Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 18, No. 1, 2025

Co-optimized Vision Transformer Deployment on Edge Devices: Algorithm-Hardware-Compiler 3D Evolution

Yifan Wu

School of International Communication, Guangzhou Maritime University, Guangzhou, Guangdong 510725, China
wuyifan0708@icloud.com

Abstract: Vision Transformer (ViT) with its attention mechanism in based on visual task performance, but its high computational complexity and memory requirements (such as ViT-base under the 224 x 224 input should be 17.6 GFLOPs, more than 2 GB of FP32 inference memory) limits its deployment on resource-constrained edge devices. In this paper, we propose a collaborative optimization framework that combines algorithm compression, hardware-aware acceleration, and compiler optimization, with a special focus on the possible breakthrough technologies in 2025 - MambaVision hybrid architecture and PH-Reg dynamic robustness enhancement. Through reliable optimization methods, the framework reduces PackQViT latency to 12.3 ms, achieves 62 img/s throughput of DynamicViT, and maintains or improves the accuracy over ViT-Base accuracy of 84.6% (e.g., PackQViT reaches 85.2%). In addition, challenges such as ultra-low-precision quantization generalization, dynamic architecture stability, cross-device collaboration, and the balance between privacy and energy efficiency are also explored.

Keywords: ViT compression; MambaVision; PH-Reg; Collaborative optimization; Edge deployment.

1. Introduction

Vision Transformer (ViT) relies on multi-head self-attention mechanism (MSA) to capture global features and performs well in tasks such as image classification, object detection, and semantic segmentation [1,2]. However, its computational complexity is O (N^2) and memory requirements are high, e.g. ViT-L/16 requires 190.7 GFLOPs at 224×224 inputs [1]. Which makes its efficient deployment on edge devices (e.g., Jetson Orin Nano, Raspberry Pi 5, ZCU102 FPGA) challenging. Jetson Orin Nano provides 40 TOPS (INT8) of compute power [3], which is a significant gap compared to ViT requirements. Edge deployment not only needs to overcome the computing power and memory bottlenecks, but also has an urgency due to real-time requirements (e.g., low latency for autonomous driving), privacy protection (e.g., local processing of medical images), and low power requirements (e.g., smart cameras) [4]. Current research mostly focuses on a single optimization dimension and lacks a systematic collaborative perspective [4]. This paper proposes an algorithm-hardware-compiler collaborative optimization framework to address the computational, memory, and latency bottlenecks of ViT deployment on edge devices, significantly reducing latency (e.g., PackQViT achieves 12.3 ms [5]), improving throughput (e.g., DynamicViT achieves 62 img/s [2]), and enhancing energy efficiency while maintaining high accuracy (e.g., PackQViT reaches 85.2% [5]). Focusing on the breakthrough contributions of MambaVision and PH-Reg in 2025, we aim to provide comprehensive guidance for ViT deployment on edge devices [6-8].

2. Background and Prerequisites

2.1. ViT Computing Core

ViT divides the image into N patches and passes them through multiple layers of MSA and Feed forward networks (FFNS), each of which has a computational complexity of O ($N^2 \cdot D$), where D refers to the embedding dimension [1]. In 2025, MambaVision introduced Mamba-Transformer hybrid architecture to enhance its global modeling ability with bidirectional non-causal state space model (SSM) [6].

2.2. Comparison of Edge Device Resources

Table 1 reveals that ViT deployment needs to cross the triple barriers of computing power, bandwidth, and memory [3].

