
VITP	A PTQ quantized model without any advanced quantization schemes
VITO	A full-precision model without any quantization
VITF- *	The model utilizing precise post-training quantization
VITFA- *	Improved PTQ-based precise post-training quantization
VITP-S	Add Similarity-Aware function
VITP-SRB	Add Similarity-Aware function, DHE methods, Bias-Correction function
VITP-SRBA	Add Similarity-Aware function, DHE methods, Bias-Correction function, Alternating-Optimization function

Note: The asterisk (*) denotes the number of training rounds (or epochs).

5.1. Comparative Studies

As shown in **Figure 4**, for the CIFAR-10 dataset (represented by the blue and pink bars), as the model version progresses from 1 to 12, both VITF and VITFA show a trend of increasing accuracy. In earlier versions (such as version 1), there is a noticeable gap in accuracy between the two models. However, as the version number rises,

the accuracy values of both models grow steadily, and in later versions (like version 12), they reach relatively high levels, with VITFA demonstrating a slightly better performance in some versions.

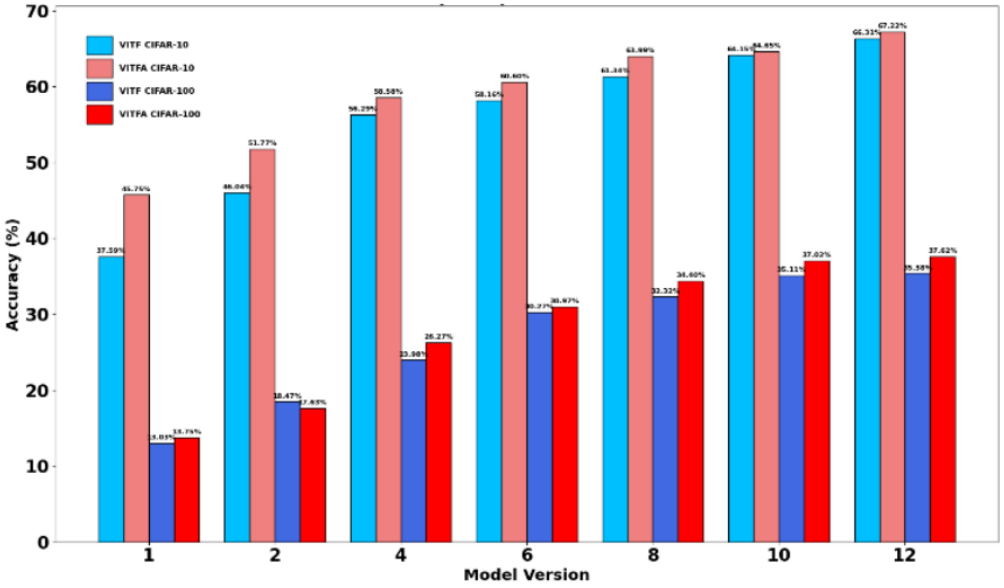

Figure 4. VITF and VITFA Accuracy Comparison on CIFAR-10 and CIFAR-100.

Regarding the CIFAR-100 dataset (indicated by the dark blue and red bars), similar to the CIFAR-10 case, the accuracy of both models also improves with the advancement of model versions. Nevertheless, the overall accuracy on CIFAR-100 is lower than that on CIFAR-10, which is likely due to the greater complexity and larger number of classes in the CIFAR-100 dataset. Additionally, VITFA maintains a competitive edge over VITF in terms of accuracy across various versions for CIFAR-100 as well.

In short, both the VITF and VITFA have the advantage of model version upgrades in accuracy on both data sets, and also, VITFA tends to be more accurate than VITF. In addition, the more complicated classification tasks are depicted by the difference in performance between the datasets. **Tables 3** and **4** balance between model size and accuracy of both methods.

Table 3. Compare test data on CIFAR-10.

Model ID	Accuracy	Model Size
VITP	6.33%	9.4 MB
VITO	47.21%	16.8 MB
VITF-1	37.59%	11.96 MB
VITFA-1	45.75%	12.8 MB

Table 4. Compare test data on CIFAR-100.

