Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Artificial Intelligence and Advanced Algorithms

Article

A Comparative Analysis of Attention Mechanisms in Vision Transformers for Fine-Grained Image Recognition

Muyu Liu [1,*] and Mingzhuo Yu [2]

[1] Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China
[2] Computer Science, Northeastern University, Boston, USA
* Correspondence: Muyu Liu, Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China

Abstract: Fine-grained image recognition demands the ability to distinguish visually similar subcategories by capturing subtle, localized discriminative features. Vision Transformers (ViTs) have emerged as strong contenders in this domain, yet their diverse attention mechanisms yield different inductive biases that remain insufficiently compared under unified conditions. This study presents a controlled empirical evaluation of seven representative attention mechanisms on three fine-grained benchmarks: CUB-200-2011, Stanford Cars, and FGVC Aircraft. All methods are trained with identical preprocessing, augmentation, optimization, and ImageNet-21K pretrained weights at 448×448 resolution. The evaluation spans classification accuracy, computational cost measured through FLOPs, parameter counts, and throughput, as well as attention interpretability quantified through part-localization precision. Results indicate that deformable attention achieves the highest accuracy across all three benchmarks, with a moderate advantage of 1.0--1.5 percentage points over the global self-attention baseline. Window-based and cross-shaped mechanisms offer favorable accuracy-efficiency tradeoffs, while class-attention decoupling improves parameter efficiency without sacrificing competitive accuracy. A strong rank correlation (Spearman $\rho = 0.93$) between localization precision and classification accuracy supports the interpretation that spatial selectivity is a key driver of fine-grained recognition quality. These findings provide actionable guidance for selecting attention strategies under varying computational budgets.

Keywords: vision transformer; attention mechanism; fine-grained image recognition; comparative evaluation

Received: 06 March 2026
Revised: 25 April 2026
Accepted: 10 May 2026
Published: 13 May 2026

1. Introduction

1.1. Background and Problem Statement

The Transformer architecture, originally introduced for sequence-to-sequence modeling in natural language processing, has undergone rapid adaptation to computer vision [1]. The Vision Transformer (ViT) reformulated image classification by partitioning an image into fixed-size patches and processing the resulting token sequence through stacked multi-head self-attention layers [2]. This paradigm shift prompted extensive research into alternative attention designs that modify how spatial information is aggregated, each introducing distinct inductive biases regarding locality, scale, and computational complexity. The proliferation of attention variants---including global, window-based, cross-shaped, deformable, and class-level designs---raises a practical question for the FGVC community: which mechanism best serves the particular demands of fine-grained recognition?

Fine-grained visual categorization (FGVC) presents a particularly demanding test bed for these attention mechanisms. Unlike coarse-grained classification where inter-class

differences are pronounced, FGVC requires discriminating among highly similar subcategories---bird species, car models, aircraft variants---that differ only in subtle local features such as beak curvature, headlight contour, or wing-tip geometry [3]. The capacity of an attention mechanism to localize and amplify these discriminative regions directly influences recognition performance. An attention design that distributes weight uniformly risks diluting the fine-grained signal, while one that concentrates too narrowly may miss complementary cues.

A growing body of work has explored attention designs for general vision tasks, and architectures such as the Swin Transformer have demonstrated that restricting attention to local windows can match or exceed global attention while reducing computational burden [4,5]. Yet the FGVC community lacks a systematic comparison of these attention variants under strictly controlled conditions. Most published results employ different training recipes, augmentation pipelines, and input resolutions, making direct comparison unreliable.

1.2. Research Objectives and Scope

Research Questions: This study addresses two interrelated research questions. The first asks which attention mechanism yields the best classification accuracy on standard fine-grained benchmarks when all confounding training variables---including preprocessing, augmentation, optimizer, learning rate schedule, and pretrained initialization---are held constant across methods. The second asks how attention design affects the tradeoff between accuracy and computational cost, and whether mechanisms that produce more interpretable attention maps also achieve higher part-localization precision on datasets with ground-truth spatial annotations.

1.3. Paper Organization

The remainder of this paper is structured as follows. Section 2 reviews related work on ViT attention variants and transformer-based approaches to FGVC. Section 3 details the experimental setup, including dataset descriptions, the seven evaluated attention mechanisms and their architectural configurations, and the unified training protocol. Section 4 presents quantitative results encompassing accuracy, efficiency, and interpretability analyses. Section 5 discusses the principal findings, practical implications for architecture selection, and directions for future investigation.