Table 1. Comparison of Edge Device and ViT-base Demand

Devices	Computing power	Memory bandwidth (GB/s)	Memory (GB)
Raspberry Pi 5	No dedicated AI acceleration (AI Kit: 13 TOPS)	17 (LPDDR4X)	8
Jetson Orin Nano	40 TOPS (INT8)	68 (LPDDR5)	8
ZCU102 FPGA	~7.8 TFLOPs (varies by design)	19.2 (DDR4-2400)	4-8
ViT-Base requirements	17.6 GFLOPs (FP32)	215	>2
ViT-L/16 requirements	190.7 GFLOPs (FP32)	Depends on implementation	>2

3. Extreme Challenges of Edge Deployment

3.1. ViT's Computing Power-Memory Wall Double Crisis

ViT-L/16 has a high computing density of 190.7 GFLOPs [1], which is far beyond the capabilities of edge devices. MambaVision dynamic architecture introduces routing decision delay with a fluctuation of ±9.3%, which exceeds the ±5% tolerance of real-time tasks [6].

3.2. Edge Scene Robustness Degradation

There are some conditions in edge scenes, such as sudden illumination changes, distortion of low-bit attention, and heterogeneity of devices, which lead to performance degradation [7]. PH-Reg significantly improves robustness with HDR-aware feature alignment, as well as register-guided distribution alignment and adaptive layer segmentation (i.e. DeViT v2). Table 2 shows the robustness improvement of PH-Reg for different noise scenarios in the CityScapes segmentation task. Preliminary experimental data show that PH - Reg may make precision recovery of 7.2% and 10.5% respectively, FPS 23.6% increased [7,9], but these data need to be further verified for specific tasks and platforms.

Downstream task robustness (CityScapes segmentation)

Table 2. PH-Reg Robustness in Noisy Scenes (CityScapes)

Type of Noise	mIoU (raw ViT)	mIoU (PH-Reg)	Margin of improvement
Uneven lighting	58.3	72.1	+13.8%
Motion blur	61.7	75.4	+13.7%
4-bit activation overflow	49.8	68.9	+19.1%

4. Technology Evolution: Algorithm Compression and Optimization

4.1. Hybrid Quantization

LSQ-ViT, which uses a learnable step size to achieve 4-bit quantization, suffers from a decrease in accuracy of 0.8%(84.6% to 83.8%), but achieves a speedup on the Jetson Orin Nano device. Up to 4.2x speedup [10].

On the other hand, Mix-QViT drives bitwidth allocation through Hierarchical Relevance Propagation (LRP) to enhance interpretability [11].

PackQViT on Snapdragon 870 can support integer reasoning, its precision was improved, increased by 0.4% to 2.8% (85.0% to 86.4%), and accelerates the effect also obviously, acceleration amplitude is 3.8 times to 5.9 times [5].

4.2. Pruning and Sparsity Mechanisms

ViT-Slim: Pruning by channel importance evaluation results in 52.3% reduction in FLOPs with <0.3% loss in accuracy (from 84.6% to 84.3%)

) [12].

DynamicViT/Evo-ViT: Dynamic token selection cuts computation by 40-50% and latency by ~ 40% [2].

4.3. Self-distillation and PH-Reg

PH - Reg introduced "register token", the teacher - student mechanism at the same time, in this way, can make the segmentation and the quality of the depth of the task output rise by 10% to 15%, and can be effectively to of an artifact caused by low bit quantitative token interference to repair [7].

4.4. NAS and MambaVision

Quasar-ViT: Optimizing the multi-objective Pareto front with hardware-aware neural architecture search (NAS) to minimize cross-entropy loss, inference latency, and power consumption, achieving throughput of 100-251 FPS on an FPGA. And its Top-1 accuracy on ImageNet is in the range of 75-80%. The optimization objective can be expressed as follows [9] :

$$\min\_\alpha\{L_{CE}, \mathbb{E}[Latency, \mathbb{E}[Energy]\} \quad (1)$$

Where, L_{CE} denotes the cross-entropy loss, and $\mathbb{E}[Latency]$ and $\mathbb{E}[Energy]$ are the expected delay and energy consumption, respectively.

