Plug-and-Play Attention Linearization for Pretrained Transformers

Kenan Kassab[1], Alexey Kashevnik[1],
Ammar Ali[1,2], and Stamatios Lefkimmiatis[1,2]

[1] ITMO University
[2] MWS AI

Abstract. We introduce LATTE (**L**inearized **A**ttention with **T**unable **T**aylor Series **E**xpansion), a training-free attention linearization method that selectively applies linearization to a subset of self-attention blocks in a pretrained vision transformer. LATTE approximates the softmax function via a second-order Taylor expansion centered on tunable anchor points, thereby enabling a hybrid architecture that seamlessly integrates standard self-attention with linear attention. This method preserves model performance without requiring any fine-tuning or retraining but only minimal calibration on the target dataset. LATTE is highly scalable, linearizing more than half of the attention layers retains more than 95% of the original model's performance and reduces attention GFLOPs by up to **15×** (e.g., from 16.384 to 1.082 at T=1000 for the CLIP attention block). We evaluate LATTE across diverse benchmarks, including image classification, object detection, and vision–language multimodal tasks, demonstrating consistent superiority over current state-of-the-art training-free linearization methods. Code will be released here.

1 Introduction

Transformer architecture [34] has reshaped modern machine learning, forming the foundation of state-of-the-art systems across diverse tasks in computer vision [9] [25] [27], natural language processing [8] [33], and multimodal reasoning [29] [35]. Its success stems primarily from the self-attention mechanism, which enables flexible modeling of long-range dependencies through pairwise token interactions. However, this expressiveness comes at a significant computational cost. Standard softmax-based attention exhibits *quadratic* time and memory complexity with respect to sequence length N, $\mathcal{O}(N^2 d)$ for d-dimensional embeddings. This bottleneck severely limits scalability in applications involving high-resolution images, long documents, or real-time inference.

To mitigate this limitation, a significant amount of research effort has been devoted to accelerating attention. Several prior works have focused on kernelized linear attention [6, 20], which approximates the softmax via feature maps, reducing complexity to $\mathcal{O}(Nd^2)$. Alternative directions include selective state space models (e.g., Mamba [11]), sparse attention [5], and hardware-aware optimizations [7]. While effective, many of these approaches either sacrifice modeling capacity, require architectural redesign, or critically depend on full retraining or fine-tuning, which is often impractical in resource-constrained settings.

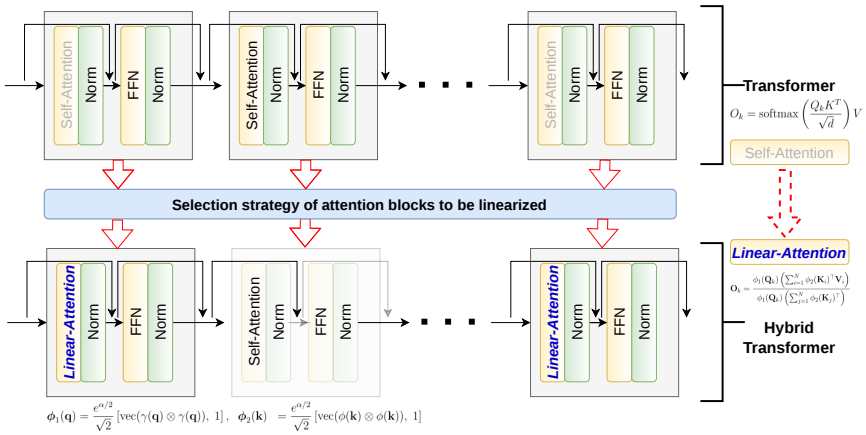

Fig. 1: A subset of the transformer blocks of a given pretrained model are selected for linearization. LATTE is applied on these blocks and their self-attention function is approximated with a second-order Taylor expansion around tunable anchor points. The anchor points are estimated using a small set of calibration data.

In this work, we revisit kernel-based linear attention through the lens of Taylor series expansion approximations, building upon recent insights from QT-ViT [36]. We demonstrate that, with careful design, linear attention can achieve near-perfect fidelity to the original transformer *without any model retraining or fine-tuning*. We define this *training-free* approach as a layer-wise calibration procedure that avoids end-to-end backpropagation. Each attention block is adjusted independently using a task-agnostic reconstruction loss on a small calibration data, while preserving the pretrained model parameters. Our key insight is that the quality of the approximation hinges on the choice of expansion point: rather than using a fixed or heuristic anchor, we introduce *tunable, input-adaptive Taylor expansion points* that are calibrated on a small dataset. This enables accurate local linearization of the softmax while preserving global representational capacity. Our main contributions are:

1. We propose a training-free linearization method that extends second-order Taylor-based attention, by introducing *tunable anchor points* that enable a flexible approximation of self-attention without the need for fine-tuning.
2. We introduce *Joint Post-Squaring Normalization* (JPN), a simple normalization adjustment that replaces QT-ViT's separate query/key post-normalization. JPN jointly normalizes squared query and key vectors using their combined ℓ_2-energy, thereby preserving directional information while inversely modulating attention scores by the joint magnitude of the token pair.
3. We perform extensive experiments across diverse architectures (CLIP, DETR, DINOv2) and tasks (image classification, object detection and multi-modality), where our proposed strategy demonstrates consistent generalization and robustness.

Collectively, we propose LATTE (**L**inearized **A**ttention with **T**unable **T**aylor Series **E**xpansion), which aims to bridge the performance gap between exact and linear attention, offering a practical pathway toward efficient transformer deployment.

2 Related Work

To reduce the quadratic complexity of the original self-attention mechanism, numerous approaches have emerged, focusing on efficient approximations of attention. The first family of techniques restricts the attention scope to limited or sparse locations. These include window-based methods such as Swin Transformer [25], dilated attention [12], and block-sparse attention patterns [5], which reduce computation by limiting token interactions. While effective for vision and long-sequence tasks, these methods sacrifice global context and require careful design of attention windows or sampling strategies.

Another line of research aims to transform attention into a linear form by approximating the softmax function using kernel-based or other decomposable transformations. The central idea is to express the exponential similarity in softmax attention as an inner product of two feature maps. This enables the reordering of matrix multiplications and thereby reduces the complexity from $\mathcal{O}(N^2d)$ to $\mathcal{O}(Nd^2)$. The authors in [20] proposed the Linear Transformer, which replaces the exponential kernel with positive activation functions, enabling streaming computation. Performer [6] introduced the Fast Attention via Orthogonal Random features (FAVOR+) technique, employing orthogonal random features to approximate the softmax kernel. In [28], the authors further studied random-feature approximations for efficient attention. Subsequent work explored more expressive kernels and architectural variants: EfficientViT [3] combines ReLU kernels with grouped convolutions to enhance spatial modeling in vision tasks; Flatten Transformer [14] emphasizes informative tokens through ReLU-based focused attention; and Hydra Attention [1] applies the hydra trick to assign unique feature maps per attention head.