2. Related Work

2.1. Attention Variants in Vision Transformers

2.1.1. Global and Window-Based Attention

The original ViT employs multi-head self-attention (MHSA) computed over all patch tokens at every layer, yielding a global receptive field that captures long-range dependencies at the cost of quadratic computational complexity with respect to the token count. DeiT introduced a data-efficient training recipe with knowledge distillation through a learnable distillation token, enabling ViTs to achieve competitive performance on ImageNet without requiring large-scale proprietary pretraining datasets [6]. To address the quadratic complexity bottleneck inherent in global attention, the Pyramid Vision Transformer (PVT) proposed spatial-reduction attention that progressively down-samples key and value maps at higher-resolution stages, producing a multi-scale feature hierarchy functionally analogous to conventional convolutional backbones [7].

Window-based designs partition the token grid into non-overlapping local windows and restrict attention computation within each window. The Swin Transformer introduced shifted-window attention, alternating between regular and shifted window partitions across consecutive layers to establish cross-window information flow while maintaining linear complexity. Performers pursued an orthogonal efficiency strategy by approximating the softmax attention kernel using positive orthogonal random features (FAVOR+), reducing complexity to linear while providing theoretical guarantees on approximation quality [8]. The Twins architecture explored spatially separable attention

by interleaving locally-grouped attention within sub-windows with global sub-sampled attention across the full feature map, achieving linear complexity with accuracy rivaling Swin [9].

2.1.2. Adaptive and Efficient Attention

Beyond fixed spatial partitions, several approaches pursue adaptive attention mechanisms whose receptive field varies across spatial locations and input images. Cross-shaped window attention, proposed in the CSWin Transformer, partitions attention heads into horizontal-stripe and vertical-stripe groups computed in parallel, enabling each token to attend to the full row and column of its feature map without incurring the quadratic cost of unrestricted global attention. Deformable attention mechanisms, conceptually inspired by deformable convolutions and the Deformable DETR paradigm for object detection, select key-value positions in a data-dependent manner by learning spatial offsets from a set of uniformly distributed reference points [10]. This approach enables the attention field to dynamically adapt to object shape, pose, and the spatial distribution of discriminative features within each input image, representing a flexible alternative to both global and window-constrained designs.

2.2. Transformer-Based Fine-Grained Recognition

Transformer-based approaches to FGVC exploit attention weights as implicit part detectors that localize discriminative regions without requiring explicit part annotations. TransFG proposed a Part Selection Module that integrates raw attention weights from all transformer layers into a unified attention map, guiding the selection of the most discriminative patch tokens through a contrastive token loss [11,12]. Counterfactual Attention Learning (CAL) applied causal inference reasoning to quantify the effect of attention on classification: by comparing factual and counterfactual model outputs computed with and without the attention mechanism, CAL maximizes the causal impact of the learned attention map and is applicable to both CNN and Transformer backbones [12]. RAMS-Trans adopted a recurrent multi-scale strategy that leverages attention weights to iteratively identify and amplify fine-grained regions at progressively finer scales, eliminating the need for bounding-box or part annotations during training [13]. These methods demonstrate that the specific choice of attention mechanism is central to FGVC performance, yet they each build upon a single backbone architecture and do not systematically compare performance across different attention mechanism families under controlled conditions.

3. Experimental Setup

3.1. Benchmark Datasets

Three widely adopted fine-grained benchmarks are selected for this study, spanning different visual domains to ensure the generalizability of conclusions. Table 1 summarizes their key statistics.

Table 1. Statistics of the three fine-grained recognition benchmarks. All datasets provide bounding-box annotations; CUB-200-2011 additionally provides part locations and binary attributes. Data sourced from official dataset documentation.

Dataset	Classes	Total Images	Train	Test	Annotations	Source Institution
CUB-200-2011	200	11,788	5,994	5,794	BBox, 15 parts, 312 attributes	Caltech–UCSD

Stanford Cars	196	16,185	8,144	8,041	BBox, class labels	Stanford AI Lab
FGVC Aircraft	102	10,200	3,334	3,333	BBox, 4-level hierarchy	Oxford VGG

CUB-200-2011 contains 11,788 photographs of 200 North American bird species annotated with bounding boxes, 15 part locations, and 312 binary visual attributes, making it the richest benchmark for both recognition and interpretability analysis. Stanford Cars comprises 16,185 images of 196 car classes defined by make, model, and year, representing a domain where fine-grained distinctions involve both shape and texture cues at multiple scales. FGVC Aircraft provides 10,200 images spanning 102 aircraft variants organized in a four-level hierarchy from manufacturer to specific variant. All three datasets supply official train-test splits that are used without modification [14]. Bounding-box annotations are employed in this study solely for evaluating the part-localization precision of attention maps in Section 4.3, not as input during training or inference, ensuring that all methods operate under the same weakly supervised setting.