Model ID	Accuracy	Model Size
VITP	5.34%	9.43 MB
VITO	15.42%	16.82 MB
VITF-1	10.75%	11.98 MB
VITFA-1	13.75%	12.84 MB

As shown in **Tables 3** and **4**, VITP and VITO have a big difference in accuracy. This is because PTQ is sensitive to information loss and distribution shift, so accuracy drops a lot. But VITF and VITFA models keep much higher accuracy even after model compression. Specifically, VITFA loses only about 2% accuracy on CIFAR-10 and CIFAR-100. This shows that fine-grained post-training quantization keeps accuracy well while compressing the model.

Figure 5 shows how the VITFA model's accuracy compares to the VITF model. It shows the difference in their accuracy and compares the accuracy of the VITFA and VITF models on the CIFAR-10 and CIFAR-100 datasets for different model versions.

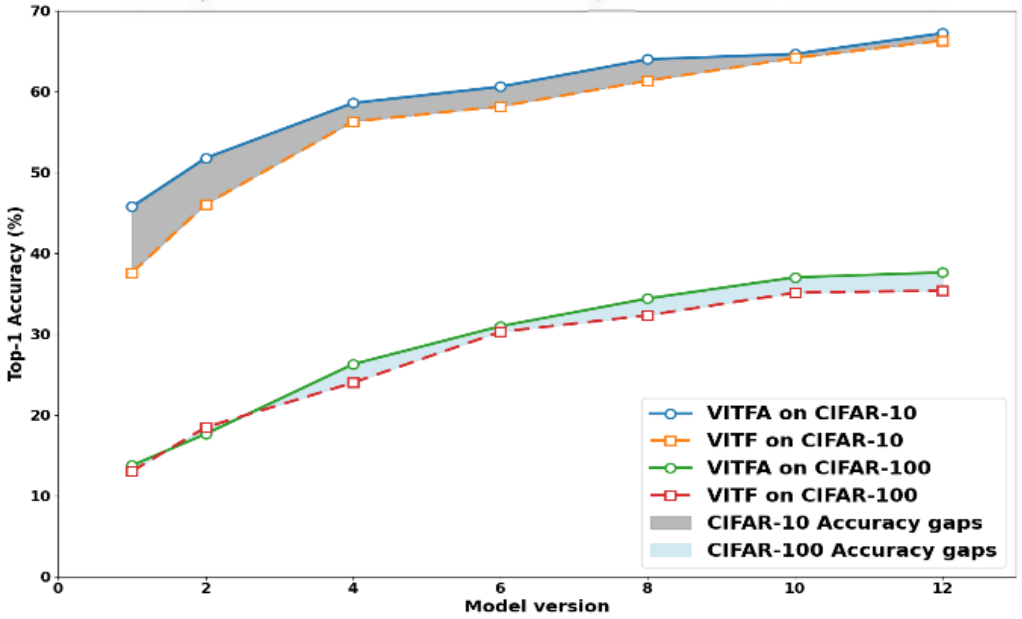

Figure 5. Comparison of VITFA and VITF Accuracy on CIFAR-10 and CIFAR-100.

On the CIFAR-10 dataset, both VITFA (blue line) and VITF (orange line) become more accurate with the increase in the number of model versions. Their accuracy increases as their values decrease. In all versions, VITFA is slightly more accurate than VITF is. VITFA is 66.1% at version 12, and VITF is at 65.4%.

VITFA (green line) and VITF (red line) also exhibit increased accuracy in subsequent model versions on the CIFAR-100 dataset. They have a lower accuracy rate than on CIFAR-10, which probably can be attributed to a greater difficulty in the tasks. VITFA is the best performer of all versions when compared to VITF. VITFA and VITF have an accuracy of 37.62 and 35.39 at version 12, respectively. Accuracy differences between the two models within each set are also represented with the help of the shaded areas in the chart. These fields demonstrate that, despite the improvement of both models, VITFA will be more accurate. There is a relatively small fluctuation in the gap size as the model version increases, as there are small variations.

Based on the comparison of the attention visualization maps, we observe the following clear results:

The conventional post-training quantization techniques damage the attention mechanism of Vision Transformers. The distribution of attention of the baseline PTQ model is poor. The concentration is diffused. Boundaries are blurry. The background noise is of great quality. Such impairment of attention reduces the semantic comprehension of the model. The correlation between the weight of attention and the image content is weakened.