A complementary strategy relies on truncated series expansion to approximate the exponential term in softmax attention. Taylor expansion is among the most widely adopted techniques, enabling linear complexity by decomposing the exponential kernel into polynomial components. MB-TaylorFormer [19] employs first-order Taylor approximations for image restoration, with later variants incorporating norm-preserving mappings to mitigate truncation errors. Taylor-Shift [26] builds upon this by introducing a Taylor-Softmax function to balance accuracy and efficiency. QT-ViT [36] enables acceleration of a second-degree Taylor expansion via approximation methods. Unlike prior linear attention models, QT-ViT retains softmax-level performance without requiring distillation or residual correction, outperforming previous EfficientViT baselines and achieving state-of-the-art accuracy–speed trade-offs.

Recent works also improve attention efficiency through architectural redesigns: SLAB [13] simplifies linear attention with progressive re-parameterized batch normalization, while FasterViT [16] adopts hierarchical attention to acceler-

ate vision transformers while maintaining strong representation quality. Several studies explore the elimination or replacement of softmax with alternative operations. SimA [21] removes softmax entirely, normalizing attention via an ℓ_1-norm for efficient inference. Rectified Linear Attention (ReLA) [39] substitutes softmax with ReLU, leveraging sparsity to accelerate convergence. Hybrid methods combine both linear and quadratic formulations: Agent Attention [15] integrates softmax and linear branches within the same architecture, while Castling-ViT [37] employs dual attention pathways during training but switches to the linear path at inference, thus combining global modeling capacity with computational efficiency.

Despite their computational and memory advantages, linearized and kernel-based approaches face significant challenges. The quality of kernel approximation depends critically on the number and selection of feature mappings, which can lead to instability or degraded performance across tasks. Moreover, proper normalization and integration of positional encodings remain non-trivial under kernelization, often resulting in deviations from the exact softmax formulation. Another major practical limitation is that most of the existing methods require full retraining of the models that feature the modified attention mechanisms, which is an expensive process that hinders the application of such strategies on large-scale pretrained models.

To bridge the gap between self and linear attention in pretrained transformer models, we propose a training-free approach to linearize the softmax attention mechanism. Building upon recent advances in QT-ViT, our method enhances linearization by dynamically estimating Taylor expansion points adapted to the input data distribution. This adaptive strategy improves the stability and accuracy of the approximation while maintaining an optimal balance between computational efficiency and representational fidelity. As a result, our method achieves superior speed–accuracy trade-offs across various Transformer architectures without the need for expensive model retraining or fine-tuning.

3 Method

Before delving into the details of our proposed strategy, we briefly review the QT-ViT method. We then propose a generalization of the quadratic Taylor expansion for softmax attention by introducing a tunable, input-dependent expansion point. We further discuss the role of normalization in stabilizing gradients, describe the transformer block selection criteria for linearization, and finally present LATTE, the proposed training-free transformer linearization approach.

3.1 QT-ViT Overview [36]

QT-ViT improves linear attention in vision transformers by approximating the softmax-based similarity function with a second-order (quadratic) Taylor expansion:

$$\mathrm{Sim}(\mathbf{q},\mathbf{k}) = \exp\left(\frac{\langle \mathbf{q},\mathbf{k}\rangle}{\sqrt{d}}\right) \approx 1 + \frac{\langle \mathbf{q},\mathbf{k}\rangle}{\sqrt{d}} + \frac{\langle \mathbf{q},\mathbf{k}\rangle^2}{2d}. \tag{1}$$

The quadratic term $\langle \mathbf{q},\mathbf{k}\rangle^2$ is expressed via the Kronecker product as $\langle \mathrm{vec}(\mathbf{q}\otimes\mathbf{q}), \mathrm{vec}(\mathbf{k}\otimes\mathbf{k})\rangle$, which enables kernel decomposition but initially leads to $\mathcal{O}(Nd^3)$

complexity. To retain efficiency, the authors propose a compact approximation that preserves only the self-multiplication terms $\{x_i^2\}$ along with optional linear and constant terms, yielding the kernel

$$\phi(\mathbf{x}) = \left[\alpha \cdot \{x_i^2\}_{i=1}^d,\, \beta \cdot \{x_i\}_{i=1}^d,\, \gamma\right]^\mathsf{T} \in \mathbb{R}^{2d+1}. \tag{2}$$

This reduces the attention complexity to $\mathcal{O}(Nd^2)$ while capturing higher-order interactions missing in first-order linear attention. The method does not require knowledge distillation or high-order residuals and achieves state-of-the-art accuracy efficiency trade offs across vision tasks.

3.2 Tunable Taylor Series Expansion

We propose a generalization of the quadratic Taylor expansion for softmax attention by introducing a calibration-tuned, activation-dependent expansion point $\alpha \in \mathbb{R}$, rather than expanding around a fixed point (e.g., zero). This allows the approximation to adapt dynamically to input sequence statistics, thereby improving the flexibility of the linearized attention kernel across inputs.

Let $f(x) = e^x$ denote the exponential function underlying the attention similarity kernel. Its second-order Taylor series expansion around an arbitrary point α is given by:

$$f(x) \approx f(\alpha) + f'(\alpha)(x-\alpha) + \frac{1}{2}f''(\alpha)(x-\alpha)^2. \tag{3}$$

Since $f'(x) = f''(x) = e^x$, it holds: $e^x \approx \frac{e^\alpha}{2}\left[(x+1-\alpha)^2 + 1\right]$.

Applying this approximation to the scaled dot-product $x = \langle \mathbf{q}, \mathbf{k}\rangle / \sqrt{d}$, we obtain the approximated similarity:

$$\mathrm{Sim}(\mathbf{q}, \mathbf{k}) \approx \frac{e^\alpha}{2}\left[\left(\frac{\langle \mathbf{q}, \mathbf{k}\rangle}{\sqrt{d}} + 1 - \alpha\right)^2 + 1\right]. \tag{4}$$

While Eq. (4) can be decomposed into asymmetric query-key mappings, we demonstrate that this construction can be generalized to a **symmetric kernel** through a unified feature mapping. This simplifies implementation, enhances interpretability, and preserves computational efficiency.