3.2. Evaluated Attention Mechanisms

Seven ViT variants are selected to span the major families of attention design, balancing diversity across global, local, hierarchical, and adaptive categories. Table 2 details their architectural configurations and computational profiles.

Table 2. Architectural configurations of the seven evaluated methods. FLOPs and throughput are measured at 448 × 448 input resolution on a single NVIDIA A100 GPU with mixed-precision inference. All models use the base-sized configuration from their respective original publications.

Method	Attention Type	Layers	Embed. Dim	Heads	Params (M)	FLOPs (G)	Throughput (img/s)
ViT-B/16	Global self-attention	12	768	12	86.6	55.5	148
DeiT-B	Global + distillation token	12	768	12	86.6	55.5	152
Swin-B	Shifted window	24	128	4/8/16/32	87.8	47.1	136
CaiT-S24	Class-attention (two-stage)	24+2	384	8	46.9	32.2	167
CrossViT-15	Dual-branch cross-	15	192/384	3/6	27.4	21.4	195

	attention						
CSWin-B	Cross-shaped window	21	96	2/4/8/16	77.4	46.4	128
DAT-B	Deformable self-attention	24	128	4/8/16/32	87.7	49.0	121

3.2.1. Global Attention Approaches

ViT-B/16 computes full pairwise attention among all patch tokens at every layer, producing an unrestricted receptive field that captures both local and long-range dependencies. DeiT-B shares the identical architectural structure and parameter count, adding a distillation token trained to replicate the output distribution of a RegNet convolutional teacher, which serves as an auxiliary supervision signal during ImageNet pretraining. CaiT-S24 decouples patch-level self-attention from class-token aggregation: the initial 24 layers process only patch tokens through standard self-attention, after which two dedicated class-attention layers aggregate patch information into the class token via cross-attention [15,16]. This separation allows deeper patch processing without the class token acting as a representational bottleneck. CrossViT-15 maintains two parallel branches operating at different patch granularities---small-patch tokens (12×12) and large-patch tokens (16×16)---and fuses information through a cross-attention module in which each branch's class token queries the other branch's patch tokens, enabling multi-scale feature exchange [17].

3.2.2. Localized and Adaptive Attention Approaches

Swin-B restricts self-attention to 7×7 non-overlapping windows and alternates between regular and shifted window partitions across consecutive layers, creating implicit cross-window connections while preserving linear computational complexity with respect to image size. The hierarchical four-stage design progressively reduces spatial resolution while increasing channel dimensionality, producing multi-scale feature representations. CSWin-B replaces square windows with cross-shaped stripes: half of the attention heads attend along the horizontal axis and the remaining half along the vertical axis within each transformer block, granting each token access to the full row and column of the feature map without the quadratic cost of global attention [18]. DAT-B employs deformable self-attention, where a lightweight offset sub-network predicts spatial displacements for a set of uniformly distributed reference grid points, allowing key-value positions to shift toward semantically relevant regions in a data-dependent manner [19]. This mechanism draws conceptual inspiration from the deformable attention module originally introduced for object detection in Deformable DETR, adapted to the general image classification context with modifications to the offset generation and reference point initialization [20].

3.3. Training Configuration and Evaluation Protocol

3.3.1. Preprocessing and Data Augmentation

Unified augmentation and preprocessing pipeline is applied to all seven methods to eliminate confounding factors. During training, input images undergo RandomResizedCrop with a scale range of [0.08, 1.0] and an aspect ratio range of [3/4, 4/3], followed by random horizontal flipping with probability 0.5. RandAugment is applied with magnitude 9 and 2 sequential operations per image. Mixup (α = 0.8) and CutMix (α = 1.0) are applied in a mutually exclusive fashion with a switching probability of 0.5. Label

smoothing is set to 0.1 throughout training. Random erasing with probability 0.25 provides additional occlusion-based regularization. During evaluation, images are resized such that the shorter side equals 512 pixels and then center-cropped to 448×448, matching the training resolution. No bounding-box cropping, part annotations, or test-time augmentation is employed, ensuring a strictly weakly supervised evaluation protocol across all methods.