We use our Dynamic Hybrid Enhancement approach, which is more efficient in retention of attention. The attention of the DHE model resembles the full-precision model considerably. It maintains concentration on attention and minimizes quantization noise. This shows our design helps. Dynamic boundary adjustment, difference-sensitive weight matrix, and multi-head normalization minimize the loss of information during the process of quantization.

We used CIFAR-10 and CIFAR-100. We also tested the DHE involvement with varying patterns of attention. This is due to the fact that DHE preserves attention structures as demonstrated in **Figure 6**. This applies to pictures that have a complicated background and several objects. The attention maps of DHE are similar to the full-precision model. This is due to the fact that the dynamic boundary adjustment reverts to its threshold depending on the distribution of attention values. This makes the ranking of attention the same across various data.

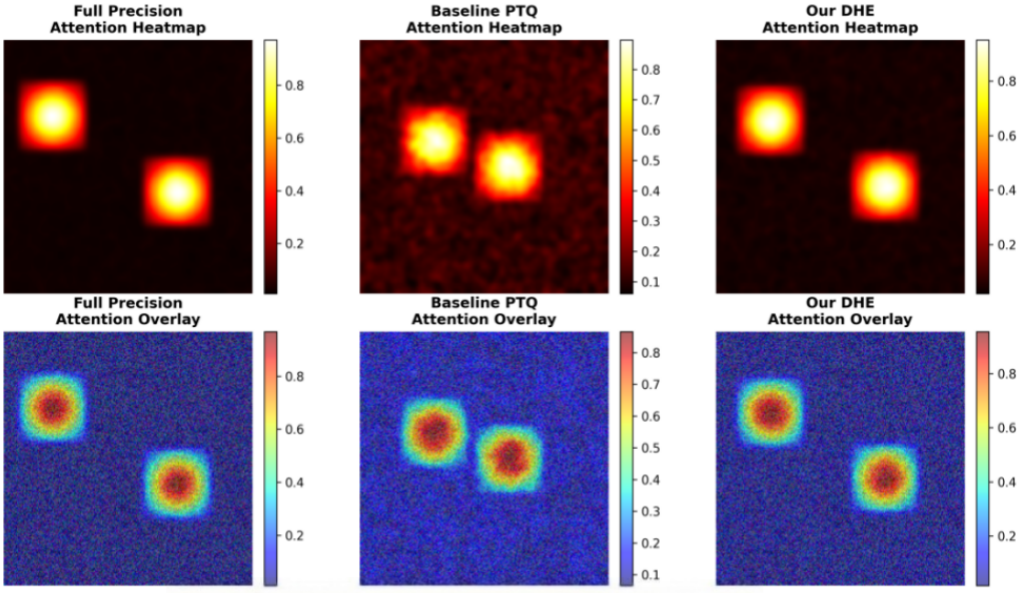

Figure 6. Attention Map Visualization Comparison (Full Precision vs Baseline PTQ vs. Our DHE Method).

5.2. Ablation Studies

This section evaluates the effects of incorporating the similarity-aware module, Dynamic Hybrid Enhancement (DHE) methods, and bias correction module into the VITP model, while also assessing the impact of training iterations on accuracy. The objective is to validate the feasibility of the proposed DHE methods.

To empirically validate the paradigm shift from numerical reconstruction to semantic preservation, this paper proposes the Attention Ranking Preservation Rate (ARPR) as a core evaluation metric.

$$ARPR = \frac{1}{H \times S \times S} \sum_{h=1}^{H} \sum_{i=1}^{S} \sum_{j=1}^{S} I(rank\left(A_{h,i,j}^{fp}\right) = rank(A_{h,i,j}^{quant})) \tag{14}$$

Equation (14): $A_{h,i}^{fp} \in \mathbb{R}^S$ and $A_{h,i}^{quant} \in \mathbb{R}^S$ represent the attention weight vectors (over S keys) corresponding to the h-th attention head and the i-th query position in the full-precision and quantized models, respectively.