Let $m := \sqrt{1-\alpha} \in \mathbb{R}$ (assuming $1 - \alpha \geq 0$). We define a symmetric augmented feature map $h_m(\mathbf{x}) \in \mathbb{R}^{d+1}$ as:

$$h_m(\mathbf{x}) = \begin{bmatrix} \dfrac{\mathbf{x}}{d^{1/4}} \\ m \end{bmatrix}. \tag{5}$$

The inner product between query and key embeddings under this mapping satisfies:

$$\langle h_m(\mathbf{q}), h_m(\mathbf{k})\rangle = \frac{\langle \mathbf{q}, \mathbf{k}\rangle}{\sqrt{d}} + m^2 = \frac{\langle \mathbf{q}, \mathbf{k}\rangle}{\sqrt{d}} + (1-\alpha), \tag{6}$$

which exactly recovers the linear term inside the squared bracket of Eq. (4). Substituting Eq. (6) into the kernel expression, we obtain a symmetric bilinear form:

$$\mathrm{Sim}(\mathbf{q}, \mathbf{k}) = \frac{e^{\alpha}}{2} \left[\langle h_m(\mathbf{q}), h_m(\mathbf{k}) \rangle^2 + 1 \right]. \tag{7}$$

To retain computational efficiency, we must decompose the squared inner product $\langle h_m(\mathbf{q}), h_m(\mathbf{k}) \rangle^2$ into feature maps. Using the Kronecker product identity $\langle \mathbf{a}, \mathbf{b} \rangle^2 = \langle \mathrm{vec}(\mathbf{a} \otimes \mathbf{a}), \mathrm{vec}(\mathbf{b} \otimes \mathbf{b}) \rangle$, we expand the term:

$$h_m(\mathbf{x}) \otimes h_m(\mathbf{x}) = \left[\frac{\mathbf{x} \otimes \mathbf{x}}{d^{1/2}}; \frac{\mathbf{x}}{d^{1/4}} m; \frac{\mathbf{x}}{d^{1/4}} m; m^2 \right] \in \mathbb{R}^{(d+1)^2}. \tag{8}$$

Utilizing the full $(d+1)^2$-dimensional feature space would reintroduce quadratic complexity in d, undermining the efficiency of linear attention. To circumvent this, we employ a structured approximation that retains only the diagonal entries of the quadratic block ($\mathbf{x} \otimes \mathbf{x} \approx \mathbf{x} \odot \mathbf{x}$) and collapses the two identical linear blocks into a single summed term. This aggregation strategy effectively reconstructs the squared binomial expansion for each dimension i. Specifically, it combines the quadratic component $\frac{x_i^2}{\sqrt{d}}$, the linear component $2m\frac{x_i}{d^{1/4}}$, and the constant offset m^2 to yield the per-dimension expression $\left(\frac{x_i}{d^{1/4}} + m\right)^2$. This diagonal Kronecker approximation preserves the kernel's second-order interactions while reducing the feature dimension to $\mathcal{O}(d)$. We thus construct the low-dimensional feature map by explicitly grouping these entries according to their functional role: (a) **Quadratic components**: For each i retain $\frac{x_i^2}{\sqrt{d}}$, (b) **Linear components**: Sum the two cross-terms involving the bias: $2m\frac{x_i}{d^{1/4}}$, (c) **Constant component**: Global offset m^2.

Combining these, the effective per-dimension contribution becomes:

$$\left(\frac{x_i}{d^{1/4}} + m\right)^2 = \frac{x_i^2}{\sqrt{d}} + 2m\frac{x_i}{d^{1/4}} + m^2. \tag{9}$$

Thus, under the diagonal Kronecker approximation with summation-based aggregation, the resulting symmetric feature map for input $\mathbf{x}$ is:

$$\boldsymbol{\psi}_m(\mathbf{x}) = \left[\left(\frac{x_1}{d^{1/4}} + m\right)^2, \ldots, \left(\frac{x_d}{d^{1/4}} + m\right)^2 \right]^{\top}, \tag{10}$$

up to scaling and normalization factors. The attention similarity is then computed as $\mathrm{Sim}(\mathbf{q}, \mathbf{k}) \approx \langle \boldsymbol{\psi}_m(\mathbf{q}), \boldsymbol{\psi}_m(\mathbf{k}) \rangle$ (with appropriate scaling).

Therefore, LATTE's attention mechanism can be interpreted as computing linear attention using features derived from **shifted squared transformations** of the form $(x_i + \tilde{m})^2$, where $\tilde{m} \propto m$ is a learned offset. This generalizes the fixed-bias structure of QT-ViT (where $m = 1$) to a data-adaptive expansion point. The scalar α (via m) is treated as a learnable parameter, allowing the model to adapt the expansion center based on global or local sequence statistics during calibration. This approach enables more accurate modeling of attention dynamics without increasing asymptotic complexity beyond $\mathcal{O}(Nd^2)$.

3.3 Joint Post-Squaring Normalization

Following the quadratic feature mapping, we introduce a *joint post-squaring normalization* that scales both queries and keys by the product of their ℓ_2-norms in the squared space:

$$\tilde{\mathbf{q}} = \frac{\delta_q \cdot \mathbf{q}^2}{h(\mathbf{q}, \mathbf{k})}, \quad \tilde{\mathbf{k}} = \frac{\delta_k \cdot \mathbf{k}^2}{h(\mathbf{q}, \mathbf{k})}, \tag{11}$$

where $h(\mathbf{q}, \mathbf{k}) = \left(\|\mathbf{q}^2\| + \varepsilon\right)\left(\|\mathbf{k}^2\| + \varepsilon\right)$ and $(\cdot)^2$ is applied elementwise. This ensures that the resulting attention scores are not only directionally meaningful (i.e., capturing angular alignment in the squared space) but also *inversely modulated by the joint energy* of the token pair, thereby suppressing responses from uninformative or overly dominant features. Coupled with calibration-tuned scaling factors δ_q, δ_k, this yields a stable, adaptive kernel that aligns with our tunable Taylor expansion center α and consistently improves performance across vision tasks.

To demonstrate this concretely, consider the resulting similarity score:

$$\mathrm{Sim}(\tilde{\mathbf{q}}, \tilde{\mathbf{k}}) = \frac{\delta_q \delta_k}{h(\mathbf{q}, \mathbf{k})^2} \langle \mathbf{q}^2, \mathbf{k}^2 \rangle = \underbrace{\frac{\langle \mathbf{q}^2, \mathbf{k}^2 \rangle}{h(\mathbf{q}, \mathbf{k})}}_{\text{cosine sim.}} \cdot \underbrace{\frac{\delta_q \delta_k}{h(\mathbf{q}, \mathbf{k})}}_{\text{inverse joint-energy modulation}} \tag{12}$$

The first factor captures directional agreement after squaring, while the second one acts as an adaptive gain inversely proportional to the geometric mean of the token energies $\|\mathbf{q}^2\|$ and $\|\mathbf{k}^2\|$. Consequently, even highly aligned token pairs are down-weighted if either exhibits excessive or negligible activation magnitude, mimicking the dynamic range compression inherent in softmax attention.