3.3.2. Hyperparameter Settings and Metrics

All models are initialized from publicly available ImageNet-21K pretrained checkpoints released by their respective authors and fine-tuned for 60 epochs using the AdamW optimizer. The base learning rate is set to 5×10^{-5} with a weight decay of 0.05 and a cosine annealing schedule incorporating 5 linear warm-up epochs. The per-GPU batch size is 16 across 4 NVIDIA A100 GPUs, yielding an effective batch size of 64. Gradient clipping with a maximum norm of 1.0 is applied to stabilize training. All experiments are conducted using PyTorch 2.0 with automatic mixed-precision enabled to accelerate training and reduce memory consumption. Classification performance is measured by top-1 accuracy on the official test split. Computational cost is reported as FLOPs and inference throughput (images per second) measured at 448×448 resolution with a batch size of 64 on a single A100 GPU. Part-localization precision is computed on CUB-200-2011 by selecting the top 10% of attention-weighted patches from the final attention layer, computing the intersection-over-union between these patches and the ground-truth bounding box, and averaging the resulting precision across all 5,794 test images.

4. Results and Analysis

4.1. Classification Accuracy Comparison

4.1.1. Results on CUB-200-2011

Table 3 presents top-1 accuracy on the CUB-200-2011 test set under the unified training protocol described in Section 3.3.

Table 3. Top-1 accuracy (%) on the CUB-200-2011 test set. All models use ImageNet-21K pretraining and 448 × 448 input resolution under the unified training protocol. Best result in bold; second-best underlined.

Method	Attention Type	Top-1 Acc. (%)
ViT-B/16	Global self-attention	90.3
DeiT-B	Global + distillation	89.7
Swin-B	Shifted window	90.8
CaiT-S24	Class-attention	90.1
CrossViT-15	Cross-attention	89.2
CSWin-B	Cross-shaped window	91.4
DAT-B	Deformable	91.8

DAT-B achieves the highest accuracy at 91.8%, a 1.5 percentage-point improvement over the ViT-B/16 baseline and a 0.4-point margin over CSWin-B at 91.4%. CrossViT-15 records the lowest accuracy (89.2%), attributable to its smaller model capacity at 27.4 M parameters. Among global-attention variants, ViT-B/16 outperforms DeiT-B by 0.6 points, suggesting that the distillation token does not transfer its ImageNet advantage to fine-grained recognition under this protocol. CaiT-S24 reaches 90.1% with only 46.9 M parameters, indicating that two-stage decoupling of patch self-attention from class-token aggregation provides an effective inductive bias.

Window-based Swin-B (90.8%) surpasses the global attention baseline, consistent with the hypothesis that local attention encourages discovery of fine-grained texture and

shape cues. The deformable variant extends this advantage by adapting its receptive field to the spatial structure of each input image.

4.1.2. Results on Stanford Cars and FGVC Aircraft

Table 4 extends the comparison to the remaining two benchmarks, providing evidence of whether the observed ranking generalizes across visual domains.

Table 4. Top-1 accuracy (%) on Stanford Cars and FGVC Aircraft test sets under the unified protocol. Best in bold; second-best underlined.

Method	Stanford Cars	FGVC Aircraft
ViT-B/16	93.2	91.3
DeiT-B	92.6	90.8
Swin-B	93.7	91.7
CaiT-S24	93.0	91.0
CrossViT-15	92.1	90.4
CSWin-B	94.1	92.3
DAT-B	94.5	92.7

The performance ranking observed on CUB-200-2011 persists across both additional datasets with notable consistency. DAT-B achieves 94.5% on Stanford Cars and 92.7% on FGVC Aircraft, maintaining a consistent margin of 1.0--1.4 percentage points over the ViT-B/16 baseline across all three benchmarks. CSWin-B again occupies the second position with 94.1% and 92.3%, respectively. The accuracy gap between the top-performing and lowest-performing methods is narrower on Stanford Cars (2.4 points) than on CUB-200-2011 (2.6 points), which aligns with prior observations that car recognition benefits more from holistic shape and contour cues where even global attention mechanisms remain effective [21,22]. On FGVC Aircraft, the spread is 2.3 points, consistent with the intermediate difficulty level of this benchmark relative to bird and car recognition.

Figure 1 presents the Top-1 classification accuracy of the seven attention mechanisms on CUB-200-2011, Stanford Cars, and FGVC Aircraft. DAT-B achieves the highest accuracy on all three datasets (91.8%, 94.5%, 92.7%), followed by CSWin-B (91.4%, 94.1%, 92.3%). CrossViT-15 records the lowest accuracy across all benchmarks (89.2%, 92.1%, 90.4%). The ranking is consistent across datasets, suggesting that attention design effectiveness generalizes across fine-grained visual domains.