The $rank()$ function returns the rank order of elements in a vector (e.g., the largest value is assigned rank 1 in descending order).

$I()$ is the indicator function, which takes the value 1 if the condition inside the parentheses is true, and 0 otherwise.

The denominator S is the normalization factor, representing the number of possible keys for each query position.

The ARPR directly measures the preservation degree of the relative importance of attention before and after quantization. Its value ranges from [1], with higher values indicating better semantic preservation.

The Kullback-Leibler Divergence (KL Divergence) is used to quantify the difference between the output distribution of the quantized model P_{quant} and that of the full-precision model P_{fp}:

$$D_{KL}(P_{fp}||P_{quant)}) = \sum_{i} P_{fp}(i) \log \frac{P_{fp}(i)}{P_{quant}(i)} \tag{15}$$

In the context of this paper, a lower KL divergence indicates a high degree of consistency between the class-level decision distributions of the quantized model and the full-precision model, which indirectly reflects the preservation of semantic understanding. However, as a global distribution metric, KL divergence cannot capture changes

in the internal semantic structure of the attention mechanism. Therefore, we use it complementarily with the proposed ARPR metric: KL divergence measures the similarity of "output results," while ARPR measures the preservation of "internal mechanisms."

Tables 5 and **6** evaluate the effectiveness of incorporating the similarity-aware module, DHE methods module, and bias correction module into the VITP model, along with the impact of training iterations on accuracy.

Table 5. Accuracy of ablation experiments on CIFAR-10.

Model ID	Accuracy
VITP	6.33%
VITP-S	8.31%
VITP-SRB	12.09%
VITP-SRBA	13.68%

Table 6. Accuracy of ablation experiments on CIFAR-100.

Model ID	Accuracy
VITP	5.34%
VITP-S	7.69%
VITP-SRB	9.64%
VITP-SRBA	12.49%

As shown in **Tables 7** and **8**, we assess the semantic preservation of the quantized VITP model using two complementary metrics: the Attention Ranking Preservation Rate (ARPR) and KL Divergence. This evaluation demonstrates the impact of incrementally adding the similarity-aware module (S), Dynamic Hybrid Enhancement (DHE) methods, and the bias correction module.

Table 7. KL divergence for different modules on CIFAR-10.

Model ID	ARPR	KL Divergence	Performance
VITP	72.3%	1.85	low
VITP-S	79.5%	1.42	good
VITP-SRB	89.2%	0.97	better
VITP-SRBA	94.7%	0.63	best

Table 8. KL divergence for different modules on CIFAR-100.

Model ID	ARPR	KL Divergence	Performance
VITP	68.5%	1.86	low
VITP-S	73.2%	1.49	good
VITP-SRB	83.7%	1.18	better
VITP-SRBA	90.1%	0.86	best

As illustrated in the table above, the ablation experiment shows how each individual module plays a progressive contribution to the goal of quantizing optimizing image features: S-Module (Similarity-Aware): This module can make a crucial step towards semantic preservation by achieving a performance improvement in CIFAR-10 by defining its targets to 8.31% and then increasing ARPR by 79.5 and minimizing KL divergence by 1.42. Accuracy on CIFAR-10 goes to 12.09, ARPR goes up to 89.2, and KL divergence goes down to 0.97. It uses dynamic thresholds and disparity-sensitive weighting to directly protect the core semantic structure-attention ranking: it achieves optimal accuracy of 13.68% on CIFAR-10, near-perfect ARPR (94.7%), and the lowest KL divergence (0.63). The three modules are a synergy of a system of numerical reconstruction–semantic protection–fine-tuning. The steady and impressive growth in ARPR, which is closely linked to the decrease of KL divergence, attests to the fact that the most important pathway to sustaining the semantic understanding ability of ViT, as well as the high-quality quantization performance, is to preserve sorting of attention.

In order to explore further the effect of integrating DHE methods on accuracy, **Tables 9** and **10** give a comparison of the KL divergence distributions presented at various iteration rounds.

Table 9. KL divergence for different iterations on CIFAR-10.