This mechanism naturally complements our tunable Taylor expansion centered at α (Eq. 4), where the approximation quality depends on the local behavior of e^x around α. By normalizing with respect to the joint energy, the effective input range to the kernel is stabilized, ensuring that the quadratic expansion remains well-conditioned across diverse inputs. The learnable scalars δ_q, δ_k further allow the network to compensate for this damping when beneficial, effectively coupling scale adaptation with the learned expansion center α. Hereafter, for simplicity we will refer to Joint Post-Squaring Normalization as JPN.

3.4 Transformer Block Selection

Inspired by layer-wise importance estimation strategies [10, 30], we adopt a hybrid metric that combines cosine distance and mean squared error (MSE) between intermediate activations of the original and linearized models to quantify the sensitivity of each transformer block to linearization.

Let $\mathcal{F}_\ell(\cdot)$ and $\tilde{\mathcal{F}}_\ell(\cdot)$ denote the ℓ-th transformer block in the original and linearized models, respectively. Given an input batch $\mathbf{X}$, we compute the corresponding hidden activations as $\mathbf{A}_\ell = \mathcal{F}_\ell(\mathbf{X})$ and $\tilde{\mathbf{A}}_\ell = \tilde{\mathcal{F}}_\ell(\mathbf{X})$ where $\mathbf{A}_\ell, \tilde{\mathbf{A}}_\ell \in$

$\mathbb{R}^{B\times N\times d}$, with B the batch size, N the number of tokens, and d the feature dimension. We define the layer-wise importance score $\mathcal{I}_\ell$ as:

$$\mathcal{I}_\ell = \underbrace{1 - \frac{1}{BN}\sum_{i=1}^{BN}\frac{\langle \mathbf{a}_{\ell,i}, \tilde{\mathbf{a}}_{\ell,i}\rangle}{\|\mathbf{a}_{\ell,i}\|_2\,\|\tilde{\mathbf{a}}_{\ell,i}\|_2 + \varepsilon}}_{\text{Mean cosine distance}} + \lambda \cdot \underbrace{\frac{1}{BNd}\|\mathbf{A}_\ell - \tilde{\mathbf{A}}_\ell\|_F^2}_{\text{MSE}}, \tag{13}$$

where $\mathbf{a}_{\ell,i}$ and $\tilde{\mathbf{a}}_{\ell,i}$ are the i-th token embeddings (flattened across batch and sequence dimensions), $\|\cdot\|_F$ denotes the Frobenius norm, $\varepsilon > 0$ ensures numerical stability, and $\lambda \geq 0$ balances the two terms (set to 10 in our experiments).

Intuitively, a higher $\mathcal{I}_\ell$ indicates that layer ℓ is more sensitive to linearization *i.e.*, replacing it degrades the representation significantly and should be preserved in its original form. Conversely, layers with low $\mathcal{I}_\ell$ can be safely linearized with minimal impact.

To select which layers to linearize, we first fully linearize all transformer blocks, then compute $\mathcal{I}_\ell$ for every layer $\ell = 1, \ldots, L$. We sort the layers in ascending order of $\mathcal{I}_\ell$ and retain only the top-k layers with the smallest scores for linearization, while reverting the remaining $(L - k)$ layers to their original softmax-based attention. This yields a hybrid architecture that maximizes efficiency while preserving fidelity in the most critical layers.

3.5 Calibration-Guided Transformer Linearization

To enable efficient deployment of vision transformers without full retraining, we propose LATTE, a lightweight, post-hoc procedure that selectively linearizes attention layers while preserving model accuracy. As shown in Fig. 1, LATTE first replaces all softmax attention blocks with our tunable quadratic approximation enhanced by joint post-squaring normalization (JPN). It then tunes the expansion center α, normalization scalars δ_q, δ_k, and refines the query, key, value, and output projection weights on a small unlabeled calibration set, followed by a sensitivity-aware selection of layers to retain in linearized form. We present the proposed method in an algorithmic schematic in the Appendix to provide a clear, consolidated overview.

In the fully linearized attention, the output is computed via kernelized matrix multiplication that explicitly accounts for both numerator and denominator of the softmax-like normalization. As derived in Sec. 3.2, we employ a symmetric kernel mapping $\boldsymbol{\psi}(\cdot)$ (Eq. (10)) for both queries and keys, rather than distinct mappings. To ensure the denominator $\sum_{j=1}^{N}\mathrm{Sim}(\mathbf{q}_k, \mathbf{k}_j)$ is captured within the linear framework, we augment the feature maps with calibration-tuned scalars, following the padding strategy of EfficientViT [3]. Specifically, we define the extended mappings:

$$\hat{\boldsymbol{\psi}}(\mathbf{q}) = [\boldsymbol{\psi}(\mathbf{q}),\, \beta_q]\,, \quad \hat{\boldsymbol{\psi}}(\mathbf{k}) = [\boldsymbol{\psi}(\mathbf{k}),\, \beta_k]\,, \quad \hat{\mathbf{v}} = [\mathbf{v},\, 1]\,, \tag{14}$$

where $\beta_q, \beta_k \in \mathbb{R}$ are learnable scalars (initialized as $\beta_q = \beta_k = 1$ to preserve consistency with Eq. (10)). The attention output for all queries is then expressed

as:

$$\mathbf{O} = \frac{\hat{\mathbf{Q}}\left(\hat{\mathbf{K}}^\top \hat{\mathbf{V}}\right)}{\hat{\mathbf{Q}}\hat{\mathbf{K}}^\top \mathbf{1}}, \tag{15}$$

where $\hat{\mathbf{Q}} = [\hat{\boldsymbol{\psi}}(\mathbf{q}_1), \ldots, \hat{\boldsymbol{\psi}}(\mathbf{q}_N)]^\top$, $\hat{\mathbf{K}} = [\hat{\boldsymbol{\psi}}(\mathbf{k}_1), \ldots, \hat{\boldsymbol{\psi}}(\mathbf{k}_N)]^\top$, and $\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_N]^\top$, with $\mathbf{1} \in \mathbb{R}^N$ denoting an all-ones vector. The denominator $\hat{\mathbf{Q}}\hat{\mathbf{K}}^\top \mathbf{1}$ corresponds exactly to the sum of kernel similarities $\sum_j \langle \hat{\boldsymbol{\psi}}(\mathbf{q}), \hat{\boldsymbol{\psi}}(\mathbf{k}_j) \rangle$, ensuring proper normalization analogous to softmax attention. This formulation retains $\mathcal{O}(Nd^2)$ complexity and enables stable, end-to-end calibration of the linearized model without full retraining.