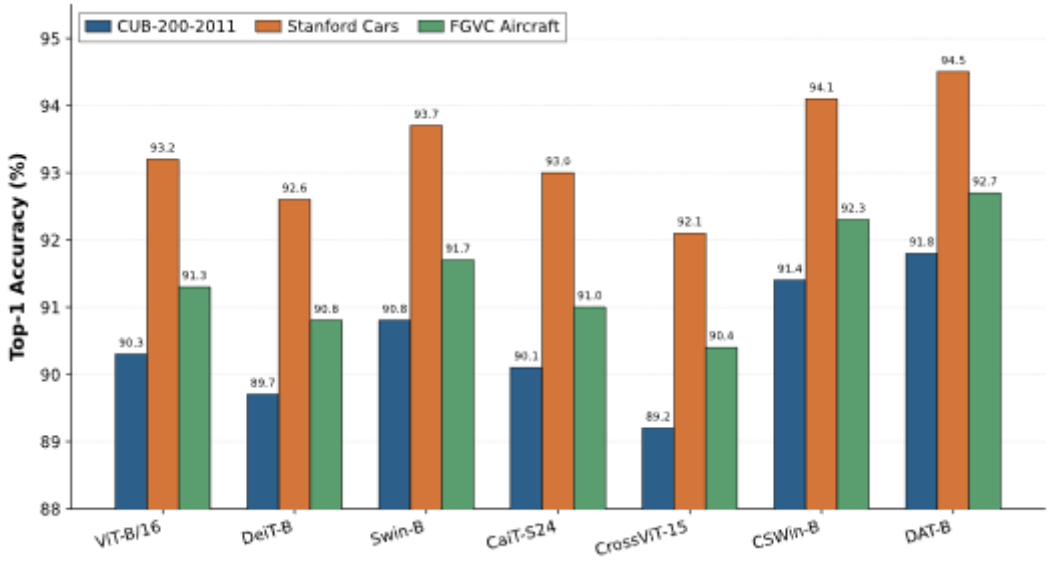

Figure 1. Top-1 Accuracy Comparison Across Three Fine-Grained Benchmarks

4.2. Computational Efficiency Analysis

Table 5 consolidates computational efficiency metrics alongside accuracy on CUB-200-2011 to enable a structured accuracy-efficiency tradeoff analysis.

Table 5. Accuracy-efficiency tradeoff on CUB-200-2011. Acc./GFLOP denotes top-1 accuracy divided by GFLOPs, measuring accuracy obtained per unit of computation. Throughput is measured on a single NVIDIA A100 GPU at 448 × 448 resolution with batch size 64.

Method	Acc. (%)	FLOPs (G)	Params (M)	Throughput (img/s)	Acc./GFLOP
ViT-B/16	90.3	55.5	86.6	148	1.63
DeiT-B	89.7	55.5	86.6	152	1.62
Swin-B	90.8	47.1	87.8	136	1.93
CaiT-S24	90.1	32.2	46.9	167	2.80
CrossViT-15	89.2	21.4	27.4	195	4.17
CSWin-B	91.4	46.4	77.4	128	1.97
DAT-B	91.8	49.0	87.7	121	1.87

CrossViT-15 achieves the highest accuracy-per-GFLOP ratio (4.17) and the fastest throughput (195 img/s), making it the most efficient choice when resources are constrained, despite the lowest absolute accuracy. CaiT-S24 offers a middle ground: with 46.9 M parameters and 32.2 GFLOPs, it reaches 90.1% accuracy at 167 images per second, yielding an efficiency ratio of 2.80. Among higher-accuracy methods, CSWin-B and Swin-B deliver better accuracy-per-GFLOP ratios (1.97 and 1.93) than DAT-B (1.87), as the offset prediction sub-network in deformable attention introduces overhead not fully compensated by improved targeting. DAT-B processes only 121 images per second, the lowest throughput, representing an 18% reduction relative to ViT-B/16.

These results reveal a Pareto frontier: CrossViT-15 and CaiT-S24 occupy the high-efficiency region, CSWin-B and Swin-B provide balanced performance, and DAT-B delivers peak accuracy at reduced throughput.