Training Epoch	Model ID	KL Divergence
1	VITF-1 VITFA-1	5.21 3.78
2	VITF-2 VITFA-2	4.15 2.94
4	VITF-4 VITFA-4	3.02 1.67
6	VITF-6 VITFA-6	2.48 1.23
8	VITF-8 VITFA-8	1.95 0.89
10	VITF-10 VITFA-10	1.32 0.71
12	VITF-12 VITFA-12	1.00 0.55

Table 10. KL divergence for different iterations on CIFAR-100.

Training Epoch	Model ID	KL Divergence
1	VITF-1 VITFA-1	5.10 3.70
2	VITF-2 VITFA-2	4.20 2.96
4	VITF-4 VITFA-4	3.29 1.87
6	VITF-6 VITFA-6	2.76 1.46
8	VITF-8 VITFA-8	2.33 1.13
10	VITF-10 VITFA-10	1.88 0.87
12	VITF-12 VITFA-12	1.58 0.69

As the table above indicates, the KL divergence of VITFA is always lower than that of the VITF model, even at the same training iterations. In addition, the KL value of VITFA decreases with an increase in the number of iterations (3.78 to 0.55 on CIFAR-10 and 3.70 to 0.69 on CIFAR-100), showing that this method is highly able to optimize and showing that the DHE techniques are effective.

To visually understand more the variations and convergence trends of the KL divergence between both types of models, that is, the VITF and VITFA models, **Figures 7** and **8** present the trending difference between the KL divergence and the regression curves of the two types of models.

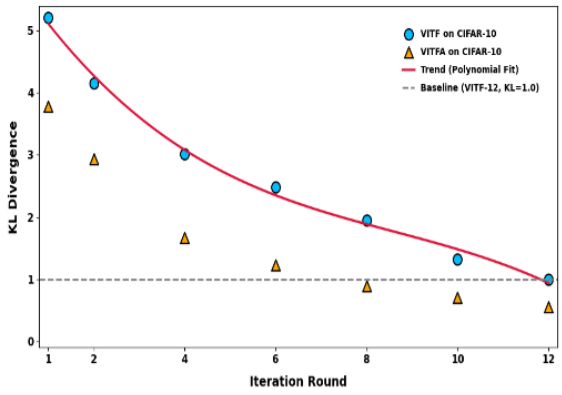

Figure 7. KL Divergence Trends across Iterations on CIFAR-10.

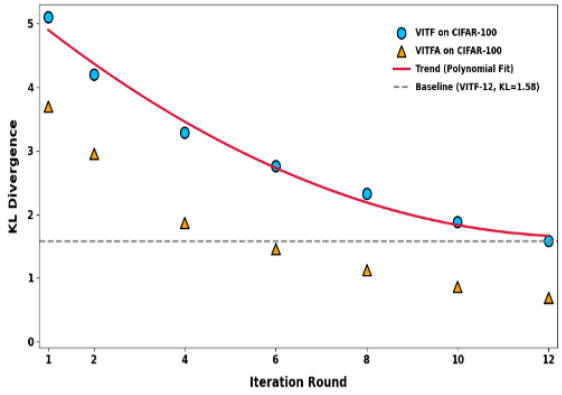

Figure 8. KL Divergence Trends across Iterations on CIFAR-100.

With the KL divergence of the two models as illustrated in **Figures 7** and **8** above, we can state that they both work to produce output at the target event, and this is a confirmation of the effectiveness with which they are quantized. The red trendline is non-linear convergence: a quick decrease in the beginning and slower gains. On the grey dashed line, the 12th round does not bring VITF to the same level as it does by VITFA; therefore, VITFA shows to be more efficient. In general, the distribution alignment of VITFA is better compared to VITF, which proves the superiority of the improved method.

The researchers in their ablation tests prove that the inclusion of the DHE methods enhances the model accuracy greatly, which proves the efficiency of this tool. Moreover, comparisons by use of KL divergence indicate that the better model is always ahead of the traditional fine-grained quantization technique at all times during the training process, which proves that the DHE methods are better than the fixed ranking loss methods.

5.3. Hyperparameter Sensitivity Comparative Experiments

In an attempt to assess fully the strength of the dynamical hybrid enrichment technique to hyperparameters fully, this section performs a comparative assessment between VITF-12 (the fixed-threshold strategy) and VITFA-12 (the dynamical hybrid enhancement strategy) in the context of the dynamic adjusting factor γ. The experiments are conducted on the validation sets of CIFAR-10 and CIFAR-100 to evaluate the effect of different values of 0.00 to 1.00 on the accuracy of the VITF and VITFA models.