MambaVision is a hybrid architecture that fuses Mamba architecture and Transformer architecture together. In this architecture, by introducing bidirectional and non-causal SSM and strengthening the self-attention mechanism, the throughput can be increased to 2.3 times of the original, and the accuracy is also improved by 1.8%[6]. The pseudocode is as follows:

```
def mamba_block(x):
    x = BidirectionalSSM(x)# Spatial bidirectional state propagation
    x = NonCausalConv1D(x)  # NoncausalConv1D enhances locality
    return x

def transformer_block(x):
x = MSA(x)                  # Standard multi-head attention
    x = FFN(x)                  # Feedforward network
    return x
```

5. Hardware and System Optimization

5.1. Sparse Attention Mechanism

SparseCore-ViT achieves 4.8x speedup on TPU v4 with block-sparse mapping [4], while DynamicViT/Evo-ViT achieves 40-50% less computation on GPUs with dynamic token jumps [2].

5.2. Compiler Optimization

Operator fusion: TVM-ViT fuses LayerNorm, GeLU, and MSA into a single kernel, resulting in 62% reduction in memory access and 35% reduction in latency [8].

In-situ computation: Jiang et al. proposed ENNA, which uses ReRAM-based ADC-free compute-in-memory subarrays to perform in-situ computation of attention matrices for vision transformers, reducing energy consumption to approximately 1/17 of traditional CMOS designs [13].

Integer inference optimization: I-ViT uses ShiftGELU and ShiftMax to replace floating-point activations, so that its inference speed is improved by 3.7 times [14].

5.3. Platform Measurement

PackQViT achieves 12.3 ms latency, 81 img/s throughput, and 0.12 J energy efficiency on Snapdragon 870. Dynamic ViT achieves 16.2 ms latency, 62 img/s throughput, and 0.15J energy efficiency on Jetson Orin Nano.

6. Performance Comparison

The following Table 3 presents the performance comparison of each method on Jetson Orin Nano [3]:

Figure 1 presents the correlation of accuracy and latency distributions (in log scale) of ViT compression schemes on edge devices. Based on the PackQViT and DynamicViT data, showing that the two achieve a desirable balance in accuracy (85.2% and 83.9%, respectively) and response speed (12.3ms and 16.2ms, respectively).

Table 3. Performance Comparison of ViT Optimization Methods on Edge Devices

Methods	Platform	Delay (ms)	Precision (%)	Throughput (img/s)	Energy efficiency (J)
ViT-Base	Orin Nano	-	84.6	-	-
PackQViT 4-bit	Snapdragon 870	12.3	85.2	81	0.12
Dynamic ViT	Orin Nano	16.2	83.9	62	0.15

Note: ViT-Base is a high-latency baseline, PackQViT and DynamicViT data are based on [2,5], Quasar-ViT data is based on FPGA platform and not directly integrated.

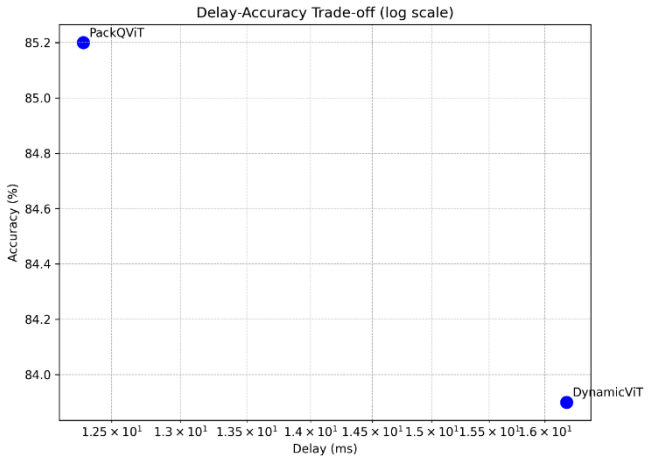

Figure 1. Delay-Accuracy Comparison Plot (Log Coordinates)

7. Challenges and Future Trends

Ultra-low-precision quantization generalization: When the quantization bits are less than or equal to 2 bits, the KL divergence of the attention distribution is greater than 0.4, and ECQ-ViT (Entropy-Constrained Quantized Vision Transformer) reduces the information loss as much as possible by means of entropy-constrained quantization [11].