3.6 Dynamic Inference Strategy

During experiments, we observed that for certain architectures, fine-tuning the full set of attention projection matrices ($\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{O}$) is unnecessary: the dominant adjustment between original and refined weights can be well-approximated by diagonal scaling. In particular, we find that re-estimating only a lightweight scaling vector $\boldsymbol{\sigma} \in \mathbb{R}^d$ applied to the output projection $\mathbf{O}$ is often sufficient to recover most of the accuracy lost due to linearization.

Motivated by this, we propose a dynamic inference strategy that enables runtime selection between softmax and linear attention based on input sequence length N. Specifically, for high-resolution inputs (large N), where softmax attention incurs $\mathcal{O}(N^2)$ cost, we deploy the linearized variant with the pre-calibrated $\boldsymbol{\sigma}$; for low-resolution inputs (small N), we retain the original softmax attention to preserve fidelity. Although this behavior is not universal across all models (see the Appendix), when applicable, it yields near-identical accuracy to the original transformer model while achieving significant speedups on high-resolution images, making it particularly suitable for practical, resolution-adaptive deployment scenarios.

4 Experiments

We evaluate the efficacy of our proposed method (LATTE) on a range of transformer based vision architectures, including DETR [4], RT-DETR [40], CLIP [29], and SigLip [38] (see the Appendix for additional models). The experiments aim to empirically demonstrate performance improvements over the state-of-the-art QT-ViT method. We analyze the individual and combined contributions of the two key components Tunable Taylor Expansion (TTE) and Joint Post-Squaring Normalization (JPN) w.r.t to model accuracy, inference speed, and computational efficiency.

Our study stands out for its emphasis on training-free linearization, a setting that has received little focus in previous studies. All experiments in this work are carried out in a training-free paradigm. Notably, we are the first to propose and validate an attention linearization strategy specifically tailored for this scenario, whereas earlier approaches (such as QT-ViT, Performer, and EfficientViT) were

Table 1: Performance comparison of LATTE (ours) and QT-ViT across vision transformer architectures. All linearized variants replace 67% of the original attention layers.

Model	Method	Linearized Ratio	Object Detection	
			mAP@0.5	mAP@0.5:0.95
DETR	Self-Attention	0%	62.3	42.0
	QT-ViT (baseline)	67%	52.5	35.9
	LATTE (Ours)	**67%**	**59.2**	**39.9**
RT-DETR	Self-Attention	0%	65.9	49.0
	QT-ViT (baseline)	67%	54.5	40.6
	LATTE (Ours)	**67%**	**62.9**	**46.7**
			Image-Text Retrieval	
			Text Recall	Image Recall
CLIP	Self-Attention	0%	79.4	61.1
	QT-ViT (baseline)	67%	42.0	31.2
	LATTE (Ours)	**67%**	**77.1**	**59.7**
SigLip	Self-Attention	0%	90.0	76.5
	QT-ViT (baseline)	67%	82.3	69.0
	LATTE (Ours)	**67%**	**87.9**	**75.8**

first created and assessed exclusively in training-from-scratch or fine-tuned environments. To ensure fair comparison, we re-implemented all baselines with the same training-free approach, which included identical calibration data, layer selection criteria, and assessment pipelines. This demonstrates our method's generalization capabilities without introducing extra training variables, establishing a new benchmark for post-hoc transformer acceleration. Importantly, this training-free paradigm enables scaled deployment in resource-constrained situations, allowing for efficient inference on edge devices, low-power systems, or institutions lacking access to large-scale computational resources.

Models and Datasets. We primarily evaluate our method on RT-DETR and CLIP, while also reporting comparative results on DETR and SigLip. For multimodal evaluation, we use the COCO Captions dataset and report image-to-text and text-to-image recall metrics for CLIP and SigLip. For object detection, we use COCO [23] and benchmark DETR and RT-DETR using mean average precision at IoU thresholds 0.50 (mAP^{50}) and 0.50–0.95 ($mAP^{50–95}$) as evaluation metrics.

4.1 Main results

Comparison with QT-ViT. We compare our training-free transformer linearization (LATTE) against QT-ViT across multiple vision transformer architectures. As shown in Table 1, LATTE consistently outperforms QT-ViT in terms of accuracy and robustness. The gains are attributed to the enhanced attention representation enabled by the integration of TTE and JPN. We replace 67% of the transformer layers with linearized counterparts in DETR, RT-DETR, CLIP, and SigLip. Object detection experiments are conducted on the COCO dataset, while zero-shot image–text retrieval is evaluated on the MSCOCO Captions dataset.

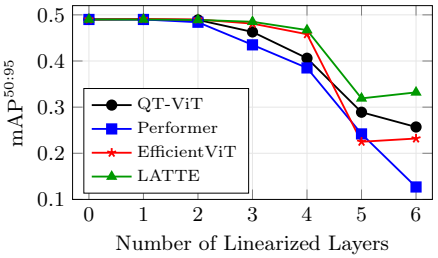

Fig. 2: Comparison of $mAP^{50:95}$ across different linear attention methods as the number of linearized layers increases. Our method maintains significantly higher performance compared to others.

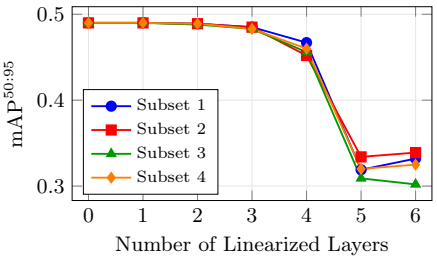

Fig. 3: Performance of the proposed linearized attention method across diverse calibration subsamples from the COCO training set.

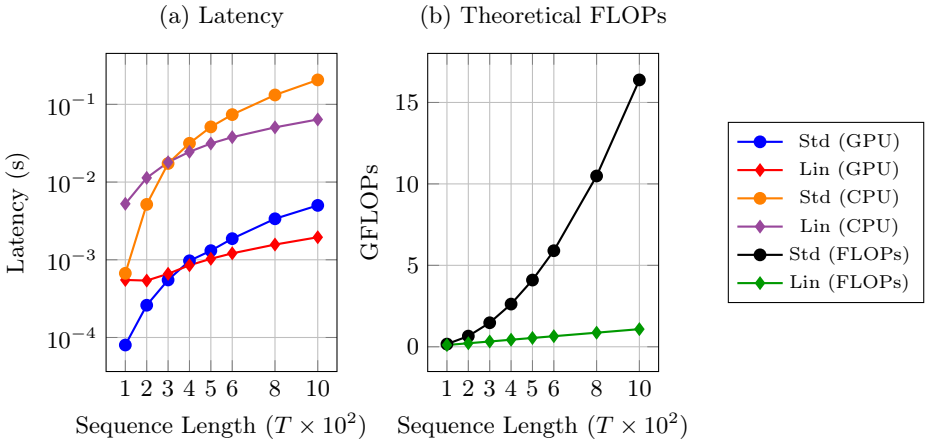

Fig. 4: Performance comparison of standard vs. linearized attention (LATTE). The x-axis shows sequence length T. (a) Latency on GPU and CPU; (b) theoretical FLOPs. Notably, LATTE achieves a **15**× speedup over standard attention at $T = 1000$.