Figure 9 shows how different levels of 0.00 to 1.00 affect the results of both models. It is interesting to note that both of these methods maintain maximum performance at 0.10, meaning that at this value of the hyperparameter, both of the quantization strategies will provide optimal results. Nevertheless, there is a great difference between the response patterns of the two methods to the changes in γ.

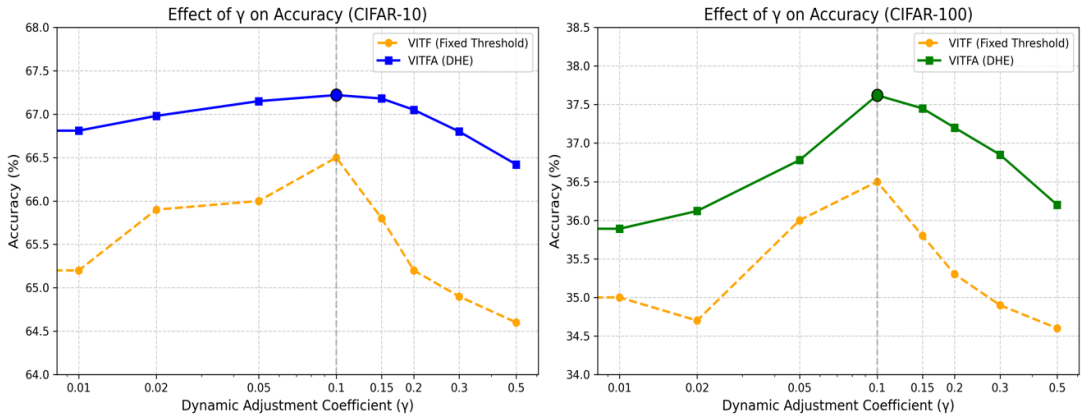

Figure 9. Comparison of the Impact of the Dynamic Adjustment Coefficient γ on the Accuracy of VITF and VITFA.

VITF has a significant volatility on the CIFAR-10 dataset, whose accuracy curve is a unimodal, sharp curve, with its peak closer to 66.50 at 0.10 afterwards. A change from 0.10 to 0.15 in gamma leads to a 0.70 percentage change in the accuracy; a change from 0.10 to 0.05 gamma leads to a −0.50 percentage change in their accuracy. Such a radical change indicates that the fixed threshold technique is highly sensitive to γ drift. By contrast, VITFA is represented by a smooth curve that is unimodal in nature, with the highest accuracy of 67.22 being again at γ = 0.10, but the variation both upwards and downwards is much smoother: the fall in accuracy is only 0.04% as γ increases to 0.15 and also only 0.07% as 0.10 to 0.05. Such gradual dynamics of change of VITFA indicate that its proactive adaptation system can adequately mitigate the effect of the changes in Equation (7).

Two-fold dissimilarities between the two approaches are even more evident on the CIFAR-100 data. VITF is at its best with 36.50% accuracy obtained at 0.10 gamma, but this declines drastically as the gamma changes to 36.00% when gamma is smaller than 0.10 and to 35.80% when gamma is bigger than 0.10. VITFA, in its turn, exhibits the highest possible accuracy of 37.62 at 0.10 and is quite stable throughout the experimented range of 0.05 to 0.20: the difference in accuracy does not exceed 1.5, where 0.10–0.20 is a range. More specifically, it is remarkable that as γ is bigger than 0.20, the performance that VITFA will witness becomes a slower process, but VITF will experience a faster downward trend.

Overall, the comparative experiments involving hyperparameter sensitivity not only confirm the better performance of VITFA, but it is more crucial to note that it is much more enhanced in terms of robustness. With the inclusion of a dynamic hybrid enhancement mechanism, VITFA can be adjusted better to the changes in hyperparameters, which makes the model less reliant on the exact parameter settings.