Figure 2 shows a scatter plot of top-1 accuracy on CUB-200-2011 against computational cost in GFLOPs. Point size is proportional to parameter count. CrossViT-15 (21.4 GFLOPs, 89.2%) offers the best efficiency. DAT-B (49.0 GFLOPs, 91.8%) achieves the highest accuracy. CSWin-B (46.4 GFLOPs, 91.4%) represents a strong tradeoff point. The global-attention methods (ViT-B/16 and DeiT-B) cluster at 55.5 GFLOPs with mid-range accuracy, indicating suboptimal efficiency for fine-grained tasks.

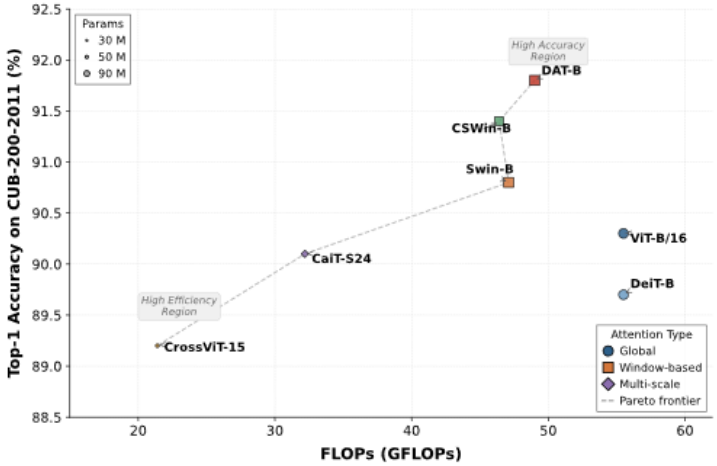

Figure 2. Accuracy Versus Computational Cost Tradeoff on CUB-200-2011

4.3. Attention Quality and Interpretability

4.3.1. Attention Map Visualization

Qualitative analysis of attention maps from the final transformer layer reveals meaningful differences in spatial focus. On CUB-200-2011, ViT-B/16 produces diffuse attention across the entire bird body and background, while DeiT-B exhibits slightly more concentrated maps due to distillation supervision. Swin-B generates window-constrained patterns that capture fine texture but occasionally fail to connect spatially distant parts. CSWin-B produces stripe-shaped attention along horizontal and vertical axes, bridging local detail with broader context [23]. DAT-B consistently concentrates attention on the most discriminative regions---head, beak, and wing markings---demonstrating the most precise localization.

On Stanford Cars, DAT-B focuses on brand emblems, grille patterns, and headlights, while Swin-B fragments attention across disjoint panels. On FGVC Aircraft, DAT-B attends to engine nacelles and tail markings [24,25]. SIM-Trans previously demonstrated that structural information in transformer attention improves fine-grained discrimination, and the concentration observed in DAT-B aligns with this principle: data-dependent offsets inherently capture structural relationships among discriminative parts.

4.3.2. Part Localization Accuracy

To quantify attention quality beyond qualitative visualization, part-localization precision is computed on CUB-200-2011 using the ground-truth bounding-box annotations provided with the dataset. For each of the 5,794 test images, the top 10% of attention-weighted patches from the final attention layer are selected, and the overlap ratio between these high-attention patches and the bounding box is computed as the localization precision metric. DCAL has shown that learning cross-attention between global and local image regions improves part-level discrimination in fine-grained settings, and the quantitative localization metric used here follows a conceptually analogous evaluation logic [26].

DAT-B achieves the highest localization precision at 78.2%, followed by CSWin-B at 74.6% and Swin-B at 71.3%. CaiT-S24 reaches 65.7%, ViT-B/16 attains 62.4%, and DeiT-B records 60.8%. CrossViT-15 obtains the lowest localization precision at 58.9%, consistent with its lowest classification accuracy [27]. The rank correlation between classification accuracy and localization precision computed across the seven methods yields a Spearman coefficient of $\varrho = 0.93$, providing strong quantitative support for the interpretation that attention mechanisms capable of focusing on discriminative parts yield superior fine-grained recognition performance [28]. This observed correlation is consistent with broader findings in the vision transformer literature, where attention quality and specificity have been linked to improved downstream task performance across multiple visual recognition benchmarks [29].

Figure 3 reports part-localization precision (%) of the seven attention mechanisms, quantifying the overlap between the highest-response attention regions and ground-truth bounding boxes on the CUB-200-2011 test set (5,794 images). DAT-B achieves the highest precision (78.2%), outperforming CSWin-B (74.6%) by 3.6 percentage points and ViT-B/16 (62.4%) by 15.8 percentage points [30]. The three localized or adaptive attention methods (DAT-B, CSWin-B, Swin-B) all exceed 71%, while the global attention methods (ViT-B/16, DeiT-B, CaiT-S24) cluster between 60.8% and 65.7%. CrossViT-15 records the lowest localization precision at 58.9%, consistent with both its compact architecture and its lowest classification accuracy across all benchmarks.