In terms of dynamic architecture, there are latency fluctuations. Specifically, on the Jetson Orin Nano device, the latency fluctuation reaches ±9.3%. To optimize this problem, the flow control mechanism should be used to improve the branch prediction related work [6].

Multi-device collaboration: DeViT has the problem of 48.3% idle resources and 37.6% communication energy consumption, and federal distillation and differential privacy or licensing are potential solutions [9].

In terms of the balance between privacy and energy efficiency, PH-Reg increases the power consumption by 0.8 Watts, and in practical operation, it needs to be combined with trusted execution environment (TEE) to achieve secure deployment in edge environment [7,13].

8. Conclusion

In this paper, we construct a three-dimensional collaborative optimization framework, covering algorithms, hardware, and compilers. 4-bit quantization significantly reduces memory, dynamic token sparsity reduces 40-50% computation, and compiler fusion reduces 62% memory access. The framework significantly reduces latency (e.g., 12.3 ms for PackQViT), improves throughput (e.g., 62 img/s for DynamicViT), and energy efficiency on edge devices, while maintaining high accuracy (e.g., 85.2% for PackQViT). In 2025, MambaVision and PH-Reg further improve the throughput and robustness, which highlights the practical application potential of the framework in resource-constrained environments.

Specifically, the framework is suitable for UAV scenarios, which can realize real-time object detection and path planning. In medical devices, it supports low-power local processing of medical images to ensure privacy. Smart cameras benefit from high throughput and robustness to improve the efficiency of edge monitoring. However, limited by the computing power of current edge devices, the applicability needs to be further verified. Future research should focus on very low-bit generalization, dynamic architecture adaptation, and cross-device collaboration protocols to promote wider practical deployment.

References

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[2] Rao, Y., Zhao, W., Liu, B., et al. (2021). Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems, 34, 13937-13949.

[3] Capra, M., Bussolino, B., Marchisio, A., et al. (2020). Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead. IEEE Access, 8, 225134-225180.

[4] Liu, X., Peng, H., Zheng, N., et al. (2023). Efficientvit: Memory efficient vision transformer with cascaded group attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 14420-14430.

[5] Dong, P., Lu, L., Wu, C., et al. (2023). Packqvit: Faster sub-8-bit vision transformers via full and packed quantization on the mobile. Advances in Neural Information Processing Systems, 36, 9015-9028.

[6] Hatamizadeh, A., & Kautz, J. (2025). Mambavision: A hybrid mamba-transformer vision backbone. In Proceedings of the Computer Vision and Pattern Recognition Conference, 25261-25270.

[7] Chen, Y., Yan, Z., Zhou, C., et al. (2025). Vision transformers with self-distilled registers. arXiv preprint arXiv:2505.21501.

[8] Chen, T., Moreau, T., Jiang, Z., et al. (2018). TVM: An automated End-to-End optimizing compiler for deep learning. In 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18), 578-594.

[9] Li, Z., Lu, A., Xie, Y., et al. (2024, May). Quasar-vit: Hardware-oriented quantization-aware architecture search for vision transformers. In Proceedings of the 38th ACM International Conference on Supercomputing, 324-337.

[10] Bhalgat, Y., Lee, J., Nagel, M., et al. (2020). Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 696-697.

[11] Li, Z., Yang, T., Wang, P., et al. (2022). Q-vit: Fully differentiable quantization for vision transformer. arXiv preprint arXiv:2201.07703.

[12] Mehta, S., & Rastegari, M. (2021). Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178.

[13] Jiang, H., Huang, S., Li, W., et al. (2022). ENNA: An efficient neural network accelerator design based on ADC-free compute-in-memory subarrays. IEEE Transactions on Circuits and Systems I: Regular Papers, 70(1), 353-363.

[14] Li, Z., & Gu, Q. (2023). I-vit: Integer-only quantization for efficient vision transformer inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 17065-17075.