6. Conclusion

ViTs have performed much better than other computer vision models in vision tasks, but are memory-intensive and computationally expensive, limiting their use on resource-limited devices. Hence, the given work discusses the large drop in accuracy of the post-training quantization of Vision Transformers and provides a new framework of Dynamic Hybrid Enhancement (DHE), providing a paradigm shift from the traditional approach of numerically recreating an image. The proposed DHE first defines preservation as the main optimization goal of ViT quantization and introduces a unified model, which combines dynamic boundary adjustment, a disparity-sensitive weight matrix, and multi-head normalization to address distribution adaptability, semantic importance awareness, and optimization balance. Specifically, DHE has no additional inference overhead and proposes intermediate evaluation criteria like Attention Ranking Preservation Rate (ARPR) to offer new instruments for studying the inner processes of quantization techniques. The proposed DHE achieves the best performance on CIFAR-10 and CIFAR-100, which can demonstrate that it is more useful in practice.

At present, the proposed DHE does not assume finer-grained channel-wise or group-wise quantization. The further work will be concentrated on the optimization of the quantization granularity and on the creation of a mathematically sound bit-width allocation module. These innovative technologies, used in combination with the suggested DHE approach, would also lead to even stronger and more efficient quantized ViT models. In addition, its generalization to larger-scale datasets (e.g., ImageNet) and cross-domain tasks (e.g., medical imaging) requires further validation. In the future, we will explore: (1) adaptive bit-width allocation and finer-grained quantization; (2) extension to other Transformer variants; (3) automated hyperparameter optimization; and (4) end-to-end deployment validation on edge devices.

Author Contributions

Conceptualization, X.M. and Y.C.; methodology, X.M. and Y.C.; software, X.M.; validation, X.M.; formal analysis, Y.C.; investigation, X.M.; resources, Y.C.; data curation, X.M.; writing—original draft preparation, X.M.; writing—review and editing, X.M. and Y.C.; visualization, X.M.; supervision, Y.C.; project administration, Y.C.; funding acquisition, Y.C. Both authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets ANALYZED for this study can be found at the Canadian Institute for Advanced Research http://www.cs.toronto.edu/~kriz/cifar.html.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Li, Y.; Ma, Z.; Wang, Y.; et al. Survey of Vision Transformers (ViT). *Comput. Sci.* **2025**, *52*, 194–209. [CrossRef]
2. Guo, M.-H.; Xu, T.-X.; Liu, J.-J.; et al. Attention Mechanisms in Computer Vision: A Survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
3. Vaishnav, M.; Cadene, R.; Alamia, A.; et al. Understanding the Computational Demands Underlying Visual Reasoning. *Neural Comput.* **2022**, *34*, 1075–1099. [CrossRef]
4. Liu, Z.; Wang, Y.; Han, K.; et al. Post-Training Quantization for Vision Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28092–28103.
5. Wang, P.; Chen, Q.; He, X.; et al. Optimization-Based Post-Training Quantization With Bit-Split and Stitching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 2119–2135. [CrossRef]
6. Jiang, Y.; Sun, N.; Xie, X.; et al. ADFQ-ViT: Activation-Distribution-Friendly Post-Training Quantization for Vision Transformers. *Neural Netw.* **2025**, *186*, 107289. [CrossRef]
7. Nguyen, P.T.Q.; Khanh, T.C.; Ergu, Y.A.; et al. Q-Drop: Optimizing Quantum Orthogonal Networks with Statistic Pruning and Dynamic Dropout. In Proceedings of the IEEE International Conference on Communications, Montreal, QC, Canada, 8–12 June 2025; pp. 2394–2399. [CrossRef]
8. Zhong, Y.; Zhou, Y.; Chao, F.; et al. MBQuant: A Novel Multi-Branch Topology Method for Arbitrary Bit-Width Network Quantization. *Pattern Recognit.* **2025**, *158*, 111061. [CrossRef]
9. Fang, J.; Shafiee, A.; Abdel-Aziz, H.; et al. Post-Training Piecewise Linear Quantization for Deep Neural Networks. In *Computer Vision –ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., et al., Eds.; Springer: Cham, Switzerland, 2020; 12347, pp. 66–89. [CrossRef]
10. Nagel, M.; Baalen, M.; Blankevoort, T.; et al. Data-Free Quantization through Weight Equalization and Bias Correction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 1325–1334. [CrossRef]
11. Fan, H.; Cheng, S.; de Nazelle, A.J.; et al. An Efficient ViT-Based Spatial Interpolation Learner for Field Reconstruction. In *Computational Science –ICCS 2023*; Mikyška, J., de Mulatier, C., Paszynski, M., et al., Eds.; Springer: Cham, Switzerland, 2023; 10476, pp. 430–437. [CrossRef]
12. Fan, H.; Cheng, S.; de Nazelle, A.J.; et al. ViTAE-SL: A Vision Transformer-Based Autoencoder and Spatial Interpolation Learner for Field Reconstruction. *Comput. Phys. Commun.* **2025**, *308*, 109464. [CrossRef]
13. Zhang, S.; Han, Q.; Wang, H.; et al. Federated Learning with Dual Dynamic Quantization Optimization in Smart Agriculture. *Internet Things* **2025**, 101798. [CrossRef]
14. Han, S.; Zhou, W.; Lu, J.; et al. NROWAN-DQN: A Stable Noisy Network with Noise Reduction and Online Weight Adjustment for Exploration. *Expert Syst. Appl.* **2022**, *203*, 117343. [CrossRef]
15. Xiao, G.; Lin, J.; Seznec, M.; et al. SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models. *arXiv preprint* **2022**, *arXiv:2211.10438*. [CrossRef]
16. Jiang, R.; Zhang, Y.; Wang, L.; et al. AIQViT: Architecture-Informed Post-Training Quantization for Vision Transformers. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 17635–17643. [CrossRef]
17. Patro, B.N.; Namboodiri, V.P.; Agneeswaran, V.S. SpectFormer: Frequency and Attention Is What You Need in a Vision Transformer. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 26 February–6 March 2025. [CrossRef]