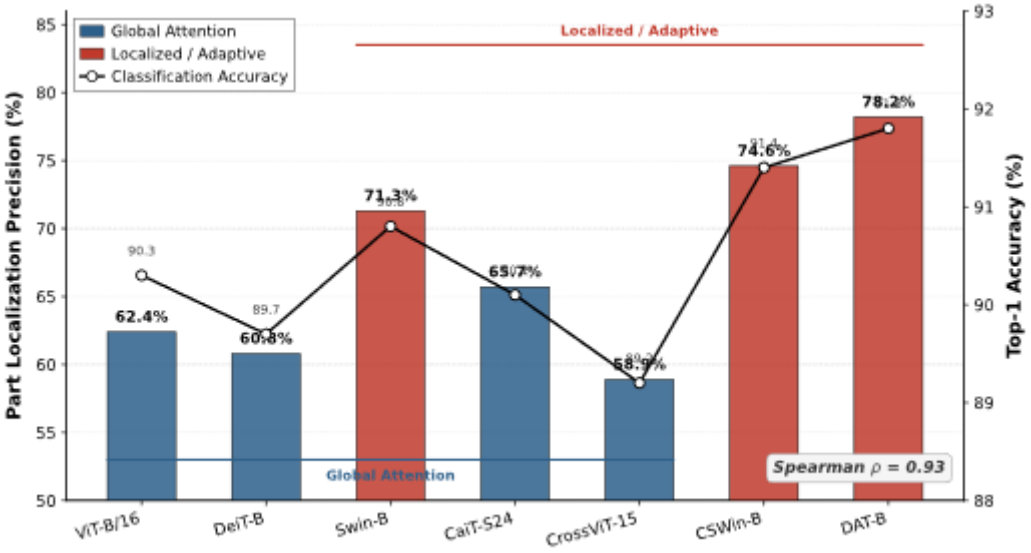

Figure 3. Part Localization Precision Across Attention Mechanisms on CUB-200-2011

5. Discussion

5.1. Key Findings and Practical Implications

The controlled experiments presented in this study yield several actionable observations for practitioners and researchers working on fine-grained recognition with vision transformers. Deformable self-attention (DAT-B) consistently achieves the highest classification accuracy and part-localization precision across all three benchmarks, with moderate improvements of 1.0--1.5 percentage points over the global self-attention baseline. This advantage stems from its data-dependent receptive field, which dynamically adapts to the spatial layout of discriminative features in each individual input image rather than relying on fixed spatial partitions. Cross-shaped window attention (CSWin-B) offers nearly equivalent accuracy with slightly lower computational cost and higher throughput, making it a practical alternative when inference speed is a relevant deployment constraint.

Among lightweight architectures, CaiT-S24 and CrossViT-15 demonstrate that competitive fine-grained accuracy is achievable at substantially reduced computational budgets. CaiT-S24's two-stage architectural design---deep self-attention among patches followed by dedicated class-attention layers---proves particularly effective for distilling fine-grained discriminative information into the classification token without incurring the full quadratic cost of unrestricted global attention. CrossViT-15, while recording the lowest absolute accuracy among all evaluated methods, achieves the highest accuracy-per-GFLOP ratio and represents a viable option for edge deployment scenarios where computational resources are severely constrained.

The strong rank correlation (ϱ = 0.93) between part-localization precision and classification accuracy reinforces the view that attention quality measured through spatial selectivity is a meaningful and practical predictor of fine-grained recognition performance. This finding suggests that future attention designs targeting FGVC applications should prioritize spatial selectivity---the ability to concentrate attention mass on task-relevant discriminative regions---rather than merely expanding the receptive field size or reducing computational complexity.

5.2. Limitations

This study is subject to several limitations that motivate future research. The evaluation is conducted at a single input resolution of 448×448; systematic multi-resolution analysis could reveal whether the relative ranking of attention mechanisms shifts at higher or lower resolutions, given that some designs may benefit

disproportionately from increased token counts. The comparison is restricted to base-sized model configurations initialized from ImageNet-21K pretrained weights; extending the analysis to models pretrained on larger-scale datasets or using self-supervised pretraining objectives may alter the relative advantages observed among the different attention types.