Comparison with alternative linearization methods. We evaluate our method against state-of-the-art linearized attention approaches including QT-ViT, EfficientViT, and Performer under identical experimental conditions on the COCO dataset using RT-DETR. As illustrated in Fig. 2, LATTE consistently achieves higher accuracy across increasing levels of layer linearization. When linearizing the first two layers, all approaches perform similarly; however, beyond the transformer stack's midpoint i.e., when more than half of the layers are linearized a clear divergence emerges. At this point, LATTE demonstrates significantly greater stability and higher retained performance, consistently outperforming competing methods. This highlights the enhanced representational capacity and robustness of our linearization strategy under aggressive architectural simplification.

Latency. We conduct a computational analysis comparing our linearized attention mechanism with standard softmax-based attention in terms of floating-point

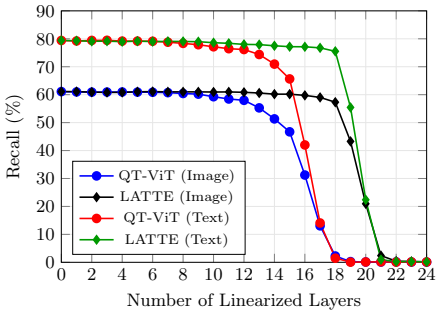

Fig. 5: Impact of progressive layer linearization on multimodal recall. Comparing image and text retrieval between QT-ViT and LATTE.

operations (FLOPs) and inference latency on the CLIP architecture (see the Appendix for details on FLOP calculations). Fig. 4 summarizes the benchmark results for various sequence lengths, including both theoretical FLOP counts and actual latency on CPU and GPU hardware. The results show that the linearized attention mechanism exhibits significantly lower computational complexity, scaling linearly with sequence length T, vs. quadratic for standard attention, yielding growing speedups on long sequences especially on CPU enabling efficient edge deployment.

4.2 Ablation Studies

Effect of calibration data composition. We assess the robustness of our linearized attention method to variations in calibration data composition. To this end, we evaluate multiple subsamples from the COCO training set as calibration datasets and report performance on a fixed evaluation split using the RT-DETR architecture. As illustrated in Fig. 3, performance across all subsampled configurations is consistent, with minimal variation in accuracy. These findings indicate that our strategy is insensitive to subpopulation shifts in calibration data while maintaining robust generalization across diverse data distributions. This insensitivity highlights the reliability of the proposed linearization strategy, demonstrating that careful calibration data selection is not critical but preferable.

Per-layer Linearization Analysis. We present a comparative evaluation of our method against QT-ViT on the MSCOCO Captions benchmark for image–text retrieval. Fig. 5 depicts the performance of each approach when integrated into the CLIP architecture. Notably, our method achieves superior efficiency–accuracy trade-offs: it successfully linearizes 75% of the transformer layers (18 out of 24) while retaining 94% of the original Image Recall and 95% of Text Recall performance. In contrast, QT-ViT attains comparable recall levels (95% Image Recall and 96% Text Recall) but to achieve this is required to linearize only 50% of the attention layers (12 out of 24). This illustrates that LATTE enables more comprehensive architectural simplification while preserving accuracy.

TTE and JPN Interplay. We conduct an ablation study to assess the individual and combined contributions of Joint Post-Squaring Normalization (JPN) and Tunable Taylor Expansion (TTE) to the success of our attention linearization method. Table 2 summarizes the results across several vision transformer

Table 2: Ablation study of (JPN) and (TTE) across vision transformer models. The individual impact of JPN and TTE, as well as their combined effect is reported.

Model	Linearized Ratio	JPN	TTE	Preserved Text Recall	Image Recall
CLIP	0%	✗	✗	100%	79.4	61.1
	67%	✗	✗	51%	42	31.2
	67%	✓	✗	78%	65.1	47.9
	67%	✗	✓	86%	71.5	52.8
	67%	✓	✓	98%	77.1	59.7
SigLip	0%	✗	✗	100%	90.0	76.5
	67%	✗	✗	90%	82.3	69.0
	67%	✓	✗	91%	83.2	69.6
	67%	✗	✓	96%	86.3	73.3
	67%	✓	✓	99%	87.9	75.8

architectures, reporting performance when each component is used alone or in combination. Our findings show that TTE yields robust and significant gains in linearized model performance, underscoring its importance in preserving representational fidelity. In contrast, JPN provides a moderate but consistent improvement, indicating a complementary though less dominant effect. Notably, the best results are achieved when JPN and TTE are combined: this configuration consistently maintains over 95% of the original pretrained model performance while enabling linearization of more than half of the transformer layers. These results demonstrate the synergistic benefits of integrating both enhancements for efficient and accurate vision-language modeling. Additional results across different architectures are provided in the Appendix.

Cross-Dataset and Cross-Task Generalization Analysis. To analyze the impact of dataset characteristics on performance, we evaluate the robustness and generalization of LATTE across a wide range of benchmarks. Table 3 shows the performance of linearized CLIP models on various tasks and domains, including SugarCrepe [18] and CIFAR-10 [22]. Tthe Appendix provides the full results on Crossmodal3600 [31] and MSCOCO for text—image retrieval, as well as EuroSAT [17], and Food101 [2] for image classification. For each dataset, we linearize the maximum number of transformer layers while maintaining at least 95% of the original pretrained model performance. Using the same linearization ratio, we directly compare our approach to QT-ViT under identical settings. Notably, our method consistently surpasses QT-ViT across all tested datasets and tasks. These results highlight LATTE's strong generalization capability, confirming its effectiveness and efficiency across diverse data distributions and downstream applications.