18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv preprint* **2020**, arXiv:2010.11929. [CrossRef]
19. Liu, Y.; Wu, Y.H.; Sun, G.; et al. Vision Transformers with Hierarchical Attention. *Mach. Intell. Res.* **2024**, *21*, 670–683. [CrossRef]
20. Li, Y.; Wang, J.; Dai, X.; et al. How Does Attention Work in Vision Transformers? A Visual Analytics Attempt. *IEEE Trans. Vis. Comput. Graph.* **2023**, *29*, 2888–2900. [CrossRef]
21. Qin, H.; Zhou, D.; Xu, T.; et al. Factorization Vision Transformer: Modeling Long-Range Dependency with Local Window Cost. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 3151–3164. [CrossRef]
22. Chen, X.; Zhao, L.; Zou, D. How Transformers Utilize Multi-Head Attention in In-Context Learning? A Case Study on Sparse Linear Regression. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 119573–119613. [CrossRef]
23. Zhang, Y.; Xu, B.; Zhao, T. Convolutional Multi-Head Self-Attention on Memory for Aspect Sentiment Classification. *IEEE/CAA J. Autom. Sin.* **2020**, *7*, 1038–1044. [CrossRef]
24. Han, D.; Pu, Y.; Xia, Z.; et al. Bridging the Divide: Reconsidering Softmax and Linear Attention. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 79221–79245. [CrossRef]
25. Yao, Z.; Yazdani Aminabadi, R.; Zhang, M.; et al. ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers. In Proceedings of the 36th Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 27168–27183. Available online: https://dl.acm.org/doi/10.5555/3600270.3602240
26. Wu, D.; Wang, Y.; Fei, Y.; et al. A Novel Mixed-Precision Quantization Approach for CNNs. *IEEE Access* **2025**, *13*, 49309–49319. [CrossRef]
27. Ramzi, E.; Audebert, N.; Rambour, C.; et al. Optimization of Rank Losses for Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4317–4329. [CrossRef]
28. Shi, H.; Cheng, X.; Mao, W.; et al. P^2-ViT: Power-of-Two Post-Training Quantization and Acceleration for Fully Quantized Vision Transformer. *IEEE Trans. Very Large Scale Integr. Syst.* **2024**, *32*, 1704–1717. [CrossRef]

Copyright © 2026 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.