The interpretability evaluation relies on bounding-box overlap as the spatial precision metric, which captures coarse object-level localization but does not assess attention alignment at the granularity of individual semantic parts. Future studies could leverage the 15-part keypoint annotations available in CUB-200-2011 to evaluate part-level precision at a finer spatial granularity. Exploring hybrid attention strategies---where different transformer layers employ different attention types depending on their depth in the network---represents a promising research direction that may yield configurations capable of balancing broad context aggregation at early layers with precise fine-grained local focus at later layers. Investigating the interaction between attention mechanism design and self-supervised pretraining objectives---such as masked image modeling and contrastive learning---may further clarify whether the advantages observed in this supervised transfer setting extend to emerging pretraining paradigms.

References

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, 2017.
2. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021.
3. X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, and S. Belongie, "Fine-grained image analysis with deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8927–8948, 2022.
4. M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331–368, 2022.
5. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pp. 10012–10022, 2021.
6. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, pp. 10347–10357, 2021.
7. W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pp. 568–578, 2021.
8. K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Q. Davis, A. Mohiuddin, Ł. Kaiser, D. B. Belanger, L. J. Colwell, and A. Weller, "Rethinking attention with Performers," in *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021.
9. X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," in *Advances in Neural Information Processing Systems 34*, pp. 9355–9366, 2021.
10. Y. Li, "Performance benchmarking and optimization strategies for depth estimation algorithms in unstructured environments," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 2, pp. 32–43, 2026.
11. P. T. Chung, "Comparative evaluation of machine learning algorithms for spectrophotometric dental shade classification," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 1, pp. 204–214, 2026.
12. J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, and C. Wang, "TransFG: A transformer architecture for fine-grained recognition," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pp. 852–860, 2022.
13. Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pp. 1025–1034, 2021.
14. Y. Hu, X. Jin, Y. Zhang, H. Hong, J. Zhang, Y. He, and H. Xue, "RAMS-Trans: Recurrent attention multi-scale transformer for fine-grained image recognition," in *Proceedings of the 29th ACM International Conference on Multimedia (MM 2021)*, pp. 4239–4248, 2021.
15. H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pp. 32–42, 2021.

16. C.-F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, pp. 357–366, 2021.
17. Q. Zhang, "Adaptive differential privacy mechanism for federated document classification: A gradient-clipping optimization approach," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 672–678, Dec. 2025.
18. Y. Wang, "Practical AI approaches for community infection early warning: From public data to actionable insights," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 1545–1552, Dec. 2025.
19. M. Han, "Privacy-preserving collaborative learning across healthcare institutions: An adaptive approach with gradient compression and dynamic privacy budget allocation," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 679–684, Dec. 2025.
20. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin Transformer: A general vision transformer backbone with cross-shaped windows," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 12124–12134, 2022.
21. Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision Transformer with deformable attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 4794–4803, 2022.
22. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*, 2021.
23. Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Computer Vision -- ECCV 2022, Lecture Notes in Computer Science*, vol. 13684, pp. 459–479, Springer, 2022.
24. Y. Wang, "Accuracy evaluation of machine learning-based hospital resource demand forecasting during infectious disease surges: A comparative analysis," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 1, pp. 314–327, 2026.
25. H. Sun, X. He, and Y. Peng, "SIM-Trans: Structure information modeling transformer for fine-grained visual categorization," in *Proceedings of the 30th ACM International Conference on Multimedia (MM 2022)*, pp. 5853–5861, 2022.
26. Y. Zhang, "A comparative study of machine learning methods for automated customer service dialogue quality assessment," *Journal of Science, Innovation & Social Impact*, vol. 2, no. 1, pp. 328–338, 2026.
27. H. Zhu, W. Ke, D. Li, J. Liu, L. Tian, and Y. Shan, "Dual cross-attention learning for fine-grained visual categorization and object re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*, pp. 4692–4702, 2022.
28. K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 87–110, 2023.
29. M. Zhong, "Privacy-preserving federated learning for collaborative risk monitoring across financial institutions: Balancing regulatory compliance and intelligence sharing," *Journal of Sustainability, Policy, and Practice*, vol. 2, no. 2, pp. 44–54, 2026.
30. P. T. Chung, "Data mining methods for biomechanical property prediction of biomedical materials based on optimized feature dimensionality reduction," in *Proceedings of the 2025 6th International Conference on Computer Science and Management Technology*, pp. 174–180, Dec. 2025.