Sensitivity to λ and ε. We further evaluate LATTE's sensitivity to λ and ε. For λ, we perform a grid search on RT-DETR at 67% linearization using $\lambda \in \{0.1, 1, 2, 5, 10, 50, 100\}$. The results are summarized in Table 4. The results show that performance is stable for $\lambda \geq 5$, where the selected layers and final accuracy remain unchanged. This supports our default choice of $\lambda = 10$. We also evaluate $\varepsilon \in \{10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ and observe negligible sensitivity, with

Table 3: Performance of linearized CLIP model across diverse datasets and tasks. Reported values correspond to the maximum number of layers linearized while retaining $\geq 95\%$ of the original pretrained model performance. We denote task-specific performance by P_t and P_i for text and image modalities, respectively. For CIFAR-10, P is classification accuracy (Top-1/Top-5), while for SugarCrepe it is retrieval recall (text/image).

Dataset	Method	Linearized Ratio	P_t	P_i
SugarCrepe	Self-Attention	0%	95.51	96.12
	QT-ViT (baseline)	75%	5.30	10.20
	LATTE (ours)	**75%**	**95.91**	**95.51**
CIFAR-10	Self-Attention	0%	95.60	99.63
	QT-ViT (baseline)	54%	86.88	97.39
	LATTE (ours)	**54%**	**90.56**	**98.45**

mean $\text{mAP}_{50:95}$ of 0.4265 and standard deviation of 0.001291. These results indicate that LATTE is robust to reasonable choices of hyperparameters λ and ε.

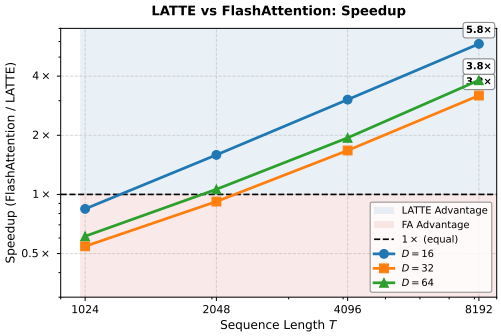

Fig. 6: Latency comparison between FlashAttention (FA) and LATTE. Values above 1× indicate a latency advantage for LATTE.

λ	mAP50:95
0.1	0.44
1	0.452
2	0.452
> 5	0.459

Table 4: Sensitivity analysis of λ on RT-DETR at 67% linearization.

Latency Comparison with FlashAttention. In the main paper, LATTE is primarily compared against standard softmax attention. Here, we additionally compare LATTE with FlashAttention [7] to assess its latency against an optimized attention implementation. Figure 6 reports the speedup ratio between FlashAttention and LATTE across different sequence lengths and feature dimensions. The results show that LATTE becomes increasingly advantageous as the sequence length grows. At long sequences, LATTE achieves clear speedups over FlashAttention across different feature dimensions, reaching up to 5.8×, 3.2×, and 3.8× for the tested settings. This confirms that LATTE remains competitive not only against standard attention, but also against highly optimized attention kernels, especially in long-sequence regimes.

While our primary focus is training-free attention linearization, we also evaluate LATTE in a training-from-scratch setting on CIFAR-10 (ViT [9] model) and COCO (RT-DETR and YOLOv12 [32]) as a proof of concept. The results,

provided in the Appendix, demonstrate that our linearized attention can be effectively trained from scratch, indicating broader applicability beyond post-hoc deployment. Although we focused on vision tasks, LATTE extends to attention-based models beyond vision. Experiments on RoBERTa [24] (see the Appendix) confirm its applicability, particularly for long sequences where linearization reduces inference time.

The Appendix provides additional ablations supporting LATTE's robustness, covering out-of-distribution calibration and calibration overhead on RT-DETR and CLIP.

4.3 Limitations

Our method outperforms existing attention linearization approaches in a training-free setting across a wide range of models and tasks. However, the current strategy for selecting appropriate attention layers to be linearized, as defined by Eq. (13), is suboptimal. Specifically, layer selection is based on local activation differences measured via MSE and cosine similarity between the original and linearized blocks activations. Layers exhibiting minimal output deviation (i.e., low MSE and high cosine similarity) are prioritized for linearization under the assumption that they induce negligible downstream performance degradation. While this heuristic yields strong results for architectures such as CLIP and RT-DETR, it does not generalize equally well to other transformer backbones, such as DINOv2 (see the Appendix). This limitation highlights the need for more holistic, architecture-aware layer selection frameworks potentially incorporating cross-layer interaction modeling or task-specific sensitivity analysis to enable robust and generalizable linearization across diverse transformer designs.

5 Conclusion

We introduced LATTE, a scalable and efficient method for approximating softmax-based self-attention in vision transformers without requiring expensive retraining or fine-tuning. Building on the foundational framework of QT-ViT, we propose two complementary enhancements Joint Post-Squaring Normalization (JPN) and Tunable Taylor Series Expansion Centered on Sequence Dependent Points (TTE) that jointly mitigate the representational degradation inherent in linearized attention mechanisms. These components operate as lightweight, architecture-agnostic post-processing modules, enabling significant performance recovery while preserving the $\mathcal{O}(Nd^2)$ computational complexity of linear attention. Through extensive experiments across diverse benchmarks, we demonstrate that LATTE consistently outperforms state-of-the-art linearization methods, including QT-ViT, while retaining over 95% of the original model's performance even when more than half of the attention layers are linearized. Crucially, our approach achieves this with minimal additional parameters, making it particularly well-suited for deployment in resource-constrained environments.

Acknowledgments This research was supported as part of the ITMO University Research Projects in AI Initiative.

References

1. Bolya, D., Fu, C., Dai, X., Zhang, P., Hoffman, J.: Hydra attention: Efficient attention with many heads. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision - ECCV 2022 Workshops - Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII. Lecture Notes in Computer Science, vol. 13807, pp. 35–49. Springer (2022). https://doi.org/10.1007/978-3-031-25082-8_3, https://doi.org/10.1007/978-3-031-25082-8_3
2. Bossard, L., Guillaumin, M., Gool, L.V.: Food-101 - mining discriminative components with random forests. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 8694, pp. 446–461. Springer (2014). https://doi.org/10.1007/978-3-319-10599-4_29, https://doi.org/10.1007/978-3-319-10599-4_29
3. Cai, H., Gan, C., Han, S.: Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. CoRR **abs/2205.14756** (2022). https://doi.org/10.48550/ARXIV.2205.14756, https://doi.org/10.48550/arXiv.2205.14756
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12346, pp. 213–229. Springer (2020). https://doi.org/10.1007/978-3-030-58452-8_13, https://doi.org/10.1007/978-3-030-58452-8_13
5. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. CoRR **abs/1904.10509** (2019), http://arxiv.org/abs/1904.10509
6. Choromanski, K.M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlós, T., Hawkins, P., Davis, J.Q., Mohiuddin, A., Kaiser, L., Belanger, D.B., Colwell, L.J., Weller, A.: Rethinking attention with performers. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=Ua6zuk0WRH
7. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/N19-1423, https://doi.org/10.18653/v1/n19-1423
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale.

In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy
10. Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., Roberts, D.A.: The unreasonable ineffectiveness of the deeper layers. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=ngmEcEer8a
11. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. CoRR **abs/2312.00752** (2023). https://doi.org/10.48550/ARXIV.2312.00752, https://doi.org/10.48550/arXiv.2312.00752
12. Guo, D., Li, K., Zha, Z., Wang, M.: Dadnet: Dilated-attention-deformable convnet for crowd counting. In: Amsaleg, L., Huet, B., Larson, M.A., Gravier, G., Hung, H., Ngo, C., Ooi, W.T. (eds.) Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019. pp. 1823–1832. ACM (2019). https://doi.org/10.1145/3343031.3350881, https://doi.org/10.1145/3343031.3350881
13. Guo, J., Chen, X., Tang, Y., Wang, Y.: SLAB: efficient transformers with simplified linear attention and progressive re-parameterized batch normalization. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 16802–16812. PMLR / OpenReview.net (2024), https://proceedings.mlr.press/v235/guo24a.html
14. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 5938–5948. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.00548, https://doi.org/10.1109/ICCV51070.2023.00548
15. Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., Song, S., Huang, G.: Agent attention: On the integration of softmax and linear attention. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part L. Lecture Notes in Computer Science, vol. 15108, pp. 124–140. Springer (2024). https://doi.org/10.1007/978-3-031-72973-7_8, https://doi.org/10.1007/978-3-031-72973-7_8
16. Hatamizadeh, A., Heinrich, G., Yin, H., Tao, A., Álvarez, J.M., Kautz, J., Molchanov, P.: Fastervit: Fast vision transformers with hierarchical attention. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=kB4yBiNmXX
17. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. **12**(7), 2217–2226 (2019). https://doi.org/10.1109/JSTARS.2019.2918242, https://doi.org/10.1109/JSTARS.2019.2918242
18. Hsieh, C., Zhang, J., Ma, Z., Kembhavi, A., Krishna, R.: Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023), http://papers.nips.cc/paper_files/

paper/2023/hash/63461de0b4cb760fc498e85b18a7fe81-Abstract-Datasets_and_Benchmarks.html
19. Jin, Z., Qiu, Y., Zhang, K., Li, H., Luo, W.: Mb-taylorformer V2: improved multi-branch linear transformer expanded by taylor formula for image restoration. IEEE Trans. Pattern Anal. Mach. Intell. **47**(7), 5990–6005 (2025). https://doi.org/10.1109/TPAMI.2025.3559891, https://doi.org/10.1109/TPAMI.2025.3559891
20. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 5156–5165. PMLR (2020), http://proceedings.mlr.press/v119/katharopoulos20a.html
21. Koohpayegani, S.A., Pirsiavash, H.: Sima: Simple softmax-free attention for vision transformers. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024. pp. 2595–2605. IEEE (2024). https://doi.org/10.1109/WACV57701.2024.00259, https://doi.org/10.1109/WACV57701.2024.00259
22. Krizhevsky, A.: Learning multiple layers of features from tiny images. University of Toronto (05 2012), https://cave.cs.toronto.edu/kriz/cifar.html
23. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014). https://doi.org/10.1007/978-3-319-10602-1_48, https://doi.org/10.1007/978-3-319-10602-1_48
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9992–10002. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00986, https://doi.org/10.1109/ICCV48922.2021.00986
26. Nauen, T.C., Palacio, S., Dengel, A.: Taylorshift: Shifting the complexity of self-attention from squared to linear (and back) using taylor-softmax. In: Antonacopoulos, A., Chaudhuri, S., Chellappa, R., Liu, C., Bhattacharya, S., Pal, U. (eds.) Pattern Recognition - 27th International Conference, ICPR 2024, Kolkata, India, December 1-5, 2024, Proceedings, Part VI. Lecture Notes in Computer Science, vol. 15306, pp. 1–16. Springer (2024). https://doi.org/10.1007/978-3-031-78172-8_1, https://doi.org/10.1007/978-3-031-78172-8_1
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024), https://openreview.net/forum?id=a68SUt6zFt
28. Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N.A., Kong, L.: Random feature attention. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=QtTKTdVrFBB

29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
30. Shopkhoev, D., Ali, A., Zhussip, M., Malykh, V., Lefkimmiatis, S., Komodakis, N., Zagoruyko, S.: Replaceme: Network simplification via depth pruning and transformer block linearization. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 19487–19517. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/1c10d0c087c14689628124bbc8fa69f6-Paper-Conference.pdf
31. Thapliyal, A.V., Pont-Tuset, J., Chen, X., Soricut, R.: Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. pp. 715–729. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.EMNLP-MAIN.45, https://doi.org/10.18653/v1/2022.emnlp-main.45
32. Tian, Y., Ye, Q., Doermann, D.S.: Yolov12: Attention-centric real-time object detectors. CoRR **abs/2502.12524** (2025). https://doi.org/10.48550/ARXIV.2502.12524, https://doi.org/10.48550/arXiv.2502.12524
33. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. CoRR **abs/2307.09288** (2023). https://doi.org/10.48550/ARXIV.2307.09288, https://doi.org/10.48550/arXiv.2307.09288
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
35. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. CoRR **abs/2409.12191** (2024). https://doi.org/10.48550/ARXIV.2409.12191, https://doi.org/10.48550/arXiv.2409.12191

36. Xu, Y., Li, C., Li, D., Sheng, X., Jiang, F., Tian, L., Barsoum, E.: Qt-vit: Improving linear attention in vit with quadratic taylor expansion. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024 (2024), http://papers.nips.cc/paper_files/paper/2024/hash/971f1e59cd956cc094da4e2f78c6ea7c-Abstract-Conference.html
37. You, H., Xiong, Y., Dai, X., Wu, B., Zhang, P., Fan, H., Vajda, P., Lin, Y.C.: Castling-vit: Compressing self-attention via switching towards linear-angular attention at vision transformer inference. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 14431–14442. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01387, https://doi.org/10.1109/CVPR52729.2023.01387
38. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 11941–11952. IEEE (2023). https://doi.org/10.1109/ICCV51070.2023.01100, https://doi.org/10.1109/ICCV51070.2023.01100
39. Zhang, B., Titov, I., Sennrich, R.: Sparse attention with linear units. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. pp. 6507–6520. Association for Computational Linguistics (2021). https://doi.org/10.18653/V1/2021.EMNLP-MAIN.523, https://doi.org/10.18653/v1/2021.emnlp-main.523
40. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 16965–16974. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01605, https://doi.org/10.1109/CVPR52733.2024.01605